A Simple Diagnostic to Investigate Instrument Validity and Heterogeneous Effects when using a Single Instrument[☆]

Steven G. Dieterle[1,*], Andy Snell[1]

University of Edinburgh, School of Economics, 30 Buccleuch Place, Edinburgh, United Kingdom, EH8 9JT

Abstract

Many studies that use instrumental variables are based on a first stage linear in the instrument. Using only linear first stages may miss important information about effect heterogeneity and instrument validity. Analyzing fifteen studies using linear first stages, we find ten with significant nonlinearities. Six of these ten have statistically different second stage estimates. Additional analysis is necessary when results are sensitive to first stage choice. We provide a framework to reconcile these differences by determining those patterns of heterogeneity that are consistent with instrument validity. If these patterns violate economic reasoning, then the validity of the instrument is questioned.

Keywords: Instrumental Variables, Heterogeneous Treatment Effects, Robustness Check
JEL: C26

1. Introduction

Economists often employ Instrumental Variable (IV) techniques when faced with the difficult task of estimating causal effects in non-experimental settings. The first order issue is to find plausibly exogenous instruments. Given that the necessary exogeneity assumption is effectively untestable, in most cases instrument validity is argued on heuristic grounds. On top of validity concerns, interpretation of IV estimates is made more difficult by allowing for unmodeled heterogeneity in responses, a concept made popular in economics due to the

[☆] We would like to thank seminar participants at CMPO Bristol University, Heriot-Watt University, Michigan State University, the Tinbergen Institute, the University of Bath, University College Dublin, and University of York, as well as two anonymous referees for helpful comments. All errors are the responsibilities of the Authors. An earlier version of the paper was circulated with the title "It's Hip to be Square: Using Quadratic First Stages to Investigate Instrument Validity and Heterogeneous Effects."

[*] Corresponding author

Email addresses: `steven.dieterle@ed.ac.uk` (Steven G. Dieterle), `A.J.Snell@ed.ac.uk` (Andy Snell)

influential work of Imbens and Angrist (1994) and Heckman and Vytlacil (1999).

While there are many ways to implement an IV strategy, one of the most common among applied economists is to use Two Stage Least Squares (2SLS) with the first stage linear in a single instrument.[1] However, using only linear first stages may obscure important information on the nature of heterogeneous effects that can, in turn, augment the heuristic arguments made for instrument validity. We argue that the sensitivity of 2SLS estimates to simple changes in the first stage is an important piece of information that should be routinely reported along with other common diagnostics, like the first stage F-statistic. In this paper, we adapt the heterogeneous effects framework in order to characterize and assess previously undocumented dimensions of heterogeneity that result from using different first stage functions of a single instrument.

To start, we identify cases where the results are sensitive to the first stage functional form by following a basic textbook approach to overidentification testing. In particular, we start by extending the first stage to include a squared term in the instrument.[2] We then test for significance of the quadratic first stage relative to the linear. Finally, we test the sensitivity of the 2SLS estimates to the choice of linear or quadratic first stage using a standard overidentification test— treating the squared instrument as an overidentification restriction. Surprisingly, this simple and nearly costless to implement procedure proves to be empirically relevant when applied to papers relying on first stages linear in a single instrument. Across the fifteen papers we study here, we find evidence of significant nonlinearities in ten papers. Six of these ten studies have cases where the significant quadratic first stage is associated with a statistically significant difference in the 2SLS estimates of interest.

The obvious question- and primary focus of this paper becomes: what should we do

[1] This focus on linear first stages is understandable given that the properties of the estimator are well understood relative to nonparametric approaches. For instance, Hansen (2009) notes the "worrisome" issue that many nonparametric approaches are "incomplete" due to ambiguity over bandwidth selection, an issue "critical to implementation." In addition, it is closely connected to the counter factual outcomes framework used in program evaluation with binary treatment and instruments. Furthermore, in traditional treatments of IV first stage choice only impacts efficiency and not consistency, while with heterogeneous effects different first stages estimate arbitrarily different weighted average partial effects. Researchers may also be cautious of the "forbidden regression" problem of using fitted values from a nonlinear, say Probit, first stage directly in the second stage (Angrist and Pischke, 2009). Coupled with concerns over weak instruments with overidentification, these considerations make the linear first stage choice appealing.

[2] While the arguments we make will also hold for higher order polynomials (and other functional forms), we find that the quadratic first stage is sufficient to uncover evidence of nonlinearity in most cases even when higher order terms would improve the fit. Furthermore, by choosing the quadratic first stage we avoid generating weak instrument problems by adding only one overidentification restriction and we have a simple test that can be uniformly applied across cases to avoid data mining.

when our results are sensitive to the choice of functional form for a single instrument in the first stage?

In a classic treatment of 2SLS with homogeneous effects, different functions of the instrument will affect efficiency, but should identify the same population parameter (Angrist, Graddy, and Imbens, 2000; Heckman, Urzua, and Vytlacil, 2006; Wooldridge, 2010). Therefore, the sensitivity can be cast as evidence of an invalid instrument.[3] Alternatively, the sensitivity may be evidence of unmodeled heterogeneity with different first stages identifying different weighted averages of underlying responses (Angrist, Graddy, and Imbens, 2000; Heckman, Urzua, and Vytlacil, 2006). Such heterogeneity may come from a number of sources including nonlinearity in the second stage relationship, as well as more complex forms due to non-separable errors, or individual level functional form differences.

We provide a framework for extracting information about potential heterogeneity from using different first stages. Building on prior work by Angrist, Graddy, and Imbens (2000), we show that the difference in the estimators (linear and quadratic first-stage) is driven completely by applying different weights to the underlying heterogeneous partial effects at different values of the instrument. Furthermore, we show that the weight ratio at each value of the instrument is easily estimated using only the first stage fitted values without imposing any additional assumptions on the most general heterogeneous effect models. Combined with subsample estimation, the weight ratios allow the researcher to infer the relative pattern of the average partial effects across the distribution of the instrument that would be consistent with a valid instrument.

We argue that the pattern of heterogeneity uncovered by our approach should be checked for a reasonable economic explanation. If it can be matched to a sensible economic story, then we can strengthen our understanding of the question being studied. The results may also justify pursuing more complex estimation approaches, such as nonparametric IV (Newey, 2013) or Local IV (Heckman and Vytlacil, 1999), that tackle effect heterogeneity head on. However, if the pattern does not match a sensible economic story, then the results should be interpreted with caution as it raises concerns over the validity of the instrument.

To illustrate the usefulness of the proposed approach, we compare linear and quadratic first stages for two well-published papers relying on continuous instruments for identification:

[3] This interpretation can be extended to more general cases where heterogeneous effects are independent of the instrument (Heckman, Urzua, and Vytlacil, 2006).

Becker and Woessmann's 2009 paper on the effects of Protestantism on economic prosperity and Acemoglu, Johnson, and Robinson's influential 2001 paper exploring the relationship between institutions and growth. We highlight these two papers as, in each case, we find evidence that adding the square of the instrument to the first stage is important for the final estimates. When exploring the heterogeneous effects explanation for Becker and Woessmann (2009), we find that the implied effects actually change sign (from positive to negative) across the instrument distribution suggesting a very important pattern of heterogeneity. Again this pattern should be matched with a sensible economic story to help bolster the argument for instrument validity.

Since the key papers were chosen to illustrate the important conclusions that may be drawn when non-linear first-stages seem to matter, we also present a survey exercise applying our approach to an objectively chosen set of thirteen papers drawn from *American Economic Association* journals. That we find rejections in over half of the papers underscores the importance of applying this approach generally.

We readily note that while the use of nonlinear transformations of instruments is not, in-and-of-itself, novel, our approach is. This paper is the first to compare estimates from different first stages to show how nonlinearity in the first stage can be exploited to enhance the heuristic arguments for instrument choice by uncovering patterns of heterogeneity with respect to the instrument. Importantly, the patterns of heterogeneity uncovered here typically go unnoticed in empirical work. Our approach also compliments recent work by Lochner and Moretti (2011) and Løken, Mogstad, and Wiswall (2012) that considers the importance of nonlinear second stages for typical instrumental variable estimators. The key point of distinction here is our focus on using the nonlinearity in the first stage to test the sensitivity of 2SLS estimates.

The paper proceeds as follows: section 2 discusses the motivation for considering higher order terms in the homogeneous effects setting; section 3 applies this approach to the two key examples; section 4 shows how to characterize the weight ratios in a heterogeneous effects framework and applies this to the Becker and Woessmann (2009) example; section 5 summarizes the literature survey exercise; and section 6 concludes.

2. Quadratic Overidentification Test

To motivate our approach, we begin with a simple text-book treatment of instrumental variables. Later, we will consider the implications under a more general heterogeneous effects setting. Following Wooldridge (2010), start with a linear model for y in terms of x in the population:

$$y = \mathbf{x}\beta + u \tag{2.1}$$

$$\text{where } \mathbf{x} = (1, x_2, ..., x_K) \text{ is a vector of covariates}$$

Further denote our instrument vector by $\mathbf{z} = (1, x_2, ..., x_{K-1}, z)$, where we assume one endogenous regressor (x_K) and a single excluded instrument (z).[4] Under the following conditions the Two-Stage Least Squares (2SLS) estimate, $\hat{\beta}$, is consistent for β:

[A1] $$E(u|\mathbf{z}) = 0$$

[A2] $$\text{rank } E(\mathbf{z}'\mathbf{x}) = K$$

Assumption [A1]— mean independence— is the key assumption needed for consistent 2SLS estimation.[5] Here we opt for the mean independence assumption instead of assuming that z and u are uncorrelated. While mean independence is a stronger assumption, when arguing for the validity of an instrument the distinction between uncorrelatedness and mean independence is seldom pursued by researchers. Indeed, when relying on a "natural experiment" for identification it is typical to rely on arguments that implicitly evoke a notion of independence. Further, it is often difficult in such cases to derive a sensible economic argument for why an instrument is plausibly uncorrelated with the error term, but may not be mean independent. For instance, in the Acemoglu, Johnson, and Robinson (2001) paper we study in Section 3.2 the authors discuss the required exogeneity assumption by stating the following:

> "The exclusion restriction implied by our instrumental variable regression is that, conditional on the controls included in the regression, the mortality rates of Eu-

[4] Throughout, we will generally refer to the outcome, endogenous explanatory variable, and instrument as y, x, and z, respectively.

[5] [A2] is the rank condition requiring z to be linearly related to the endogenous regressor (x_K).

> ropean settlers more than 100 years ago *have no effect* [emphasis added] on GDP per capita today, other than their effect through institutional development."

This statement is much stronger than simply suggesting uncorrelatedness. In contrast, arguments for an instrument that is uncorrelated but not mean independent would likely require a more particular discussion of the data generating process.[6]

Under mean independence, not only is z a valid instrument, but so is any function of z. This fact motivates a simple test of the sensitivity to first stage choice based on a standard overidentification test. Namely, replace the linear-in-z first stage with a quadratic-in-z and conduct the overidentification test. Obviously, one could consider other first stage functions of z,[7] however we chose to focus on the quadratic-in-z first stage as it is simple to implement uniformly across cases (i.e. low cost to the researcher and avoids data-mining) while still capturing a key component of potential nonlinearities.[8]

Rejecting the null in this case implies that the two instruments lead to statistically different estimates of β.[9] Formally, this is a rejection of the linear-homogeneous-effects model in equation (2.1) under the mean independence assumption. If the source of the rejection is a violation of mean independence, then following our discussion on the distinction between mean independence and the weaker uncorrelatedness assumption needed for identification the validity of the instrument would be questioned. However, given that the rejection may come from either a failure of mean independence or misspecification in equation (2.1), we prefer to interpret the result more generally as evidence of sensitivity to first stage choice. Regardless of the source of the rejection, this sensitivity is very important for understanding

[6] For instance, a common statistical example of uncorrelated but dependent variables is if X is symmetrically distributed around the origin and $Y = X^2$,then X and Y are clearly dependent but $Cov(Y, X) = E[YX] = E[X^3] = 0$ and Y and X are uncorrelated.

[7] One could use higher order polynomials, creating categorical dummy variables, or account for a non-continuous x (Probit fitted values as the instrument when x is binary).

[8] If the second stage is properly specified, one could choose the "best" fitting first stage for efficiency reasons. However, in the heterogeneous effects framework, the concept of the "best" (rather than best fitting) first stage function becomes much less clear.

[9] An important consideration for our test is the bias of 2SLS. It is well known that 2SLS estimates are consistent but not unbiased and that this bias is most severe when instruments are weak and there are several overidentification restrictions (Angrist and Pischke, 2009). In our context, we might be concerned that adding z^2 may introduce or exacerbate a weak instrument problem. To account for weak instruments, we follow two common approaches. Following Stock, Wright, and Yogo (2002), we report first stage F-stats. We also estimate β by Limited Information Maximum Likelihood (LIML). In overidentified models LIML and 2SLS estimates have the same probability limit but different small sample properties. In particular, under certain assumptions 2SLS is biased toward OLS while LIML is roughly "median-unbiased" (Angrist and Pischke, 2009). A comparison of the 2SLS and LIML estimates provides a useful "eyeball test" of the weak instrument problem.

the economic conclusions that can be drawn from the estimates. Robustness to first stage choice is just as interesting, as it provides additional justification for estimating model (2.1) under mean independence.

A particularly important form of misspecification could come from unmodeled heterogeneous response to x. Unmodeled heterogeneity could take many forms including nonlinearity, non-separable errors, or individual differences in functional relationship between y and x. This leads directly to the modern heterogeneous effect interpretations of linear 2SLS estimates found in Imbens and Angrist (1994) and Heckman and Vytlacil (1999). In Section 4, we will use a heterogeneous effects framework in order to characterize the implied patterns of heterogeneity that are consistent with the difference in coefficient estimates when using a linear-in-z or quadratic-in-z first stage. At this point, we simply want to emphasize the fact that sensitivity to the first stage warrants additional investigation.

As a final note, it is worth clarifying two features of our approach. First, as will become clear in the heterogeneous effects framework, we are not suggesting that the quadratic first stage is “preferred” in any way to the linear. In the homogeneous effects setting, it would likely be best to follow the literature on identifying the optimal instrument vector in order to improve efficiency. In the heterogeneous effects framework in section 4, preference for a particular first stage is much less clear as they provide, arguably arbitrarily, different weighted averages of heterogeneous effects. Rather, we will show that even in the heterogeneous effects world, there is valuable information to be learned by considering the quadratic-in-z result along with the commonly used linear first stage results when the two results differ. Second, it is helpful to distinguish the role of functional form and the precise specification in the first stage versus the second stage. In the second stage, functional form is directly tied to the economics of the relationship of interest. The first stage, however, is used to isolate plausibly exogenous variation in x. In the homogeneous effects framework, choosing the first stage based on best fit or on economic grounds will simply affect the efficiency, while with heterogeneous effects it will simply identify a different, and not necessarily preferred, weighted average of the effect of interest.

3. Key Examples

In this section, we provide two examples to illustrate the potential for using a quadratic of continuous instruments as discussed in section 2 to push forward economic analysis. The two papers considered, Becker and Woessmann (2009) and Acemoglu, Johnson, and Robinson (2001) were both well published (*The Quarterly Journal of Economics* and the *American Economic Review*, respectively), utilize clever and innovative approaches to answer important causal questions in economic development based on a continuous instrument, and have data that is readily available to other researchers.

3.1. Becker and Woessmann (2009): Prussia, Protestants, and Prosperity

In Becker and Woessmann (2009) (BW), the authors explore the link between Protestantism and economic prosperity in 19th century Prussia. In order to identify the causal effect of Protestantism on economic outcomes, the authors take the innovative approach of using the distance from Wittenberg as an instrument for Protestantism.

BW provide a set of 2SLS estimates based on the following general specification with county-level data:

$$y_i = \alpha + \beta PROT_i + x_i \phi + u_i \tag{3.1}$$

where y_i is one of four human capital/economic outcomes

$PROT_i$ is the share of Protestants

x_i is a set of demographic controls

The four outcomes they consider are: the Literacy Rate in 1871, the Income Tax per capita in 1877, Log Average Annual Income for Male Teachers in 1886, and the Average Population Share in Non-agriculture in 1882. The instrument for the share of Protestants is always the Distance from Wittenberg. Using this approach, BW find statistically and practically significant effects of Protestantism on each outcome. The results are replicated in column (1) of Table 3.1. For the literacy outcome, the coefficient implies an 18.9 percentage point increase in literacy by moving from a county with no Protestants to all Protestants. BW note that the effect on per capita Income Tax is roughly equivalent to 29.6% of the average income tax in their data. Finally, an all-protestant county is estimated to have Log Teacher Pay 10.5% higher and have 8.2 percentage points higher non-agricultural workforce. These effects are quite large and signify a meaningful role for Protestantism in 19th century economic development for Prussia. Importantly for their identification strategy, the first stage F-statistic for the instrument is over 74, which is suggestive of a strong first stage and is well above the Stock and Yogo (2002) rule of thumb of 10.

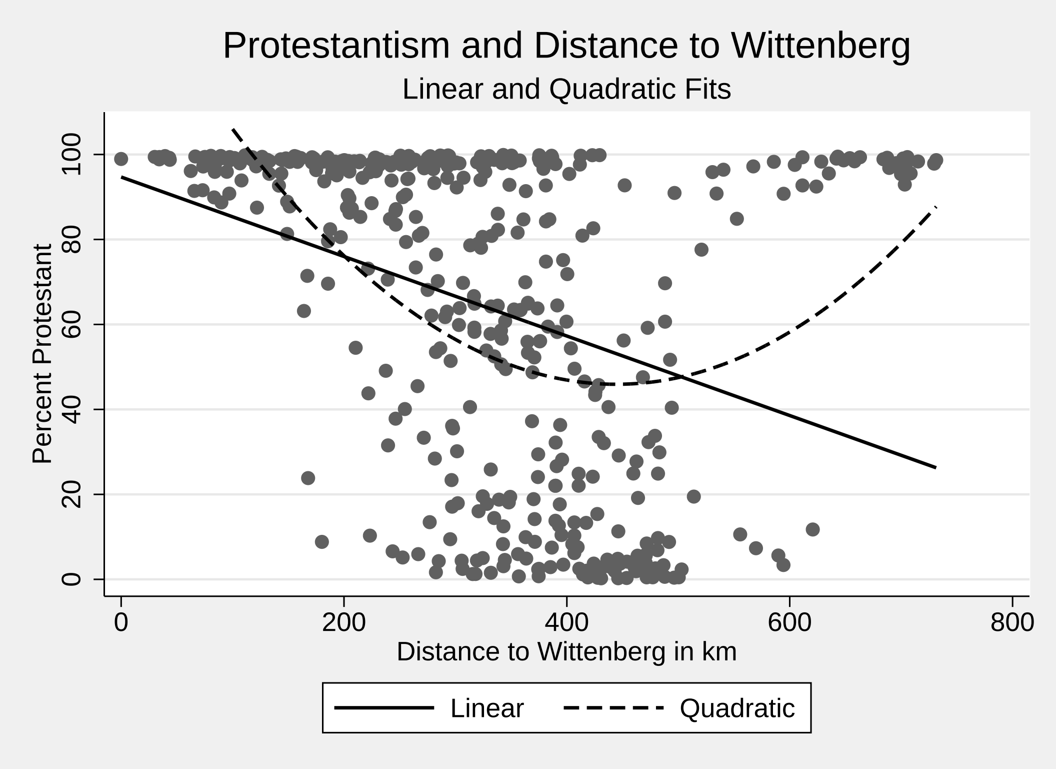
Table 3.1

Becker and Woessmann (2009) Replication and Extension				
		Linear	**Quadratic**	
Outcome	**Statistic**	*2SLS*	*2SLS*	*LIML*
Literacy Rate	$\hat{\beta}$	0.1885***	0.0932***	0.0898***
	s.e.	(0.0280)	(0.0205)	(0.0246)
	First Stage F	74.19	64.75	
	Overid p-value		0.0000	
	Z^2 p-value		0.0000	
Income Tax Per Capita	$\hat{\beta}$	0.5865**	-0.0219	-0.0647
	s.e.	(0.2326)	(0.1829)	(0.1996)
	First Stage F	75.07	66.02	
	Overid p-value		0.0000	
	Z^2 p-value		0.0000	
Log Teacher Salary	$\hat{\beta}$	0.1047**	0.0165	0.0123
	s.e.	(0.0493)	(0.0392)	(0.0406)
	First Stage F	75.07	66.02	
	Overid p-value		0.0033	
	Z^2 p-value		0.0000	
Manufacturing & Service Workers	$\hat{\beta}$	0.0821**	0.0336	0.0335
	s.e.	(0.0381)	(0.0299)	(0.0304)
	First Stage F	75.07	66.02	
	Overid p-value		0.0337	
	Z^2 p-value		0.0000	

BW provide a number of sensitivity checks to support the validity and robustness of their results. Here we add our proposed quadratic overidentification test. To help motivate the potential for the quadratic first stage to provide additional information, Figure 3.1 provides a scatter plot of the first stage relationship—Protestantism on Distance to Wittenberg—as well as both linear and quadratic fitted lines. Clearly, the quadratic fit implies a very different first stage relation than the linear suggesting that the two rely on different variation in Protestantism for identification.

Starting with the Literacy outcome, we see the estimated effect of Protestantism fall by half, from 18.9 percentage points to 9.3. Importantly, the overidentification test easily rejects with a p-value zero to four decimal points. As we argued in the previous section, this sensitivity to the choice of first stage requires additional investigation and caution when interpreting the results. For the literacy outcome, the estimated effect using the quadratic first stage is meaningfully different from the linear first stage; however, it is still positive and

Figure 3.1: Becker and Woessmann (2009) First Stage Scatter Plot



statistically different from zero. In the heterogeneous effects framework of the next section, this may lead us to conclude that there is some heterogeneity in the return to Protestantism but that the overall relationship is still intact. The quadratic-in-z 2SLS estimates for the other outcomes are perhaps more worrisome, as they become much smaller and are no longer statistically different from zero at conventional levels. For instance, the estimated effect on per capita Income Tax changes sign and is only 3% as large as the linear-in-z estimate. Allowing for heterogeneous effects in this case will lead to a very inconclusive picture of the relationship between Protestantism and Income Tax in 19th century Prussia.

It is important to highlight three key points regarding the relevance and strength of the added instrument, Distance to Wittenberg Squared. While the first stage F-stat does fall when including the squared instrument, at 64.75 it is still well above 10 and indicative of a strong first stage. In addition, the t-test for the coefficient on the squared distance is a further test of the relevance of the squared term. The p-value for that test is 0.0000, indicating that the squared term does help in predicting Protestantism. Finally, in column

(3) we present the Quadratic LIML estimate as well. That the LIML estimate of 8.98 percentage points is very close to the 2SLS estimate is also suggestive that we have not introduced a weak instrument problem.

3.2. Acemoglu, Johnson, and Robinson (2001): Settler Mortality, Institutions, and Development

Acemoglu, Johnson, and Robinson (2001) (AJR) explore the role institutions play in shaping economic development. AJR approach the problem of identifying the causal link from institutions to growth by trying to isolate variation in present day institutions that is driven by different conditions, measured by mortality rates, at the time of colonial settlement.[10]

Using cross-sectional data on 64 countries, AJR estimate a series of regressions based on the following second stage:

$$GDP_i = \alpha + \beta RISK_i + \phi LAT_i + u_i \tag{3.2}$$

where GDP_i is Log GDP per Capita in 1995

$RISK_i$ is a measure of the protection from expropriation

LAT_i is the Latitude of the country

The key explanatory variable, the protection from expropriation, is measured on a scale from 0 (lowest protection) to 10 (highest protection) with a sample mean and standard deviation of 6.5 and 1.5, respectively. Given the small sample size, AJR explore the robustness of their results by considering different subsamples and additional, albeit limited, controls. Column (1) of Table 3.2 displays the replication of a select set of AJR's baseline estimates. AJR present results both with and without the Latitude control showing little difference in the estimates of β, however for space considerations we only display the estimates when including Latitude. The coefficient estimate of 0.9957 found in row (1) for AJR's base case implies that a one standard deviation (1.5) increase in protection from expropriation leads

[10] The original AJR paper has been highly influential and has spurred a lengthy debate centered on the quality of the data used and methodological considerations (See Albouy (2012) and Acemoglu, Johnson, and Robinson (2012) for the published comment and reply). We focus here on the original data and estimation methodology. In Appendix C, we comment on the broader debate by exploring the implications of using higher order polynomials of the instrument with the alternative data and methods described in Albouy (2012).

to over a three-and-a-half-fold increase in per capita GDP ($e^{(1.5)(0.9957)} - 1 \approx 3.5$). This is certainly a sizable difference driven by institutional differences. Rows (2) and (3) display estimates based on subsamples excluding “NeoEuropes” (United States, Canada, Australia, and New Zealand) and African countries, respectively. The coefficient estimate is larger than the base case when excluding NeoEuropes and smaller when excluding Africa. Finally, the relationship remains largely intact when including continent dummy variables. In all, the estimates imply anywhere from a 75% to over a five fold increase in GDP per capita from a one standard deviation increase in protection from expropriation.

Table 3.2

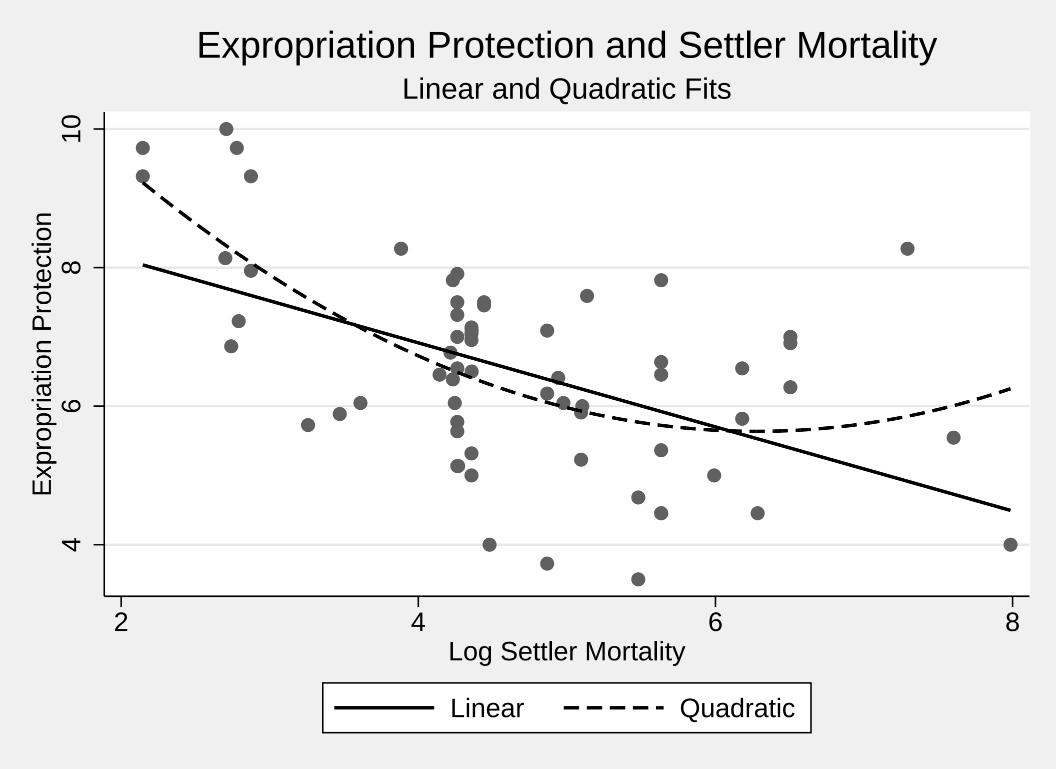
Acemoglu, Johnson, and Robinson (2001) Replication and Extension

Sample &		**Linear**	**Quadratic**	
Specification	**Statistic**	*2SLS*	*2SLS*	*LIML*
Base	$\hat{\beta}$	0.9957***	0.7356***	0.8740***
	s.e.	(0.2164)	(0.1356)	(0.1806)
	First Stage F	13.09	11.33	
	Overid p-value		0.0098	
	Z^2 p-value		0.0062	
Excluding	$\hat{\beta}$	1.2118***	0.9938***	1.1501***
Neo-Europes	s.e.	(0.3453)	(0.2491)	(0.3202)
	First Stage F	7.83	5.25	
	Overid p-value		0.1151	
	Z^2 p-value		0.1214	
Excluding	$\hat{\beta}$	0.5757***	0.5698***	0.5701***
Africa	s.e.	(0.1124)	(0.1083)	(0.1084)
	First Stage F	21.61	11.75	
	Overid p-value		0.8386	
	Z^2 p-value		0.2222	
Base w/	$\hat{\beta}$	1.1071**	0.7019***	0.8194***
Continent	s.e.	(0.4413)	(0.1712)	(0.2240)
Indicators	First Stage F	3.46	5.28	
	Overid p-value		0.0760	
	Z^2 p-value		0.0119	

All specifications include latitude as an additional covariate

Figure 3.2 plots the first stage relationship from AJR, again showing scope for the quadratic to provide a different source of variation from the linear first stage. In column (2) of Table 3.2 we present our results from using the quadratic of the mortality rate in the first stage. The overidentification test rejects at the 10% level for the base sample both with and without continent dummies. In both cases the estimated coefficient is considerably

Figure 3.2: Acemoglu, Johnson, and Robinson (2001) First Stage Scatter Plot



smaller. For the base sample, the change in the point estimate suggests a drop from a 350% to roughly a 300% increase in per capita GDP for a one standard deviation increase in protection from expropriation. The test for the sample excluding NeoEuropes does not reject at common significance levels, but with a p-value of 0.12, it is not surprising that the point estimates from the linear- and quadratic-in-z first stages are still quite different. In all three cases, the results of the overidentification test and the comparison between the original linear-in-z and quadratic-in-z estimates raises concerns over the interpretation of the results However, there are questions about the first-stage strength, with F-stats either below or barely above 10. Furthermore, the LIML estimates are all noticeably different from 2SLS, suggesting further caution.

Interestingly, the overidentification test fails to reject the null at any reasonable level (p-value=0.84) for the subsample excluding Africa. In this case, the first stage F-stat is just above 10 and the three estimates are all very close to 0.57 (corresponding to a 75% increase in per capita GDP for a one standard deviation increase in protection). In this sense, the results for the Non-African subsample are perhaps the most robust to possible violations of mean independence or misspecification. More generally, the fact that the smallest estimates from the original set of sample sensitivity checks are the most robust to first stage specification may be important for the conclusions that can be drawn. However, overidentification tests can be misleading in the sense that we will tend to fail to reject the null in the presence of bad instruments if the two instruments lead to similar biases. In this case, we might fail to reject the null when the instrument is invalid if the true first-stage relationship is approximately linear (i.e. the squared term is irrelevant once we control for the linear effect). Indeed, the p-value for the test for the coefficient on the squared instrument in the first stage is 0.22, although the small sample size certainly contributes to the weaker results.

4. Heterogeneous Treatment Effects

In this section, we analyze our proposed overidentification test within the modern heterogeneous effects framework used to interpret IV estimates. In this setting, a rejection of the overidentification test could result from estimating a different average partial effect. Generally, this has led people to conclude that "overidentification testing...is out the window in a fully heterogeneous world" (Angrist and Pischke (2009), pg. 166). However, in this case, we only change the weights applied to each partial effect in a particular way that can be estimated quite generally. If we proceed under the assumption that z is valid, we can consider the particular nature of the heterogeneous effects needed to explain the change in the point estimates. Explicitly, we derive and estimate the ratio of weights placed on partial effects at different values of the instrument by the linear- and quadratic-in-z estimates. We then use the estimated change in the weights at each value of z to uncover the pattern of heterogeneous effects that would be needed to account for the change in the point estimates. At the very least the required patterns may be economically interesting. Alternatively if the patterns are inconsistent with economic theory, then it may raise doubts over the validity of the instrument.

4.1. General Framework

Our discussion will follow closely from the framework laid out in Angrist, Graddy, and Imbens (2000) for continuous instruments.[11] Adapting the Angrist, Graddy, and Imbens (2000) setup to a more general case, we are interested in the effect of a possible endogenous x on some outcome y and hope to use an instrument, z. At this point we adopt a very general model for y and x:[12]

$$\begin{aligned} y_i &= y_i(x, z) \qquad (4.1) \\ x_i &= x_i(z) \\ &\text{where } y, x, \text{ and } z \text{ are scalars} \end{aligned}$$

Note, that this setup allows for individual specific relationships between y, x, and z. Our interest lies in interpreting the 2SLS estimates found in section 2 based on the following linear specification for y when the true model is given by (4.1):

$$y_i = \beta x_i + u_i \tag{4.2}$$

If we denote the first stage function of the instrument by $g(z)$, then under the assumptions outlined in Appendix B.1, Angrist, Graddy, and Imbens (2000) show that the IV estimator based on the ratio of covariances between y and $g(z)$ and x and $g(z)$— the probability limit of a 2SLS estimate of β from (4.2)— can be expressed as the weighted average of heterogeneous partial effects:

$$\beta_g = \frac{Cov(y_i, g(z_i))}{Cov(x_i, g(z_i))} = \int \beta(z) \cdot \lambda_g(z) dz \tag{4.3}$$

[11] We chose to follow the Angrist, Graddy, and Imbens (2000) setup over the alternative heterogeneous effects framework of Heckman, Urzua, and Vytlacil (2006) for a few reasons. First, our main goal in the current paper is to add to the heuristic arguments for instrument choice commonly made in applied literature, rather than provide an alternative heterogeneous effects estimate. The Heckman, Urzua, and Vytlacil (2006) approach may be better suited for the latter. However their approach is framed in terms of heterogeneity across the distribution of unobservables in an underlying selection equation, while the Angrist, Graddy, and Imbens (2000) setup is based on heterogeneity across the instrument distribution. This focus on the heterogeneity in terms of the instrument is clearly better suited to our goal. Additionally, the Heckman, Urzua, and Vytlacil (2006) focus on the propensity score for binary treatment is less appropriate for the current setting given our examples with continuous endogenous variables.

[12] Here we work through the case with no additional covariates. In Appendix B.4, we discuss the extension with covariates.

The partial effects, $\beta(z)$ take the following form:

$$\beta(z) = E\left[\frac{\partial y}{\partial x}(x_i(z))\right]$$

Note that the average partial effect $\beta(z)$ is the expectation over the partial effects of x on y across all units for a given value of z. That is, $\beta(z)$ is an average of unit-specific partial effects at potentially different levels of x. The motivation for Angrist and Pischke's comment that overidentification testing is uninformative in the heterogeneous effects setting can be seen here. In the traditional use of overidentification tests, researchers compare two distinct instruments. As a result, the $\beta(z)$ that we are averaging over will be different, implying a completely different estimand. By focusing on different functions of the *same* instrument the underlying partial effects we are averaging, the $\beta(z)$, are unchanged allowing us to extract useful information from an overidentification test.

To simplify notation, we normalize x and z to have mean zero. We show in Appendix B.1 that after demeaning we can write the weights, $\lambda(z)$, as follows:

$$\lambda_g(z) = \frac{\frac{\partial x}{\partial z}(z) \cdot E\left[g(\zeta)|g(\zeta) > g(z)\right] Pr\left(g(\zeta) > g(z)\right)}{Cov(x_i, g(z_i))} \tag{4.4}$$

Expressed this way, the weight consists of two main components, one determined by the form of the chosen first stage, $g(z)$, and the other the true partial effect of z on x. This second term, $\partial x/\partial z$, represents the heterogeneous responses to the instrument. Larger responses are given more weight, a concept closely tied to the characterization of Always Takers, Never Takers, and Compliers in the binary treatment and instrument setting. Just as in the binary case, the IV estimate (regardless of first stage choice) will be a weighted average for compliers only ($\partial x/\partial z \neq 0$). Always Takers and Never Takers, units not induced to change x when z changes ($\partial x/\partial z = 0$), will not contribute to the estimates no matter the choice of $g(\cdot)$ as can be seen from equation (4.4).

From here we can derive a very general result for the ratio of the weights when using two different first stage functions, $g_1(z)$ and $g_2(z)$:

$$\frac{\lambda_2(z)}{\lambda_1(z)} = \left[\frac{Cov\left(x_i, g_1(z_i)\right)}{Cov\left(x_i, g_2(z_i)\right)}\right]\left[\frac{E\left[g_2(\zeta)|g_2(\zeta) > g_2(z)\right] \cdot Pr(g_2(\zeta) > g_2(z))}{E\left[g_1(\zeta)|g_1(\zeta) > g_1(z)\right] \cdot Pr(g_1(\zeta) > g_1(z))}\right] \tag{4.5}$$

Since we consider different functions of the same instrument, $\partial x/\partial z$, cancels out in the ratio. With different instruments, the other components of the estimator, $\beta(z)$ and $\partial x/\partial z$, would change as well. Importantly, since these two components represent the very general relationships that underlie our estimates, this implies we do not need to make any further assumptions on the true model in order to compare the weights using different first stage functions.

Estimating the weight ratio is fairly straight forward.[13] We simply calculate the sample analogue to the conditional expectations, probabilities, and covariances. For any value of z we can use the fitted values from the first stage, denoted $\hat{g}(z)$, to estimate each component. We must order the observations by $\hat{g}(z)$ to estimate the conditional expectation as the mean of $\hat{g}(z)$ for all observations with a larger $\hat{g}(z)$. We can also estimate the probability as the fraction of observations with a larger fitted value. Finally, we can estimate the covariances quite generally by using the corresponding sample covariances.

Once more, we are not asserting a preference for the quadratic over the linear first stage. In the current context, it is not clear which set of weights are preferred, rather we simply want to exploit the differences to learn more about the instrument and the economic effect of interest in the second stage.

4.2. Empirical Example: Distance to Wittenberg

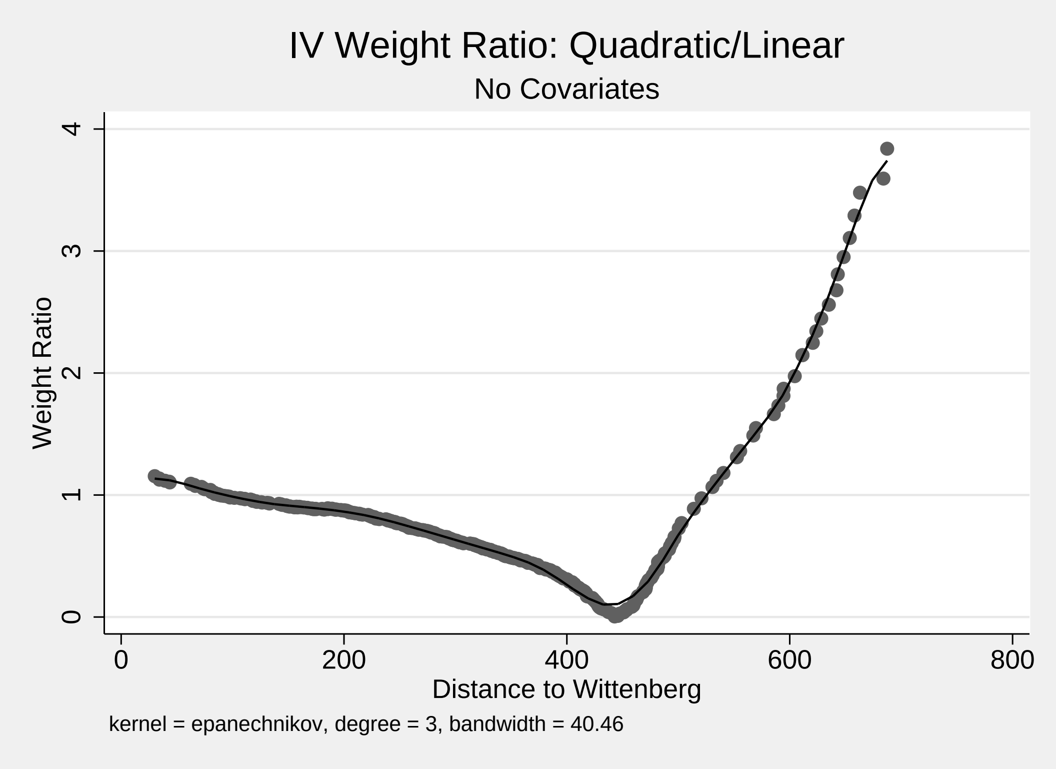
In order to illustrate the usefulness of the above approach, we return to the the example in section 3 from Becker and Woessmann (2009) looking at how the spread of Protestantism affected social and economic outcomes in 1800s Prussia. Once more, the key to the identification strategy is to use the distance from Wittenberg as an instrument for the fraction of a county that was protestant. We start with a simplified case using the basic setup from BW to estimate the effect of Protestantism on literacy, but omitting the additional control variables. To be clear, this does not represent BW's preferred approach and is done to provide a cleaner interpretation and illustration of the information that can be gathered by estimating the weight ratios.[14] Without covariates, the linear-in-z estimate is $\hat{\beta}_1 = 0.42$ while the quadratic-in-z is $\hat{\beta}_2 = 0.15$.

[13] Both Angrist, Graddy, and Imbens (2000) and Heckman, Urzua, and Vytlacil (2006) provide estimates of IV weights in the discrete case, but not in the continuous case. Our focus on the weight ratios avoids making semi-parametric assumptions on the x-z relationship. For instance, in the supplementary material for Heckman, Urzua, and Vytlacil (2006), they use a series of linear projections and Probit models to approximate and estimate weight components in the binary treatment case.

[14] Appendix B.4 discusses the case with covariates.

In Figure 4.1,[15] we plot the estimated weight ratio for all observed values of z, omitting those in the far right tail as they are quite large and obscure the general pattern.[16] The figure is helpful for making comparisons between the two estimators at a given value of z. If the weight ratio is above one, the quadratic-in-z estimate places more weight on the $\beta(z)$ at that value of the instrument while if it is below one the opposite is true. Comparisons across values of the instrument are more difficult since the absolute magnitudes of the weights depend on $\partial x/\partial z$, a term that cancels out in the ratio.

Figure 4.1: Becker and Woessmann (2009) IV Weight Ratio without Covariates



Estimated weight ratios based on the sample analogue of equation (4.5). For each observed value of z, we use the fitted values for the linear and quadratic first stages— $\hat{g}_1(z)$ and $\hat{g}_2(z)$— to estimate the sample mean, probabilities, and covariances.

[15] Figure B.1 in Appendix B.2 depicts the same weight ratios with bootstrapped 95% confidence intervals (CI).

[16] Note that very large weight ratios are not, in-and-of-themselves a sign of unreasonable weights. For instance, the very small weight ratios in Figure 4.1 near 425km would be very large if we presented the linear-to-quadratic ratio instead. Indeed, in Appendix B.3 we provide evidence that the large quadratic-to-linear weight ratios are not due to particularly unreasonable weights for either estimator based on a comparison of the components of the weight ratio.

Here we see that the quadratic first stage puts more weight on partial effects of Protestantism on literacy for counties that are either less than 100km or more than 525km from Wittenberg than the linear first stage. For instance, partial effects tied to counties 600km from Wittenberg are given roughly double the weight by the quadratic-in-z estimate than the linear-in-z. The overall impact of this doubling of the weights on the final estimate depends on the level of "compliance" with the instrument, captured by $\partial x_i/\partial z$, at 600km from Wittenberg.[17] Considering the overall pattern of weights seen in Figure 4.1 and assuming z is a valid instrument, it must be the case that the partial effects are *on average* smaller for counties when they are either very close to or farther away from Wittenberg in order to explain why $\hat{\beta}_2 < \hat{\beta}_1$. An interesting question that emerges from this is whether there is a sensible economic rationale for such a relationship to exist. That is, why might the changes in Protestantism driven by changes in distance from Wittenberg have more bite at intermediate distances? The following section discusses a procedure to further exploit first stage nonlinearities to uncover more about the pattern of heterogeneity to better address this issue.

4.3. Exploring the Pattern of Heterogeneity

While the pattern implied by the weight ratios begins to provide insight into the nature of heterogeneity, ideally we would like to be able to identify partial effects at different values of the instrument. This can be difficult when allowing fully for heterogeneity, however we propose a simple procedure to reveal more about the structure of heterogeneity. We view this as an exploratory descriptive approach, rather than a more formal estimation technique. Our goal is to help inform the heuristic arguments for instrument choice by describing and assessing the economic sensibility of the implied pattern of partial effects that is consistent with instrument validity.

One way to explore heterogeneity in partial effects of x on y at different z is to partition the sample based on the instrument and estimate separate 2SLS regressions in each region of the instrument distribution. Choosing how to partition the data is the key consideration. Rather than making an arbitrary choice, we choose to divide the data into equal size groups until the squared instrument is no longer significant at the 5% level within any of the regions.

[17] For instance, if these counties are approximately Always or Never Takers ($\partial x/\partial z \approx 0$), then doubling the weight will have no effect on the final estimate.

That is, we first split the data at the median of the instrument distribution and test the hypothesis that the coefficient on z^2 in each region is zero. If it is significant in either region (above or below the median), we then split the sample again into terciles and so on. In the BW case, this leads us to partition the sample into quartiles of the distance from Wittenberg. This approach is appealing as it separates the sample into equal sized groups where the first stage is approximately linear so that the choice of first stage becomes less likely to yield different results. Of course, estimating by 2SLS still gives a weighted average of the partial effects within each region.

Figure 4.2 displays the 2SLS coefficient estimates and 95% confidence intervals for three outcomes: Literacy Rate, Log Teacher Salary, and the Share of Manufacturing and Service workers.[18] Across all three outcomes, we find positive estimated effects of Protestantism in areas very close to Wittenberg. However, the point estimate is only statistically significant from zero at the 5% level for the Literacy Rate. In each case, the estimated effects become negative and imprecise at intermediate distances.[19] Farthest from Wittenberg, the estimates are smaller in magnitude but still negative and statistically significant.

Importantly these patterns are consistent with the estimated weight ratios and the original linear and quadratic-in-z estimates. For example, if we consider the Log Teacher Salary outcome, the original linear-in-z estimate was large and positive and the quadratic-in-z was close to zero. Panel B of Figure 4.2 suggests that the effects of Protestantism are positive close to Wittenberg and negative farther way. By placing more weight on counties far away (see figure 4.1) the quadratic puts more weight on the negative partial effects and less on the positive, resulting in a weighted average much closer to zero.

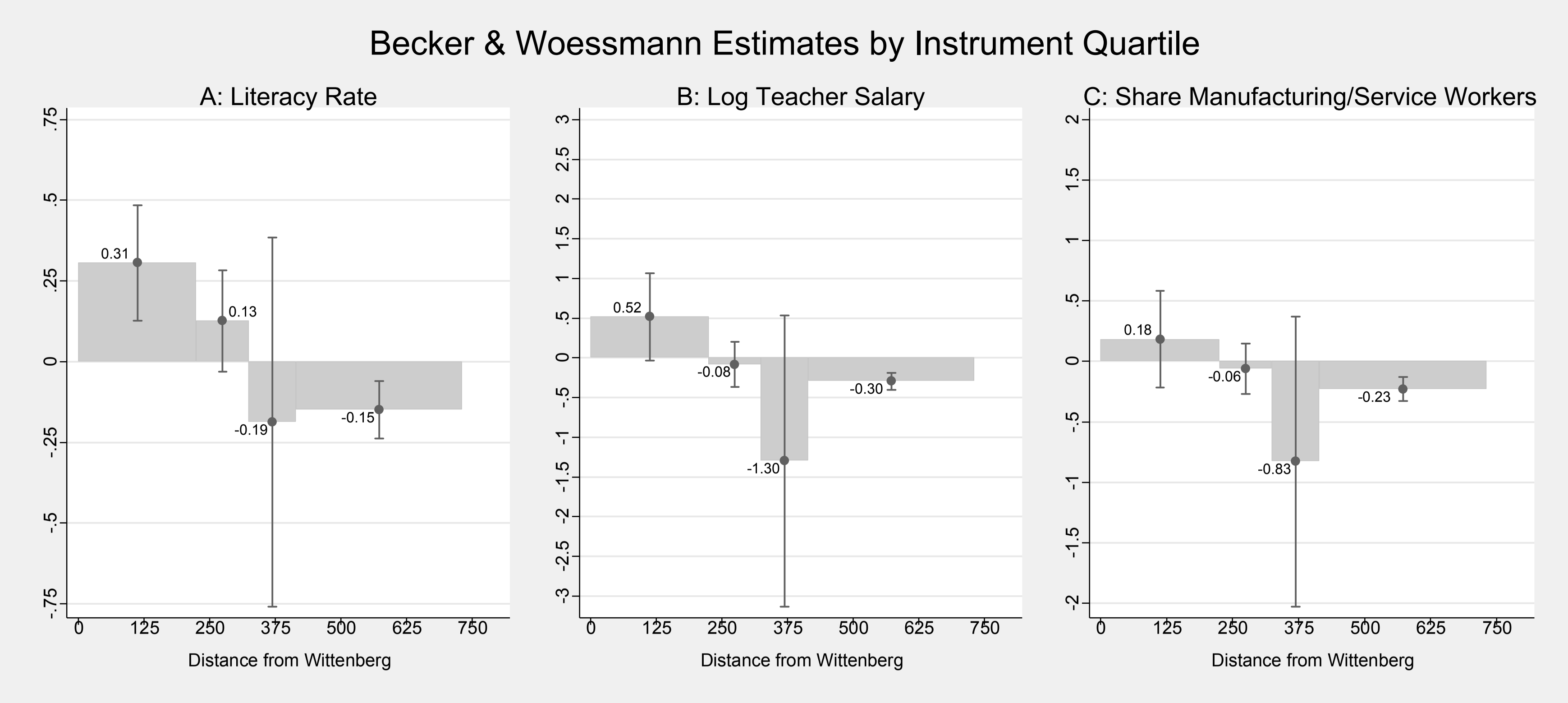
The main point of this section has been to ascertain patterns of heterogeneous effects that are consistent with the significant difference we found in the quadratic-in-z versus linear-in-z 2SLS estimates. A key question now is whether these patterns make economic sense. Much like the case when the difference between IV and OLS estimates goes in the opposite direction of what was expected (e.g IV returns to education are larger than presumably upward biased OLS estimates), it is important to consider a sensible *economic* story for

[18] The fourth outcome, Income Tax per Capita, is excluded as missing data leads to a smaller sample and different first stage.

[19] Note that the imprecisely estimated relationship in the third quartile is not driving any of the full sample results. Indeed, omitting the third quartile of the instrument distribution yields very similar linear- and quadratic-in-z estimates of 0.48 and 0.19— compared to the estimates with no covariates of of 0.42 and 0.15 noted above.

why this is the case (e.g. instruments used for returns to schooling affect those with the highest marginal cost of schooling and, therefore, the highest return on the margin). Here, it would be necessary to explain why counties close to Wittenberg had positive effects from Protestantism while those farthest away were, contrary to the motivating story, adversely affected.

Figure 4.2: Becker and Woessmann (2009) Heterogeneity Estimates



Each figure depicts the estimated coefficient on Percent Protestant from separate 2SLS regressions by quartile of the instrument, Distance from Wittenberg. In each case, the first stage specification is linear in Distance from Wittenberg.

5. Literature Survey and Extension

The examples discussed in Section 3 were chosen because they illustrate cases in which using the quadratic in a continuous instrument lead to qualitatively different results. Here, we survey the results of replicating and applying our procedure to a larger set of objectively chosen examples. This survey will help to establish the relevance of our procedure more broadly and help to highlight the range of results one might find in applying the quadratic overidentification test.

In order to collect a sufficient set of papers in an objective way, we used the American Economic Association (AEA) online journal search. This search includes the *American Economic Review* and the four *American Economic Journal* field journals: Applied, Macro, Micro, and Policy. The papers considered were chosen based on the following criteria:

1. Found by searching“instrument”
2. Have at least one estimate using 2SLS where there was one endogenous regressor and one continuous instrument
3. Data was available on the AEA website

This selection criteria yielded thirteen separate papers published between 2008 and 2013, each containing multiple examples.

For each paper, an estimate was replicated if it met our key criteria. However, since it is common to explore different sets of control variables while trying to estimate the same fundamental relationship with the same instrument, we only replicated one estimate for every y-x-z pairing. We attempted to select a “baseline” or preferred specification for each y-x-z pair. As a caveat, when a preferred specification was not made explicit in the paper, we used our own judgment. For instance, we might choose a specification as the baseline if it was used in subsequent sensitivity analysis. Importantly, the choice of specification was made *before* any replication was done to avoid making selections based on the results of the overidentification test. We argue that this procedure, while requiring some, possibly subjective, decisions on our part maintains objectivity for the purposes outlined above.

5.1. Replication and Extension Results

Appendix A provides brief summaries of each paper, highlighting the IV strategy used and main findings. Appendix Table A.1 displays the results of our replication and extension exercise. Despite limiting the analysis to one estimate per y-x-z pair, the table contains 148 separate examples relying on 54 unique first stages.[20] The 148 estimates are based on 105 outcome variables, 24 endogenous regressors, and 18 instruments across the thirteen papers. Rather than discuss each paper in depth, we choose to provide summary measures of the test results and to discuss a few cases that highlight a range of interesting results and key considerations for applying the quadratic overidentification test.

5.1.1. Overidentification Test Rejects

The motivating point of this exercise is to use the overidentification test to find cases in which the quadratic-in-z first stage gives statistically different estimates of the effect of interest. We see several cases of this across papers. The following papers have at least one specification in which the p-value for the overidentification test is less than 0.10: Alesina and Zhuravskaya (2011), Ananat (2011), Becker, Hornung, and Woessmann (2011), Brown and Laschever (2012), Chou et al. (2010), Dinkelman (2011), Lipscomb, Mobarak, and Barham (2013), and Werker, Ahmed, and Cohen (2009). Table 5.1 summarizes the results across the thirteen papers. Of the 148 separate estimates, there were 29 rejections of the null for the overidentification test at the 10% level. Of these 29 rejections, 18 occurred in conjunction with the residualized squared instrument being statistically significant in the first stage at the 10% level. These 18 cases are found across four separate papers and represent the strongest cases for using the quadratic first stage. That 11 of the overidentification rejections are not associated with a statistically significant residualized squared instrument undermines the results in these cases. However, overidentification rejection still implies that the introduction of the squared instrument does alter the first stage enough to change the estimate of the coefficient of interest in a statistically significant way and warrants additional attention.[21]

Note that the total number of possible tests of z^2 is equal to the number of unique first stages. That is, in 24 of 54 cases we find evidence that the additional instrument helps

[20] The number of unique first stages in a paper will depend on the number of x-z pairs, as well as the number of different samples in which that pair was used. The samples may differ for economic reasons, for instance running separate regressions on Black and White subsamples, or due to differential missing data when considering different outcome variables.

[21] Note that these 11 cases stem from only 7 unique first stages and the list is based on the strict cut-off at the 10% level. Interestingly, the overidentification rejections not associated with a statistically significant z^2 do not seem to be due purely to adding a weak instrument as the tests also reject when using the LIML estimates.

in predicting the endogenous regressor.[22] Of these 24 cases, eleven are associated with at least one overidentification rejection. The 24 unique first stages come from seven of the thirteen papers. In turn, four of the seven papers have a statistically significant squared term associated with an overidentification rejection. Importantly, this summary does not take into account that many of the outcome variables may be highly correlated with each other, yielding similar overidentification test results.

Table 5.1

Summary of Test Results by Paper

						Test Rejections 10%		
	Total	**First**	**Variables**					*Overid*
Paper	**Estimates**	**Stages**	Y	X	Z	*Overid*	Z^2	*&* Z^2
Acemoglu et al.	2	2	1	1	2	0	1	0
Alesina et al.	18	3	6	3	3	1	1	0
Ananat	24	3	24	1	1	2	2	1
Becker et al.	8	2	4	2	1	3	1	3
Brown et al.	6	6	1	6	1	1	3	0
Chodorow-Reich et al.	4	2	4	1	1	0	0	0
Chou et al.	8	8	4	2	1	1	0	0
Collins et al.	10	4	10	1	1	0	0	0
Dinkelman	13	5	13	1	1	3	5	3
Hunt et al.	2	2	1	2	2	0	0	0
Lipscomb et al.	20	4	20	1	1	5	0	0
Saiz et al.	1	1	1	1	1	0	1	0
Werker et al.	32	12	16	2	2	13	10	11
Total	148	54	105	24	18	29	24	18

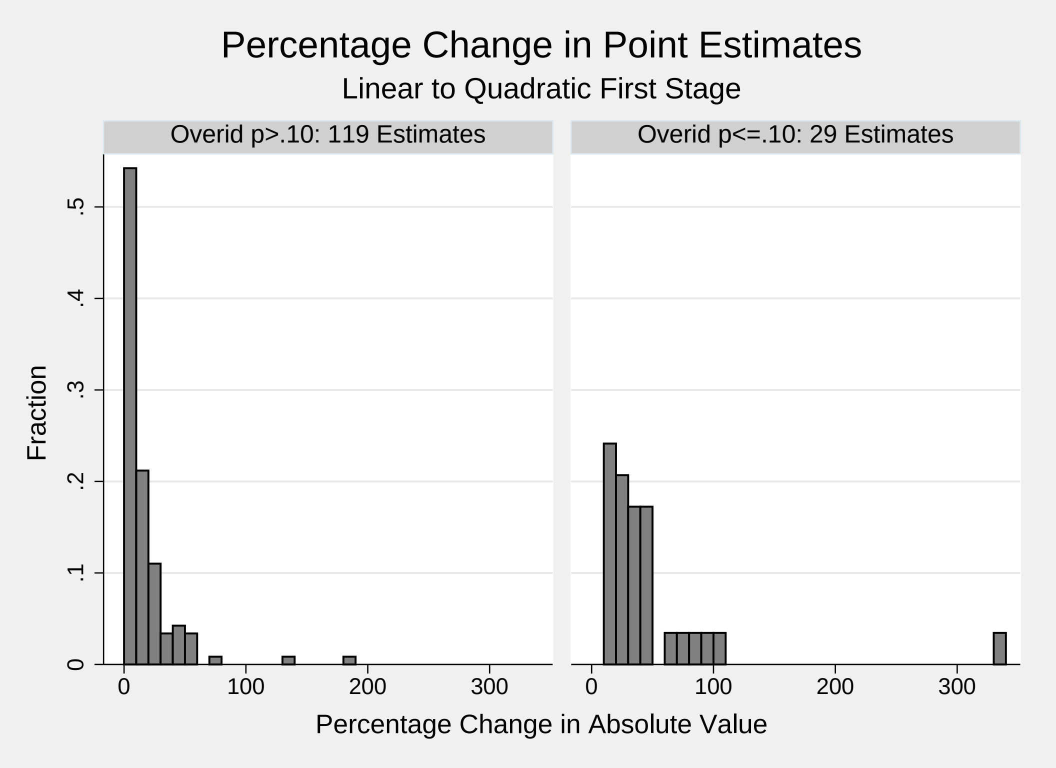
Test results compiled from replication and extension exercise found in Appendix Table A.1.

Allowing for heterogeneous effects, the importance of these rejections depends on the change in the magnitudes of the point estimates. There are several cases where the difference seems important, such as in Werker, Ahmed, and Cohen (2009), where the implied marginal effect of foreign aid as a percent of GDP on Non-capital Imports increases by nearly 0.3 percent of GDP. To provide a sense of how important the difference in point estimates is, we calculate the absolute value of the percentage change in the estimate going from the linear to quadratic first stage. Figure 5.1 displays the distribution of these percentage changes

[22] Obviously, it may be possible for higher order terms of the instrument to be statistically significant even when the squared term is not. We restrict focus here on the square to provide a uniform analysis across cases. Importantly, the squared term tends to do well empirically at picking up nonlinearities even if higher order terms may improve the fit.

separately for cases where we do not (left hand figure) or do (right hand figure) reject the null in the overidentification test. In each case, we present the fraction of estimates that fall within ten percentage point bins.

Figure 5.1: Change in Estimates by Overidentification Rejection



Absolute value of the percentage change in estimates going from a linear to quadratic first stage for the thirteen papers presented in Table 5.1. The histograms depict the fraction of estimates falling in ten percentage point bins separately for cases where the overidentification test rejects at the ten percent level and does not reject.

Starting with the cases where the overidentification test rejects, displayed in the right hand side of the figure, we see that in every case the change in the estimate is greater than ten percent. In contrast, nearly sixty percent of the cases that do not reject the overidentification test are associated with changes in the point estimate of less than ten percent. We also see several cases where the percentage change is substantial, well over fifty percent, when the overidentification test rejects the null. It is worth noting that we also observe a few cases of quite large changes in the estimates (over one-hundred percent) even when the overidentification test does not reject the null. Such cases likely deserve additional consideration, as the qualitative conclusions are not robust to the first stage specification. In sum, the differences in estimates uncovered here due to a simple change in how the first stage is specified may prove to be important more generally.

5.1.2. Significant z^2 and Overidentification Test Does not Reject

It is equally interesting when a significant squared instrument is not associated with overidentification rejection as this provides additional support for the linear-homogeneous effects model estimated under mean independence. In addition, the inclusion of the significant squared term may improve the precision of the 2SLS estimates. For example, in Ananat (2011) the standard error for the estimate of β falls when looking at the "Median Rent for Whites" outcome. This results in the estimate passing the threshold from a 5% significance level to 1%, while the estimate itself stays relatively unchanged. Perhaps more importantly, the 95% confidence interval shrinks from [-1170.948, -101.9589] to [-979.3989, -277.4901]. Such a gain in precision is nontrivial when considering the conclusions that can be drawn from the analysis.

5.1.3. Weak Instruments

It is important to consider whether we have introduced or exacerbated a weak instruments problem by adding the squared instrument. If we consider the results for the Dinkelman (2011) paper, we see that for the outcome "Change in Household Electrical Use" the estimate falls from 0.6350*** (0.2256) to 0.3576** (0.1408). However, we might be concerned with weak instruments given the first stage F-stat of 6.02 in the quadratic-in-z case. A quick glance at the LIML estimate confirms this concern, with $\hat{\beta}_{Q,LIML} = 0.6724$. Comparing all three point estimates, we see that the LIML estimate is quite close to the original linear-first-stage 2SLS estimate.

Importantly, it is not always the case that a low quadratic first stage F-stat is associated with large differences between 2SLS and LIML. For instance, the Brown and Laschever (2012) estimate for the effect of current year peer retirement on retirement decisions is 0.0219*** (0.0051) when estimated by 2SLS and 0.0220*** (0.0051) when estimated by LIML, despite a quadratic first stage F-stat of only 5.29.

Furthermore, even cases with a relatively strong first stage F-stat, may exhibit meaningful differences in the point estimates between 2SLS and LIML. For instance, several of the specifications in Lipscomb, Mobarak, and Barham (2013) are sensitive to the the

2SLS/LIML choice despite a relatively strong F-stat of 41.77. It is important to note that the F-stat is sensitive to the choice of standard error estimates.

5.1.4. Review Summary

While certainly not exhaustive, the above literature review illustrates the potential for considering alternative first stage specifications as an additional sensitivity analysis. The fact that eight out of thirteen papers had at least one rejection suggests that the application of the quadratic overidentification test may be fruitful more generally. As we have argued previously, such cases require additional caution and care in justifying the validity of the instrument. When a strong case for validity can be made, a heterogeneous effects analysis like that in section 4 can provide a more nuanced understanding of the economic relationship being studied.

6. Conclusion

We have explored the use of quadratic first stages to generate overidentifying restrictions when using continuous instruments in order to test the sensitivity of IV results to the choice of first stage. In applying this test to fifteen separate papers, we find many cases in which the overidentification test is suggestive of both statistically and economically meaningful differences in the estimated coefficients. We then show how to characterize the difference between the two estimates by the ratio of weights applied to average partial effects at different values of the instrument. Furthermore, these weight ratios are shown to be generally estimable without imposing additional assumptions. The estimated weight ratios can then be combined with the point estimates to provide additional insight into the economic relationship of interest.

Ultimately, how one should interpret the rejection is somewhat case specific. Regardless of approach, finding that higher order terms alter the conclusions drawn is both statistically and economically interesting. Either additional caution or justification for the instrument is needed or there are interesting heterogeneous effects to explore. While a failure to reject the null with the overidentification test is both encouraging, as it points to a potentially valid instrument, and intriguing, as it suggests generally homogeneous effects, it is not foolproof. Indeed, overidentification tests may fail to reject the null due to low power or if the two estimates have similar biases. In this case, the test may fail to reject even with an invalid instrument if the underlying relationship between x and z is approximated well by a linear specification, leaving little room for the higher order terms to predict x. Despite these caveats, the incredibly low cost of implementing our approach coupled with the potential benefits outlined here to further economic analysis make it appealing as a common sensitivity check to be undertaken by researchers using 2SLS.

Acemoglu, Daron, Simon Johnson, and James A Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *The American Economic Review* 91 (5):1369–1401.

———. 2012. "The colonial origins of comparative development: An empirical investigation: Reply." *The American Economic Review* 102 (6):3077–3110.

Acemoglu, Daron, Simon Johnson, James A Robinson, and Pierre Yared. 2008. "Income and Democracy." *American Economic Review* 98 (3):808–842.

Albouy, David Y. 2012. "The colonial origins of comparative development: an empirical investigation: comment." *The American Economic Review* 102 (6):3059–3076.

Alesina, Alberto and Ekaterina Zhuravskaya. 2011. "Segregation and the Quality of Government in a Cross-Section of Countries." *American Economic Review* 101 (5):1872–1911.

Ananat, Elizabeth Oltmans. 2011. "The wrong side (s) of the tracks: The causal effects of racial segregation on urban poverty and inequality." *American Economic Journal: Applied Economics* :34–66.

Angrist, Joshua and Jörn–Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricists' Companion.* Princeton, NJ: Princeton Univerisity Press.

Angrist, Joshua D, Kathryn Graddy, and Guido W Imbens. 2000. "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish." *The Review of Economic Studies* 67 (3):499–527.

Becker, Sascha O, Erik Hornung, and Ludger Woessmann. 2011. "Education and catch-up in the industrial revolution." *American Economic Journal: Macroeconomics* :92–126.

Becker, Sascha O and Ludger Woessmann. 2009. "Was Weber wrong? A human capital theory of Protestant economic history." *The Quarterly Journal of Economics* :531–596.

Brown, Kristine M and Ron A Laschever. 2012. "When They're Sixty-Four: Peer Effects and the Timing of Retirement." *American Economic Journal: Applied Economics* 4 (3):90–115.

Chodorow-Reich, Gabriel, Laura Feiveson, Zachary Liscow, and William Gui Woolston. 2012. "Does state fiscal relief during recessions increase employment? Evidence from the American Recovery and Reinvestment Act." *American Economic Journal: Economic Policy* :118–145.

Chou, Shin-Yi, Jin-Tan Liu, Michael Grossman, and Ted Joyce. 2010. "Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan." *American Economic Journal: Applied Economics* 2 (1):33–61.

Collins, William J, Katharine L Shester et al. 2013. "Slum Clearance and Urban Renewal in the United States." *American Economic Journal: Applied Economics* 5 (1):239–73.

Dinkelman, Taryn. 2011. "The effects of rural electrification on employment: New evidence from South Africa." *The American Economic Review* :3078–3108.

Hansen, Bruce E. 2009. "Lecture notes on nonparametrics." *Lecture notes* .

Heckman, James J, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The Review of Economics and Statistics* 88 (3):389–432.

Heckman, James J and Edward J Vytlacil. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the national Academy of Sciences* 96 (8):4730–4734.

Hunt, Jennifer and Marjolaine Gauthier-Loiselle. 2010. "How Much Does Immigration Boost Innovation?" *American Economic Journal: Macroeconomics* :31–56.

Imbens, Guido W and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2):467–475.

Lipscomb, Molly, Mushfiq A Mobarak, and Tania Barham. 2013. "Development effects of electrification: Evidence from the topographic placement of hydropower plants in Brazil." *American Economic Journal: Applied Economics* 5 (2):200–231.

Lochner, Lance and Enrico Moretti. 2011. "Estimating and testing non-linear models using instrumental variables." Tech. rep., CIBC Working Paper.

Løken, Katrine V, Magne Mogstad, and Matthew Wiswall. 2012. "What linear estimators miss: The effects of family income on child outcomes." *American Economic Journal: Applied Economics* 4 (2):1–35.

Moreira, Marcelo J. 2009. "Tests with correct size when instruments can be arbitrarily weak." *Journal of Econometrics* 152 (2):131–140.

Newey, Whitney K. 2013. "Nonparametric instrumental variables estimation." *The American Economic Review* 103 (3):550–556.

Saiz, Albert and Susan Wachter. 2011. "Immigration and the Neighborhood." *American Economic Journal: Economic Policy* :169–188.

Stock, James H, Jonathan H Wright, and Motohiro Yogo. 2002. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business & Economic Statistics* 20 (4).

Stock, James H and Motohiro Yogo. 2002. "Testing for weak instruments in linear IV regression."

Werker, Eric, Faisal Z Ahmed, and Charles Cohen. 2009. "How is foreign aid spent? Evidence from a natural experiment." *American Economic Journal: Macroeconomics* :225–244.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* MIT press.

Appendix A. AER and AEJ Replication Details

Appendix A.1. Brief Paper Summaries

We begin with a set of very brief summaries of each paper and then proceed to look at the results of our exercise. In each summary, we highlight the relevant IV strategy.

Appendix A.1.1. Acemoglu et al. (2008)

Acemoglu et al. explore the relationship between income levels and measures of democracy across countries. The pertinent IV strategy consists of regressing the Freedom House measure of democracy on the log of GDP per capita from five years prior. They consider two instruments: the savings rate from ten years prior and a measure of world income that has been weighted based on the trade patterns for a particular country. On the whole, they find little evidence of a causal link between income and democracy, despite the raw relationship between the two in the data.

Appendix A.1.2. Alesina and Zhuravskaya (2011)

Alesina & Zhuravskaya consider the potential effect of ethnic, religious, and linguistic segregation on the quality of government. The authors consider these segregation effects separately and examine several different outcomes. The instruments used for the three segregation measures are predicted segregation based on the composition of people's background in neighboring countries. They find evidence that countries that are more ethnically and linguistically segregated also tend to have lower quality government.

Appendix A.1.3. Ananat (2011)

Ananat looks for causal evidence of the negative link between racial segregation and the characteristics of the population. Ananat considers a range of economic outcomes related to education, migration, and income. The instrument for segregation is a segregation Herfindahl index derived from the alignment of train tracks in the 19th century. The author finds that segregation leads to higher rates of poverty for Black and larger Black-White income gaps, while lowering poverty rates for White and decreasing inequality within the White population

Appendix A.1.4. Becker, Hornung, and Woessmann (2011)

Becker et al. look at the role formal education may have played in the industrial revolution using historical data from Prussia. The authors regress several measures of factory employment (from 1849 or 1882) on contemporaneous measures of education (years of schooling or literacy rates), using school enrollment from 1816 as an instrument. They find that education was strongly related to industrialization.

Appendix A.1.5. Brown and Laschever (2012)

Brown & Laschever estimate the effect of peer retirement decisions on ones own choice to retire using administrative data on teachers from Los Angeles. The main empirical approach is to regress a retirement indicator on various peer retirement measures. The authors use the sum (across one's peers) of unexpected changes in pension wealth driven by reforms as an instrument for the peer retirement variables. They find evidence that peer retirement in the previous year increases the probability of retirement for an individual.

Appendix A.1.6. Chodorow-Reich et al. (2012)

In this paper, Chodorow-Reich et al. look for evidence that transfer payments made by the government during a recession have a positive impact on employment. Using transfers summing to $88 billion made by the US government to states in 2009, they regress employment outcomes for states on the per person payment associated with the transfer scheme. To address the possible endogeneity of the payment amounts to current economic conditions, the authors use prior Medicaid spending as an instrument for the transfer since this spending determined a portion of the transfer. They find evidence that these transfers did increase employment.

Appendix A.1.7. Chou et al. (2010)

Chou et al. focus on the connection between a child's health and the education of their parents in Taiwan. The authors regress measures of early child health on parent's years of schooling. They exploit the large scale building of over 150 new schools in 1968 to create an instrument for parental schooling based on the local intensity of the school expansion program experienced by the parent. Using this approach, the authors find evidence that parental schooling does have a positive effect on health outcomes of young children.

Appendix A.1.8. Collins, Shester et al. (2013)

This paper explores the effect of a program in the US that cleared poverty stricken urban areas to allow for redevelopment after World War II. The main estimation technique involves regressing local measures capturing labor market, housing, and population characteristics on the amount of funding from the urban renewal program. As an instrument for the funding variable, the authors use variation in the timing of State-level legislation allowing for agencies to acquire property for private development via eminent domain. By adopting this approach, the authors estimate positive effects of the program on the economic outcomes they consider.

Appendix A.1.9. Dinkelman (2011)

Dinkelman estimates the effect of electrification on employment in a developing context using data from South Africa. The author regresses measures of employment and home production on the availability of electricity due to a large scale electrification program. In order to address the possible endogeneity of the roll-out of electricity, Dinkelman uses a measure of the land gradient as an instrument exploiting the fact that expansion of the electrical grid is less costly on flatter terrain. Dinkelman finds evidence that the access to household electricity increased hours worked while increasing male wages but reducing female wages.

Appendix A.1.10. Hunt and Gauthier-Loiselle (2010)

Hunt & Gauthier-Loiselle consider the relationship between immigration and innovation in the US by regressing log patents per capita on the population share of skilled immigrants. To instrument for the skilled immigrant share, the authors predict the skilled immigrant share based on the historical distribution of immigrants across states from 1940. The authors find a large, positive effect of increasing the population share of skilled immigrants on patents.

Appendix A.1.11. Lipscomb, Mobarak, and Barham (2013)

Lipscomb et al. look at the effect of electrification on a host of economic outcomes in Brazil. The pertinent regressions involve regressing one of these economic measures on electric availability using predicted availability as an instrument. The predicted availability is based on differences in the infrastructure investment costs due to geographic concerns. This approach leads to positive estimates of electrification on economic outcomes.

Appendix A.1.12. Saiz and Wachter (2011)

Saiz & Wachter are interested in estimating the relationship between immigration and neighborhood-level economic outcomes. To do so, they regress a measure of neighborhood property value on the foreign born population using the predicted population based on a geographic diffusion “gravity pull” model as an instrument. The results suggest that increased immigration leads to slower increases in housing prices.

Appendix A.1.13. Werker, Ahmed, and Cohen (2009)

Werker et al. look at how foreign aid is spent by receiving countries by examining transfers from oil-rich OPEC countries to poorer Muslim countries. The authors consider a large number of trade, production, consumption, and price outcomes and focus on the effect of both current and lagged foreign aid as a percentage on GDP. As an instrument for foreign aid, they use oil prices interacted with a Muslim country indicator. Using this approach, they find a positive effect of aid on GDP but little effect on growth or prices.

Appendix A.2. Replication Tables

As a space-saving measure, we adopt the convention of referring to dual-authored papers with "et al." in the table. We identify each estimate by the Authors, the outcome (Y), the endogenous explanatory variable (X), and the instrument (Z). For each y-x-z pair, we report the coefficient estimates ($\hat{\beta}$) and standard errors (SE) for the linear-in-z 2SLS, quadratic-in-z 2SLS, and quadratic-in-z LIML estimators. We also report the first stage F-stat (F), the p-value from the overidentification test (Overid p), and the p-value from the t-test for the coefficient on the squared instrument (Z^2 p).

Table A.1

AER and AEJ Replications

			Linear	**Quadratic**	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
Acemoglu	*Y: Democracy Measure*	$\hat{\beta}$	-0.1196	-0.1238	-0.1247
et al. (2008)	*X: Lag Ln GDP/Cap*	SE	(0.0968)	(0.1017)	(0.1026)
	Z: Trade Wghtd World Inc	F	26.53	58.9	
		Overid p		0.4797	
		Z^2 p		0.7699	
	Y: Democracy Measure	$\hat{\beta}$	-0.0205	-0.0119	-0.012
	X: Lag Ln GDP/Cap	SE	(0.0743)	(0.0716)	(0.0717)
	Z: 2nd Lag Savings Rate	F	24.68	15.99	
		Overid p		0.6834	
		Z^2 p		0.0064	
Alesina	*Y: Control Corruption*	$\hat{\beta}$	-1.7725***	-1.7648***	-1.7650***
et al. (2011)	*X: Ethnic Segregation*	SE	(0.5911)	(0.6433)	(0.6433)
	Z: Predict X: Border Comp	F	15.59	8.17	
		Overid p		0.9428	
		Z^2 p		0.5518	
	Y: Control Corruption	$\hat{\beta}$	-1.2882	-0.9533	-0.9612
	X: Linguistic Segregation	SE	(0.8916)	(0.9089)	(0.9450)
	Z: Predict X: Border Comp	F	8.39	5	
		Overid p		0.3163	
		Z^2 p		0.0572	
	Y: Control Corruption	$\hat{\beta}$	-1.1084	-0.5165	-0.5514
	X: Religious Segregation	SE	(1.7828)	(1.5146)	(1.5704)
	Z: Predict X: Border Comp	F	16.08	9.77	
		Overid p		0.3375	
		Z^2 p		0.1588	
	Y: Gov Effectiveness	$\hat{\beta}$	-2.1435***	-2.1875***	-2.1962***
	X: Ethnic Segregation	SE	(0.5973)	(0.6621)	(0.6651)
	Z: Predict X: Border Comp	F	15.59	8.17	
		Overid p		0.676	
		Z^2 p		0.5518	

Table A.1

AER and AEJ Replications

			Linear	**Quadratic**	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Gov Effectiveness	$\hat{\beta}$	-1.4740*	-1.241	-1.2582
	X: Linguistic Segregation	SE	(0.7698)	(0.7847)	(0.7993)
	Z: Predict X: Border Comp	F	8.39	5	
		Overid p		0.3715	
		Z^2 p		0.0572	
	Y: Gov Effectiveness	$\hat{\beta}$	-1.1385	-0.5208	-0.5767
	X: Religious Segregation	SE	(1.8206)	(1.5319)	(1.6033)
	Z: Predict X: Border Comp	F	16.08	9.77	
		Overid p		0.3187	
		Z^2 p		0.1588	
	Y: Political Stability	$\hat{\beta}$	-3.6463***	-3.6662***	-3.6674***
	X: Ethnic Segregation	SE	(1.3067)	(1.2817)	(1.2824)
	Z: Predict X: Border Comp	F	15.59	8.17	
		Overid p		0.8919	
		Z^2 p		0.5518	
	Y: Political Stability	$\hat{\beta}$	-2.9193***	-3.1224***	-3.1428***
	X: Linguistic Segregation	SE	(0.7932)	(0.8758)	(0.8867)
	Z: Predict X: Border Comp	F	8.39	5	
		Overid p		0.5199	
		Z^2 p		0.0572	
	Y: Political Stability	$\hat{\beta}$	-2.1257	-3.0109	-3.1913
	X: Religious Segregation	SE	(2.1052)	(2.3062)	(2.4634)
	Z: Predict X: Border Comp	F	16.08	9.77	
		Overid p		0.0877	
		Z^2 p		0.1588	
	Y: Regulatory Quality	$\hat{\beta}$	-2.0962*	-2.2975**	-2.4452**
	X: Ethnic Segregation	SE	(1.1122)	(1.0959)	(1.1650)
	Z: Predict X: Border Comp	F	15.59	8.17	
		Overid p		0.1663	
		Z^2 p		0.5518	
	Y: Regulatory Quality	$\hat{\beta}$	-1.9511	-1.9745*	-1.9748*
	X: Linguistic Segregation	SE	(1.1878)	(1.1040)	(1.1042)
	Z: Predict X: Border Comp	F	8.39	5	
		Overid p		0.9427	
		Z^2 p		0.0572	
	Y: Regulatory Quality	$\hat{\beta}$	0.9667	1.1392	1.1403
	X: Religious Segregation	SE	(1.8284)	(1.7249)	(1.7303)
	Z: Predict X: Border Comp	F	16.08	9.77	
		Overid p		0.741	
		Z^2 p		0.1588	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Rule Law	$\hat{\beta}$	-2.4666***	-2.5561***	-2.5896***
	X: Ethnic Segregation	SE	(0.6145)	(0.6746)	(0.6888)
	Z: Predict X: Border Comp	F	15.59	8.17	
		Overid p		0.4133	
		Z^2 p		0.5518	
	Y: Rule Law	$\hat{\beta}$	-1.7969**	-1.7456**	-1.7464**
	X: Linguistic Segregation	SE	(0.7048)	(0.7472)	(0.7479)
	Z: Predict X: Border Comp	F	8.39	5	
		Overid p		0.8134	
		Z^2 p		0.0572	
	Y: Rule Law	$\hat{\beta}$	-0.8699	-0.5569	-0.5669
	X: Religious Segregation	SE	(1.7784)	(1.6053)	(1.6235)
	Z: Predict X: Border Comp	F	16.08	9.77	
		Overid p		0.5756	
		Z^2 p		0.1588	
	Y: Voice & Accountability	$\hat{\beta}$	-1.2805	-1.258	-1.2581
	X: Ethnic Segregation	SE	(0.9487)	(0.9170)	(0.9180)
	Z: Predict X: Border Comp	F	15.59	8.17	
		Overid p		0.8507	
		Z^2 p		0.5518	
	Y: Voice & Accountability	$\hat{\beta}$	-2.6487***	-2.3073***	-2.3588***
	X: Linguistic Segregation	SE	(0.8268)	(0.8144)	(0.8518)
	Z: Predict X: Border Comp	F	8.39	5	
		Overid p		0.2559	
		Z^2 p		0.0572	
	Y: Voice & Accountability	$\hat{\beta}$	0.5506	0.7959	0.7989
	X: Religious Segregation	SE	(1.8802)	(1.6262)	(1.6371)
	Z: Predict X: Border Comp	F	16.08	9.77	
		Overid p		0.6862	
		Z^2 p		0.1588	
Ananat (2011)	*Y: College Grads: Blk*	$\hat{\beta}$	-0.2969	-0.2997	-0.2997
	X: Segregation	SE	(0.2153)	(0.2040)	(0.2040)
	Z: Segregation Herf	F	13.18	11.7	
		Overid p		0.9626	
		Z^2 p		0.2164	
	Y: College Grads: Wht	$\hat{\beta}$	-0.1403	-0.2131	-0.2158
	X: Segregation	SE	(0.1496)	(0.1568)	(0.1611)
	Z: Segregation Herf	F	15.07	12.58	
		Overid p		0.2463	
		Z^2 p		0.0974	

Table A.1

AER and AEJ Replications

			Linear	**Quadratic**	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: HS Dropouts: Blk	$\hat{\beta}$	0.4310**	0.3262*	0.3259*
	X: Segregation	SE	(0.1979)	(0.1932)	(0.1966)
	Z: Segregation Herf	F	13.18	11.7	
		Overid p		0.0879	
		Z^2 p		0.2164	
	Y: HS Dropouts: Wht	$\hat{\beta}$	-0.144	-0.0852	-0.0919
	X: Segregation	SE	(0.1466)	(0.1064)	(0.1115)
	Z: Segregation Herf	F	15.07	12.58	
		Overid p		0.3408	
		Z^2 p		0.0974	
	Y: HS Grads: Blk	$\hat{\beta}$	0.6520**	0.5930*	0.5945*
	X: Segregation	SE	(0.3281)	(0.3350)	(0.3363)
	Z: Segregation Herf	F	13.18	11.7	
		Overid p		0.5181	
		Z^2 p		0.2164	
	Y: HS Grads: Wht	$\hat{\beta}$	0.4580***	0.4445***	0.4447***
	X: Segregation	SE	(0.1719)	(0.1347)	(0.1349)
	Z: Segregation Herf	F	15.07	12.58	
		Overid p		0.8667	
		Z^2 p		0.0974	
	Y: In-migrants: Blk	$\hat{\beta}$	-0.2705**	-0.2835***	-0.2835***
	X: Segregation	SE	(0.1132)	(0.0794)	(0.0795)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.8567	
		Z^2 p		0.0000	
	Y: In-migrants: Wht	$\hat{\beta}$	-0.1550**	-0.1855***	-0.1863***
	X: Segregation	SE	(0.0719)	(0.0545)	(0.0553)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.4657	
		Z^2 p		0.0000	
	Y: Inequality: 10 Wht/10 Blk	$\hat{\beta}$	2.7269***	2.2749***	2.3177***
	X: Segregation	SE	(0.8563)	(0.5658)	(0.5892)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.4149	
		Z^2 p		0.0000	
	Y: Inequality: 90 Blk/10 Wht	$\hat{\beta}$	-0.8073**	-0.6824**	-0.6887**
	X: Segregation	SE	(0.3796)	(0.2831)	(0.2868)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.5558	
		Z^2 p		0.0000	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Inequality: 90 Wht/10 Blk	$\hat{\beta}$	1.7887**	1.7048***	1.7055***
	X: Segregation	SE	(0.7487)	(0.5418)	(0.5424)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.8609	
		Z^2 p		0.0000	
	Y: Inequality: 90 Wht/90 Blk	$\hat{\beta}$	-0.1308	0.1123	0.1115
	X: Segregation	SE	(0.3082)	(0.1878)	(0.2036)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.2955	
		Z^2 p		0.0000	
	Y: Ln Gini: Blk	$\hat{\beta}$	0.8751**	0.6057***	0.6256***
	X: Segregation	SE	(0.4040)	(0.1614)	(0.1787)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.4774	
		Z^2 p		0.0000	
	Y: Ln Gini: Wht	$\hat{\beta}$	-0.3345***	-0.2263***	-0.2519***
	X: Segregation	SE	(0.0980)	(0.0785)	(0.0929)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.0118	
		Z^2 p		0.0000	
	Y: Median Rent % Inc: Blk	$\hat{\beta}$	-3.4159	-4.8608	-4.8686
	X: Segregation	SE	(5.3199)	(4.7785)	(4.8003)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.5883	
		Z^2 p		0.0000	
	Y: Median Rent % Inc: Wht	$\hat{\beta}$	-16.6657***	-14.6675***	-14.8829***
	X: Segregation	SE	(3.5978)	(3.1242)	(3.2410)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.2132	
		Z^2 p		0.0000	
	Y: Median Rent: Blk	$\hat{\beta}$	-623.6425***	-624.7007***	-624.7015***
	X: Segregation	SE	(155.0110)	(140.3437)	(140.3441)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.9887	
		Z^2 p		0.0000	
	Y: Median Rent: Wht	$\hat{\beta}$	-636.4534**	-628.4445***	-628.4822***
	X: Segregation	SE	(272.7063)	(179.0617)	(179.0827)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.9639	
		Z^2 p		0.0000	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Poverty Rate: Blk	$\hat{\beta}$	0.2584**	0.2911***	0.2917***
	X: Segregation	SE	(0.1069)	(0.0973)	(0.0978)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.6347	
		Z^2 p		0.0000	
	Y: Poverty Rate: Wht	$\hat{\beta}$	-0.1957***	-0.1694***	-0.1715***
	X: Segregation	SE	(0.0640)	(0.0483)	(0.0493)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.4345	
		Z^2 p		0.0000	
	Y: HH w/ >1 Person/Room: Blk	$\hat{\beta}$	-0.1650***	-0.1529***	-0.1532***
	X: Segregation	SE	(0.0463)	(0.0419)	(0.0421)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.5934	
		Z^2 p		0.0000	
	Y: HH w/ >1 Person/Room: Wht	$\hat{\beta}$	-0.1161***	-0.1025***	-0.1030***
	X: Segregation	SE	(0.0365)	(0.0283)	(0.0286)
	Z: Segregation Herf	F	16.57	20.47	
		Overid p		0.5106	
		Z^2 p		0.0000	
	Y: Some College: Blk	$\hat{\beta}$	-0.7862**	-0.6196**	-0.6330**
	X: Segregation	SE	(0.3159)	(0.2879)	(0.3000)
	Z: Segregation Herf	F	13.18	11.7	
		Overid p		0.1368	
		Z^2 p		0.2164	
	Y: Some College: Wht	$\hat{\beta}$	-0.1737	-0.1462	-0.1461
	X: Segregation	SE	(0.1087)	(0.0935)	(0.0940)
	Z: Segregation Herf	F	15.07	12.58	
		Overid p		0.4992	
		Z^2 p		0.0974	
Becker	*Y: Factory Employ 1849: All*	$\hat{\beta}$	0.1317*	0.1316*	0.1316*
et al. (2011)	*X: Yrs Schlng 1849*	SE	(0.0767)	(0.0769)	(0.0769)
	Z: Sch Enroll 1816	F	6206.97	3132.35	
		Overid p		0.9416	
		Z^2 p		0.2617	
	Y: Factory Employ 1849: Oth	$\hat{\beta}$	0.1351***	0.1361***	0.1360***
	X: Yrs Schlng 1849	SE	(0.0436)	(0.0441)	(0.0441)
	Z: Sch Enroll 1816	F	6206.97	3132.35	
		Overid p		0.3047	
		Z^2 p		0.2617	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Factory Employ 1849: Metal	$\hat{\beta}$	0.0447	0.0437	0.0437
	X: Yrs Sching 1849	SE	(0.0465)	(0.0462)	(0.0462)
	Z: Sch Enroll 1816	F	6206.97	3132.35	
		Overid p		0.1758	
		Z^2 p		0.2617	
	Y: Factory Employ 1849: Textile	$\hat{\beta}$	-0.0481	-0.0482	-0.0482
	X: Yrs Sching 1849	SE	(0.0334)	(0.0335)	(0.0335)
	Z: Sch Enroll 1816	F	6206.97	3132.35	
		Overid p		0.8819	
		Z^2 p		0.2617	
	Y: Factory Employ 1882: All	$\hat{\beta}$	0.1360***	0.1183***	0.1160***
	X: Literacy 1871	SE	(0.0357)	(0.0333)	(0.0339)
	Z: Sch Enroll 1816	F	69.85	37.26	
		Overid p		0.0865	
		Z^2 p		0.0043	
	Y: Factory Employ 1882: Oth	$\hat{\beta}$	0.0689***	0.0576***	0.0568***
	X: Literacy 1871	SE	(0.0127)	(0.0113)	(0.0122)
	Z: Sch Enroll 1816	F	69.85	37.26	
		Overid p		0.0029	
		Z^2 p		0.0043	
	Y: Factory Employ 1882: Metal	$\hat{\beta}$	0.0930***	0.0807***	0.0795***
	X: Literacy 1871	SE	(0.0248)	(0.0233)	(0.0237)
	Z: Sch Enroll 1816	F	69.85	37.26	
		Overid p		0.0327	
		Z^2 p		0.0043	
	Y: Factory Employ 1882: Textile	$\hat{\beta}$	-0.0259	-0.02	-0.0202
	X: Literacy 1871	SE	(0.0247)	(0.0227)	(0.0228)
	Z: Sch Enroll 1816	F	69.85	37.26	
		Overid p		0.3041	
		Z^2 p		0.0043	
Brown	*Y: Retirement*	$\hat{\beta}$	0.1823**	0.1795**	0.1796**
et al. (2012)	*X: Rate Retirement Last Yr*	SE	(0.0913)	(0.0909)	(0.0910)
	Z: Unexpected Pension Δ	F	81.85	45.74	
		Overid p		0.4951	
		Z^2 p		0.7676	
	Y: Retirement	$\hat{\beta}$	0.0098***	0.0096***	0.0096***
	X: Peer RetireesLast 2 Yr	SE	(0.0026)	(0.0026)	(0.0026)
	Z: Unexpected Pension Δ	F	34.19	26.41	
		Overid p		0.7208	
		Z^2 p		0.399	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Retirement	$\hat{\beta}$	0.0271***	0.0219***	0.0220***
	X: Peer RetireesThis Yr	SE	(0.0063)	(0.0051)	(0.0051)
	Z: Unexpected Pension Δ	F	7.2	5.29	
		Overid p		0.0593	
		Z^2 p		0.43	
	Y: Retirement	$\hat{\beta}$	0.0151***	0.0122***	0.0122***
	X: Peer RetireesSpr/Sum	SE	(0.0057)	(0.0045)	(0.0045)
	Z: Unexpected Pension Δ	F	31.31	34.56	
		Overid p		0.4121	
		Z^2 p		0.0001	
	Y: Retirement	$\hat{\beta}$	0.0191***	0.0137***	0.0138***
	X: Peer RetireesLast Sum	SE	(0.0072)	(0.0051)	(0.0052)
	Z: Unexpected Pension Δ	F	24.23	29.67	
		Overid p		0.3133	
		Z^2 p		0.0003	
	Y: Retirement	$\hat{\beta}$	0.0153***	0.0124***	0.0124***
	X: Peer RetireesLast Yr	SE	(0.0058)	(0.0045)	(0.0045)
	Z: Unexpected Pension Δ	F	28.59	32.49	
		Overid p		0.4106	
		Z^2 p		0.0001	
Chodorow-Reich	*Y: Rainy Day Fund 2009*	$\hat{\beta}$	0.0122	0.0107	0.0107
et al. (2012)	*X: FMAP Payout/Person*	SE	(0.2251)	(0.2276)	(0.2276)
	Z: Prior Medicaid Spending	F	96.67	56.05	
		Overid p		0.9155	
		Z^2 p		0.3718	
	Y: Rainy Day Fund 2010	$\hat{\beta}$	0.0835	0.0659	0.0682
	X: FMAP Payout/Person	SE	(0.1787)	(0.1744)	(0.1765)
	Z: Prior Medicaid Spending	F	96.67	56.05	
		Overid p		0.2313	
		Z^2 p		0.3718	
	Y: Employ: Gov, Edu, Hlth	$\hat{\beta}$	1.1949***	1.1528***	1.1743***
	X: FMAP Payout/Person	SE	(0.3728)	(0.3858)	(0.3879)
	Z: Prior Medicaid Spending	F	120.28	54.02	
		Overid p		0.1752	
		Z^2 p		0.2466	
	Y: Employ: Nonfarm	$\hat{\beta}$	4.6136***	4.5040***	4.5379***
	X: FMAP Payout/Person	SE	(1.5664)	(1.6447)	(1.6477)
	Z: Prior Medicaid Spending	F	120.28	54.02	
		Overid p		0.4275	
		Z^2 p		0.2466	

Table A.1

AER and AEJ Replications

			Linear	**Quadratic**	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
Chou	*Y: Infant Mortality*	$\hat{\beta}$	-0.4288*	-0.4192*	-0.4209*
et al. (2010)	*X: Father's Sching*	SE	(0.2387)	(0.2332)	(0.2409)
	Z: New Sch Build Intensity	F	5.6	3.09	
		Overid p		0.3091	
		Z^2 p		0.8389	
	Y: Infant Mortality	$\hat{\beta}$	-0.5050***	-0.5156***	-0.5174***
	X: Mother's Sching	SE	(0.1948)	(0.1940)	(0.1966)
	Z: New Sch Build Intensity	F	13.8	7.18	
		Overid p		0.4167	
		Z^2 p		0.6041	
	Y: Low Birth Weight	$\hat{\beta}$	-0.1890***	-0.1916***	-0.1914***
	X: Father's Sching	SE	(0.0730)	(0.0678)	(0.0679)
	Z: New Sch Build Intensity	F	8.58	3.98	
		Overid p		0.7021	
		Z^2 p		0.7464	
	Y: Low Birth Weight	$\hat{\beta}$	-0.1936**	-0.1938**	-0.1938**
	X: Mother's Sching	SE	(0.0775)	(0.0753)	(0.0753)
	Z: New Sch Build Intensity	F	12.72	9.19	
		Overid p		0.9691	
		Z^2 p		0.7	
	Y: Neonatal Mortality	$\hat{\beta}$	-0.2434	-0.1829	-0.1577
	X: Father's Sching	SE	(0.2615)	(0.2358)	(0.2558)
	Z: New Sch Build Intensity	F	2.76	1.68	
		Overid p		0.0753	
		Z^2 p		0.6333	
	Y: Neonatal Mortality	$\hat{\beta}$	-0.3183**	-0.3292**	-0.3283**
	X: Mother's Sching	SE	(0.1496)	(0.1464)	(0.1478)
	Z: New Sch Build Intensity	F	10.16	5.6	
		Overid p		0.4	
		Z^2 p		0.6416	
	Y: Post Neonatal Mortality	$\hat{\beta}$	-0.5837*	-0.5508*	-0.5560*
	X: Father's Sching	SE	(0.3400)	(0.3121)	(0.3275)
	Z: New Sch Build Intensity	F	5.17	3.31	
		Overid p		0.3816	
		Z^2 p		0.5926	
	Y: Post Neonatal Mortality	$\hat{\beta}$	-0.6412***	-0.6505***	-0.6551***
	X: Mother's Sching	SE	(0.2439)	(0.2443)	(0.2500)
	Z: New Sch Build Intensity	F	13.75	7.42	
		Overid p		0.3454	
		Z^2 p		0.7271	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
Collins	*Y: Employ Rate*	$\hat{\beta}$	0.0034*	0.0033	0.0033
et al. (2013)	*X: Urban Renewal Funding*	SE	(0.0020)	(0.0022)	(0.0022)
	Z: Yrs Since Legislation	F	13.79	7.85	
		Overid p		0.9579	
		Z^2 p		0.2165	
	Y: Ln Housing Units	$\hat{\beta}$	0.0011**	0.0011**	0.0011**
	X: Urban Renewal Funding	SE	(0.0005)	(0.0005)	(0.0005)
	Z: Yrs Since Legislation	F	13.21	7.37	
		Overid p		0.8645	
		Z^2 p		0.1918	
	Y: Ln Median Family Inc	$\hat{\beta}$	0.0002**	0.0002**	0.0002**
	X: Urban Renewal Funding	SE	(0.0001)	(0.0001)	(0.0001)
	Z: Yrs Since Legislation	F	13.79	7.85	
		Overid p		0.8033	
		Z^2 p		0.2165	
	Y: Ln Median Property Value	$\hat{\beta}$	0.0007**	0.0009***	0.0011**
	X: Urban Renewal Funding	SE	(0.0003)	(0.0003)	(0.0004)
	Z: Yrs Since Legislation	F	14.02	7.95	
		Overid p		0.1863	
		Z^2 p		0.2192	
	Y: Ln Pop	$\hat{\beta}$	0.0009*	0.0010*	0.0010*
	X: Urban Renewal Funding	SE	(0.0005)	(0.0005)	(0.0005)
	Z: Yrs Since Legislation	F	12.88	7.19	
		Overid p		0.8173	
		Z^2 p		0.1878	
	Y: Median Sching	$\hat{\beta}$	0.0000	0.0003	0.0003
	X: Urban Renewal Funding	SE	(0.0004)	(0.0004)	(0.0005)
	Z: Yrs Since Legislation	F	13.79	7.85	
		Overid p		0.2357	
		Z^2 p		0.2165	
	Y: % Blk	$\hat{\beta}$	0.0112	0.0091	0.0091
	X: Urban Renewal Funding	SE	(0.0100)	(0.0096)	(0.0097)
	Z: Yrs Since Legislation	F	13.79	7.85	
		Overid p		0.4478	
		Z^2 p		0.2165	
	Y: % No Plumbing	$\hat{\beta}$	-0.0017	-0.0013	-0.0015
	X: Urban Renewal Funding	SE	(0.0012)	(0.0012)	(0.0013)
	Z: Yrs Since Legislation	F	13.79	7.85	
		Overid p		0.3278	
		Z^2 p		0.2165	

Table A.1

AER and AEJ Replications

			Linear	**Quadratic**	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: % Old Units	$\hat{\beta}$	-0.0330**	-0.0279**	-0.0301**
	X: Urban Renewal Funding	SE	(0.0130)	(0.0131)	(0.0147)
	Z: Yrs Since Legislation	F	13.79	7.85	
		Overid p		0.1097	
		Z^2 p		0.2165	
	Y: Poverty Rate	$\hat{\beta}$	-0.0061	-0.005	-0.0052
	X: Urban Renewal Funding	SE	(0.0051)	(0.0052)	(0.0054)
	Z: Yrs Since Legislation	F	13.79	7.85	
		Overid p		0.2575	
		Z^2 p		0.2165	
Dinkelman (2011)	*Y: Δ Female Employ: Nonmigrant*	$\hat{\beta}$	0.1157*	0.0761	0.0899
	X: Electrification Program	SE	(0.0682)	(0.0483)	(0.0591)
	Z: Land Gradient	F	8.26	6.02	
		Overid p		0.168	
		Z^2 p		0.0995	
	Y: Δ Male Employ: Nonmigrant	$\hat{\beta}$	0.086	0.0245	0.0355
	X: Electrification Program	SE	(0.0685)	(0.0548)	(0.0736)
	Z: Land Gradient	F	8.26	6.02	
		Overid p		0.0423	
		Z^2 p		0.0995	
	Y: Δ Female Employ	$\hat{\beta}$	0.0951*	0.057	0.0664
	X: Electrification Program	SE	(0.0548)	(0.0424)	(0.0506)
	Z: Land Gradient	F	8.26	6.02	
		Overid p		0.1608	
		Z^2 p		0.0995	
	Y: Δ Flush Toilets	$\hat{\beta}$	0.067	0.0694	0.0695
	X: Electrification Program	SE	(0.0670)	(0.0599)	(0.0599)
	Z: Land Gradient	F	8.34	6.06	
		Overid p		0.9331	
		Z^2 p		0.098	
	Y: Δ HH Cook w/ Electric	$\hat{\beta}$	0.2275**	0.1279**	0.1868
	X: Electrification Program	SE	(0.1003)	(0.0646)	(0.1341)
	Z: Land Gradient	F	8.26	6.02	
		Overid p		0.0097	
		Z^2 p		0.0995	
	Y: Δ HH Cook w/ Wood	$\hat{\beta}$	-0.2754*	-0.2156**	-0.2265**
	X: Electrification Program	SE	(0.1457)	(0.1026)	(0.1106)
	Z: Land Gradient	F	8.6	6.24	
		Overid p		0.3536	
		Z^2 p		0.0976	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Δ HH Electric	$\hat{\beta}$	0.6350***	0.3576**	0.6724
	X: Electrification Program	SE	(0.2256)	(0.1408)	(0.5999)
	Z: Land Gradient	F	8.26	6.02	
		Overid p		0.0075	
		Z^2 p		0.0995	
	Y: Δ HS Matric: Female	$\hat{\beta}$	0.1297**	0.1028**	0.1093**
	X: Electrification Program	SE	(0.0577)	(0.0492)	(0.0535)
	Z: Land Gradient	F	9.77	6.99	
		Overid p		0.2707	
		Z^2 p		0.0729	
	Y: Δ HS Matric: Male	$\hat{\beta}$	0.0767	0.0477	0.0526
	X: Electrification Program	SE	(0.0503)	(0.0416)	(0.0466)
	Z: Land Gradient	F	9.77	6.99	
		Overid p		0.1349	
		Z^2 p		0.0729	
	Y: Δ Male Employ	$\hat{\beta}$	0.0355	-0.0118	-0.0121
	X: Electrification Program	SE	(0.0654)	(0.0574)	(0.0668)
	Z: Land Gradient	F	8.26	6.02	
		Overid p		0.1677	
		Z^2 p		0.0995	
	Y: Δ Water Close	$\hat{\beta}$	-0.3722	-0.3626*	-0.3628*
	X: Electrification Program	SE	(0.2466)	(0.1930)	(0.1931)
	Z: Land Gradient	F	8.34	6.06	
		Overid p		0.943	
		Z^2 p		0.098	
	Y: Ln Non-inmigrant Pop	$\hat{\beta}$	4.3489***	3.4156***	4.1645***
	X: Electrification Program	SE	(1.5732)	(1.0381)	(1.5171)
	Z: Land Gradient	F	8.26	6.02	
		Overid p		0.1239	
		Z^2 p		0.0995	
	Y: Ln Pop	$\hat{\beta}$	3.8970***	3.1736***	3.6460***
	X: Electrification Program	SE	(1.4158)	(0.9747)	(1.2666)
	Z: Land Gradient	F	8.26	6.02	
		Overid p		0.1934	
		Z^2 p		0.0995	
Hunt	*Y: Ln Patents/Cap*	$\hat{\beta}$	17.6333***	17.5881***	17.6455***
et al. (2010)	*X: Skill Immigrant: College*	SE	(5.3499)	(5.3239)	(5.3559)
	Z: Predict X: Hist Immigrant	F	26.74	13.6	
		Overid p		0.7372	
		Z^2 p		0.9633	

Table A.1

AER and AEJ Replications

			Linear	**Quadratic**	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Ln Patents/Cap	$\hat{\beta}$	18.9134	19.0753	19.0771
	X: Skill Immigrant: Post Coll	SE	(13.5348)	(13.6413)	(13.6440)
	Z: Predict X: Hist Immigrant	F	18.03	9.47	
		Overid p		0.9518	
		Z^2 p		0.7316	
Lipscomb	*Y: < 4 Yrs Edu*	$\hat{\beta}$	-21.2534***	-23.3515***	-31.3959**
et al. (2013)	*X: Electric Availability*	SE	(7.7510)	(8.5506)	(12.6881)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.106	
		Z^2 p		0.757	
	Y: Economically Active	$\hat{\beta}$	0.1728***	0.1750***	0.1763***
	X: Electric Availability	SE	(0.0499)	(0.0538)	(0.0544)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.7784	
		Z^2 p		0.757	
	Y: Formal Employ	$\hat{\beta}$	0.1836***	0.1881***	0.1933***
	X: Electric Availability	SE	(0.0514)	(0.0562)	(0.0584)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.6052	
		Z^2 p		0.757	
	Y: Formal Employ: Rural	$\hat{\beta}$	0.1647***	0.1617***	0.1634***
	X: Electric Availability	SE	(0.0545)	(0.0576)	(0.0584)
	Z: Predict X: Geo Invest Cost	F	18.37	9.53	
		Overid p		0.7357	
		Z^2 p		0.7575	
	Y: Formal Employ: Urban	$\hat{\beta}$	0.1762***	0.1874***	0.2270***
	X: Electric Availability	SE	(0.0507)	(0.0565)	(0.0748)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.2038	
		Z^2 p		0.757	
	Y: Gross Inc/Cap	$\hat{\beta}$	0.1115**	0.1194**	0.1357**
	X: Electric Availability	SE	(0.0455)	(0.0509)	(0.0583)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.3611	
		Z^2 p		0.757	
	Y: HDI: Education	$\hat{\beta}$	0.1878***	0.1891***	0.1893***
	X: Electric Availability	SE	(0.0573)	(0.0618)	(0.0619)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.8804	
		Z^2 p		0.757	

Table A.1

AER and AEJ Replications

			Linear	**Quadratic**	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: HDI: Inc	$\hat{\beta}$	0.4499***	0.4647***	0.4866***
	X: Electric Availability	SE	(0.1533)	(0.1649)	(0.1736)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.5478	
		Z^2 p		0.757	
	Y: HDI: Longrevity	$\hat{\beta}$	-0.0046	-0.0198	-0.0284
	X: Electric Availability	SE	(0.0502)	(0.0540)	(0.0796)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.0204	
		Z^2 p		0.757	
	Y: Housing Value	$\hat{\beta}$	8.8111***	9.6949***	14.4719**
	X: Electric Availability	SE	(3.0253)	(3.3639)	(6.0446)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.087	
		Z^2 p		0.757	
	Y: Human Capital	$\hat{\beta}$	11.5415	10.1088	10.7157
	X: Electric Availability	SE	(7.2985)	(8.6715)	(9.4379)
	Z: Predict X: Geo Invest Cost	F	3.5	1.8	
		Overid p		0.6533	
		Z^2 p		0.5099	
	Y: Human Development Index	$\hat{\beta}$	0.1093**	0.1250***	0.2025**
	X: Electric Availability	SE	(0.0439)	(0.0470)	(0.0902)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.019	
		Z^2 p		0.757	
	Y: Illiteracy	$\hat{\beta}$	-8.3495*	-10.2461*	-16.3831
	X: Electric Availability	SE	(4.7794)	(5.3543)	(10.4555)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.0211	
		Z^2 p		0.757	
	Y: In-Migration	$\hat{\beta}$	0.1019	0.1313	0.1561
	X: Electric Availability	SE	(0.0936)	(0.0826)	(0.1074)
	Z: Predict X: Geo Invest Cost	F	1.98	1.03	
		Overid p		0.4495	
		Z^2 p		0.5838	
	Y: Infant Mortality	$\hat{\beta}$	-11.973	-17.0172	-22.8713
	X: Electric Availability	SE	(18.0789)	(18.5353)	(30.6285)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.0096	
		Z^2 p		0.757	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Life Expectancy	$\hat{\beta}$	-1.0339	-1.0853	-1.0866
	X: Electric Availability	SE	(2.3939)	(2.4463)	(2.4512)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.8513	
		Z^2 p		0.757	
	Y: % Pop Urban	$\hat{\beta}$	0.2379**	0.2635**	0.3594**
	X: Electric Availability	SE	(0.1110)	(0.1221)	(0.1787)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.1674	
		Z^2 p		0.757	
	Y: Pop Density	$\hat{\beta}$	-23.6182	-23.3398	-23.3439
	X: Electric Availability	SE	(19.1952)	(19.9019)	(19.9060)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.8556	
		Z^2 p		0.757	
	Y: Poverty	$\hat{\beta}$	-42.1649***	-45.7629***	-60.2197***
	X: Electric Availability	SE	(13.8406)	(15.6227)	(22.6310)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.1583	
		Z^2 p		0.757	
	Y: Yrs Edu	$\hat{\beta}$	2.0216***	1.9929***	2.0099***
	X: Electric Availability	SE	(0.6686)	(0.7039)	(0.7111)
	Z: Predict X: Geo Invest Cost	F	18.44	9.56	
		Overid p		0.7662	
		Z^2 p		0.757	
Saiz	*Y: Δ Ln Neighborhood Value*	$\hat{\beta}$	-0.3227**	-0.2739***	-0.2740***
et al. (2011)	*X: Δ Foreign Born Pop*	SE	(0.1361)	(0.1029)	(0.1029)
	Z: Gravity Pull	F	30.71	165.22	
		Overid p		0.3347	
		Z^2 p		0.0002	
Werker	*Y: Auto Import: % GDP*	$\hat{\beta}$	0.2785***	0.2847***	0.2848***
et al. (2009)	*X: Foreign Aid: % GDP*	SE	(0.0846)	(0.0676)	(0.0676)
	Z: Muslim x Oil Price	F	28.12	20.73	
		Overid p		0.8801	
		Z^2 p		0.0094	
	Y: Auto Import: % GDP	$\hat{\beta}$	0.2816***	0.2644***	0.2663***
	X: Lag Foreign Aid: % GDP	SE	(0.0817)	(0.0633)	(0.0639)
	Z: Lag Muslim x Oil Price	F	31.82	23.21	
		Overid p		0.6125	
		Z^2 p		0.0156	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Auto Import: % Import	$\hat{\beta}$	-0.0479	-0.0929	-0.0988
	X: Foreign Aid: % GDP	SE	(0.1042)	(0.1023)	(0.1062)
	Z: Muslim x Oil Price	F	53.16	34.49	
		Overid p		0.0997	
		Z^2 p		0.0656	
	Y: Auto Import: % Import	$\hat{\beta}$	-0.1383	-0.1818*	-0.1919*
	X: Lag Foreign Aid: % GDP	SE	(0.0990)	(0.0961)	(0.1005)
	Z: Lag Muslim x Oil Price	F	54.97	35.44	
		Overid p		0.0695	
		Z^2 p		0.0863	
	Y: Capital Import: % GDP	$\hat{\beta}$	0.2070***	0.2508***	0.2571***
	X: Foreign Aid: % GDP	SE	(0.0754)	(0.0709)	(0.0735)
	Z: Muslim x Oil Price	F	28.12	20.73	
		Overid p		0.3251	
		Z^2 p		0.0094	
	Y: Capital Import: % GDP	$\hat{\beta}$	0.2210***	0.2530***	0.2578***
	X: Lag Foreign Aid: % GDP	SE	(0.0674)	(0.0650)	(0.0668)
	Z: Lag Muslim x Oil Price	F	31.82	23.21	
		Overid p		0.3772	
		Z^2 p		0.0156	
	Y: Capital Import: % Import	$\hat{\beta}$	-0.2676**	-0.3050**	-0.3121**
	X: Foreign Aid: % GDP	SE	(0.1193)	(0.1198)	(0.1228)
	Z: Muslim x Oil Price	F	53.16	34.49	
		Overid p		0.2143	
		Z^2 p		0.0656	
	Y: Capital Import: % Import	$\hat{\beta}$	-0.3565***	-0.3540***	-0.3540***
	X: Lag Foreign Aid: % GDP	SE	(0.1123)	(0.1114)	(0.1115)
	Z: Lag Muslim x Oil Price	F	54.97	35.44	
		Overid p		0.9224	
		Z^2 p		0.0863	
	Y: Exports	$\hat{\beta}$	0.1077	0.2243	0.2423
	X: Foreign Aid: % GDP	SE	(0.1447)	(0.1397)	(0.1513)
	Z: Muslim x Oil Price	F	45.8	32.15	
		Overid p		0.0156	
		Z^2 p		0.0326	
	Y: Exports	$\hat{\beta}$	0.1801	0.2063*	0.2075*
	X: Lag Foreign Aid: % GDP	SE	(0.1245)	(0.1246)	(0.1253)
	Z: Lag Muslim x Oil Price	F	49.64	35.23	
		Overid p		0.4464	
		Z^2 p		0.0484	

Table A.1

AER and AEJ Replications

Paper	Variable	Stat	Linear	Quadratic	
			2SLS	**2SLS**	**LIML**
	Y: Gov Final Consumption	$\hat{\beta}$	0.1061	0.0498	0.0477
	X: Foreign Aid: % GDP	SE	(0.1256)	(0.1232)	(0.1305)
	Z: Muslim x Oil Price	F	45.8	32.15	
		Overid p		0.1779	
		Z^2 p		0.0326	
	Y: Gov Final Consumption	$\hat{\beta}$	0.0061	0.0018	0.0017
	X: Lag Foreign Aid: % GDP	SE	(0.1022)	(0.1083)	(0.1083)
	Z: Lag Muslim x Oil Price	F	49.64	35.23	
		Overid p		0.8914	
		Z^2 p		0.0484	
	Y: Gross Capital Formation	$\hat{\beta}$	0.3054**	0.3877***	0.4027***
	X: Foreign Aid: % GDP	SE	(0.1356)	(0.1284)	(0.1357)
	Z: Muslim x Oil Price	F	45.8	32.15	
		Overid p		0.0568	
		Z^2 p		0.0326	
	Y: Gross Capital Formation	$\hat{\beta}$	0.4158***	0.4337***	0.4347***
	X: Lag Foreign Aid: % GDP	SE	(0.1138)	(0.1144)	(0.1148)
	Z: Lag Muslim x Oil Price	F	49.64	35.23	
		Overid p		0.5582	
		Z^2 p		0.0484	
	Y: Gross Domestic Savings	$\hat{\beta}$	-0.9573***	-0.8994***	-0.9097***
	X: Foreign Aid: % GDP	SE	(0.1653)	(0.1567)	(0.1596)
	Z: Muslim x Oil Price	F	45.8	32.15	
		Overid p		0.2398	
		Z^2 p		0.0326	
	Y: Gross Domestic Savings	$\hat{\beta}$	-0.7334***	-0.7068***	-0.7093***
	X: Lag Foreign Aid: % GDP	SE	(0.1497)	(0.1466)	(0.1474)
	Z: Lag Muslim x Oil Price	F	49.64	35.23	
		Overid p		0.5132	
		Z^2 p		0.0484	
	Y: HH Final Consumption	$\hat{\beta}$	0.8512***	0.8496***	0.8496***
	X: Foreign Aid: % GDP	SE	(0.1891)	(0.1915)	(0.1915)
	Z: Muslim x Oil Price	F	45.8	32.15	
		Overid p		0.9774	
		Z^2 p		0.0326	
	Y: HH Final Consumption	$\hat{\beta}$	0.7273***	0.7050***	0.7067***
	X: Lag Foreign Aid: % GDP	SE	(0.1727)	(0.1740)	(0.1745)
	Z: Lag Muslim x Oil Price	F	49.64	35.23	
		Overid p		0.6351	
		Z^2 p		0.0484	

Table A.1

AER and AEJ Replications

			Linear	Quadratic	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Import	$\hat{\beta}$	1.3704***	1.5115***	1.5655***
	X: Foreign Aid: % GDP	SE	(0.2125)	(0.1977)	(0.2103)
	Z: Muslim x Oil Price	F	45.8	32.15	
		Overid p		0.0578	
		Z^2 p		0.0326	
	Y: Import	$\hat{\beta}$	1.3292***	1.3468***	1.3478***
	X: Lag Foreign Aid: % GDP	SE	(0.1997)	(0.1870)	(0.1873)
	Z: Lag Muslim x Oil Price	F	49.64	35.23	
		Overid p		0.7639	
		Z^2 p		0.0484	
	Y: Ln Inflation	$\hat{\beta}$	-0.0666	-0.0437	-0.0525
	X: Foreign Aid: % GDP	SE	(0.0885)	(0.0848)	(0.0992)
	Z: Muslim x Oil Price	F	14.87	12.86	
		Overid p		0.0828	
		Z^2 p		0.62	
	Y: Ln Inflation	$\hat{\beta}$	-0.0469	-0.033	-0.0382
	X: Lag Foreign Aid: % GDP	SE	(0.0800)	(0.0778)	(0.0899)
	Z: Lag Muslim x Oil Price	F	16.67	13.09	
		Overid p		0.0822	
		Z^2 p		0.7296	
	Y: Ln Undervaluation	$\hat{\beta}$	0.0283	0.0530*	0.0565*
	X: Foreign Aid: % GDP	SE	(0.0246)	(0.0280)	(0.0309)
	Z: Muslim x Oil Price	F	26.58	16.14	
		Overid p		0.0261	
		Z^2 p		0.0017	
	Y: Ln Undervaluation	$\hat{\beta}$	0.0295	0.0426	0.047
	X: Lag Foreign Aid: % GDP	SE	(0.0249)	(0.0272)	(0.0305)
	Z: Lag Muslim x Oil Price	F	27.65	14.86	
		Overid p		0.0452	
		Z^2 p		0.0881	
	Y: Net Errors & Omissions	$\hat{\beta}$	-0.3649***	-0.3497***	-0.3502***
	X: Foreign Aid: % GDP	SE	(0.1178)	(0.1102)	(0.1104)
	Z: Muslim x Oil Price	F	26.21	21.17	
		Overid p		0.8098	
		Z^2 p		0.009	
	Y: Net Errors & Omissions	$\hat{\beta}$	-0.3242***	-0.3397***	-0.3404***
	X: Lag Foreign Aid: % GDP	SE	(0.0908)	(0.0851)	(0.0854)
	Z: Lag Muslim x Oil Price	F	29.48	23.43	
		Overid p		0.7499	
		Z^2 p		0.0216	

Table A.1

AER and AEJ Replications

			Linear	**Quadratic**	
Paper	**Variable**	**Stat**	**2SLS**	**2SLS**	**LIML**
	Y: Noncap Import: % GDP	$\hat{\beta}$	0.3780**	0.6329***	0.7486***
	X: Foreign Aid: % GDP	SE	(0.1528)	(0.1576)	(0.2138)
	Z: Muslim x Oil Price	F	28.12	20.73	
		Overid p		0.0055	
		Z^2 p		0.0094	
	Y: Noncap Import: % GDP	$\hat{\beta}$	0.5007***	0.6514***	0.7103***
	X: Lag Foreign Aid: % GDP	SE	(0.1532)	(0.1541)	(0.1755)
	Z: Lag Muslim x Oil Price	F	31.82	23.21	
		Overid p		0.0000	
		Z^2 p		0.0156	
	Y: Noncap Import: % Import	$\hat{\beta}$	0.3155*	0.3979**	0.4211**
	X: Foreign Aid: % GDP	SE	(0.1842)	(0.1873)	(0.1971)
	Z: Muslim x Oil Price	F	53.16	34.49	
		Overid p		0.0785	
		Z^2 p		0.0656	
	Y: Noncap Import: % Import	$\hat{\beta}$	0.4948***	0.5358***	0.5444***
	X: Lag Foreign Aid: % GDP	SE	(0.1732)	(0.1747)	(0.1775)
	Z: Lag Muslim x Oil Price	F	54.97	35.44	
		Overid p		0.3173	
		Z^2 p		0.0863	
	Y: per Cap GDP Growth	$\hat{\beta}$	0.2145	0.2777**	0.2873**
	X: Foreign Aid: % GDP	SE	(0.1346)	(0.1329)	(0.1385)
	Z: Muslim x Oil Price	F	45.8	32.15	
		Overid p		0.1286	
		Z^2 p		0.0326	
	Y: per Cap GDP Growth	$\hat{\beta}$	0.2203*	0.2903**	0.3009**
	X: Lag Foreign Aid: % GDP	SE	(0.1188)	(0.1200)	(0.1290)
	Z: Lag Muslim x Oil Price	F	49.64	35.23	
		Overid p		0.0322	
		Z^2 p		0.0484	

Appendix B. Weight Ratio Details

Appendix B.1. Weight Derivation

Angrist, Graddy, and Imbens (2000) set-out the following assumptions needed to interpret IV estimates of equation (4.2) in the presence of unmodelled heterogeneity:

[A1] Independence: $z_i \perp y_i(x, z), x_i(z)$

Importantly, this is not saying that y and x are unrelated to z, but rather that the particular functional forms for $y_i(x, z)$ and $x_i(z)$ are independent of the realized value of the instrument. For instance, while y and x should vary with z, when thinking about the *counterfactual* differences in y and x when z takes on different values for *the same individual*, it can not be the case that individuals with larger differences between the two states are systematically more likely to have a particular value of z.

[A2] Exclusion: $y_i(x, z) = y_i(x, z')$ for $z \neq z'$

Assumption [A2] simply states that the instrument effects y only through its effect on x. Assumptions [A1] and [A2] are akin to the mean independence assumption—$E(u|\mathbf{z}) = 0$— in section 2.

[A3] Relevance: $x_i(z)$ is a non-trivial function of z

This assumption states that the instrument does influence x and is akin to the rank condition in the linear case.

[A4] Monotonicity: Either $\frac{\partial x_i}{\partial z}(z) \leq 0$ or $\frac{\partial x_i}{\partial z}(z) \geq 0$ for all *units* (defined by i) at *any value* of the instrument

Importantly, this does not assume that x must always be either increasing or decreasing in z. Rather, if *at a particular* z increasing z increases x for some units, then it must not decrease x for other units.[23]

[23] For instance, in Figure 3.1 the quadratic fit suggests that x is not monotonic in z (decreasing then increasing in z), but this does not imply a violation of the monotonicity assumption needed here. The monotonicity assumption requires that at any particular Distance from Wittenberg, the level of Protestantism in all counties would weakly respond in the same direction to a change in distance.

Angrist, Graddy, and Imbens (2000) express the weights from Equation (4.3) as:

$$\lambda_g(z) = \frac{\frac{\partial x}{\partial z}(z) \cdot \int_z^{\infty} \left(g(\zeta) - E\left[g(z_i)\right]\right) \cdot f_z(\zeta) d\zeta}{\int \frac{\partial x}{\partial z}(\nu) \cdot \int_{\nu}^{\infty} \left(g(\zeta) - E\left[g(z_i)\right]\right) \cdot f_z(\zeta) d\zeta d\nu} \tag{B.1}$$

Following directly from the proof of Theorem 4 in Angrist, Graddy, and Imbens (2000), the denominator in equation (B.1) is simply $Cov(x_i, g(z_i))$. The integral representation is helpful for establishing that the weights sum to one. However, for our purposes it is instructive to revert back to the covariance representation as it makes for a more straightforward estimation of the ratio of weights from the two different estimators. Specifically, this covariance is directly estimable from the data with no further assumptions, while estimating the $\partial x/\partial z$ component of the denominator would require making assumptions about $x(z)$.

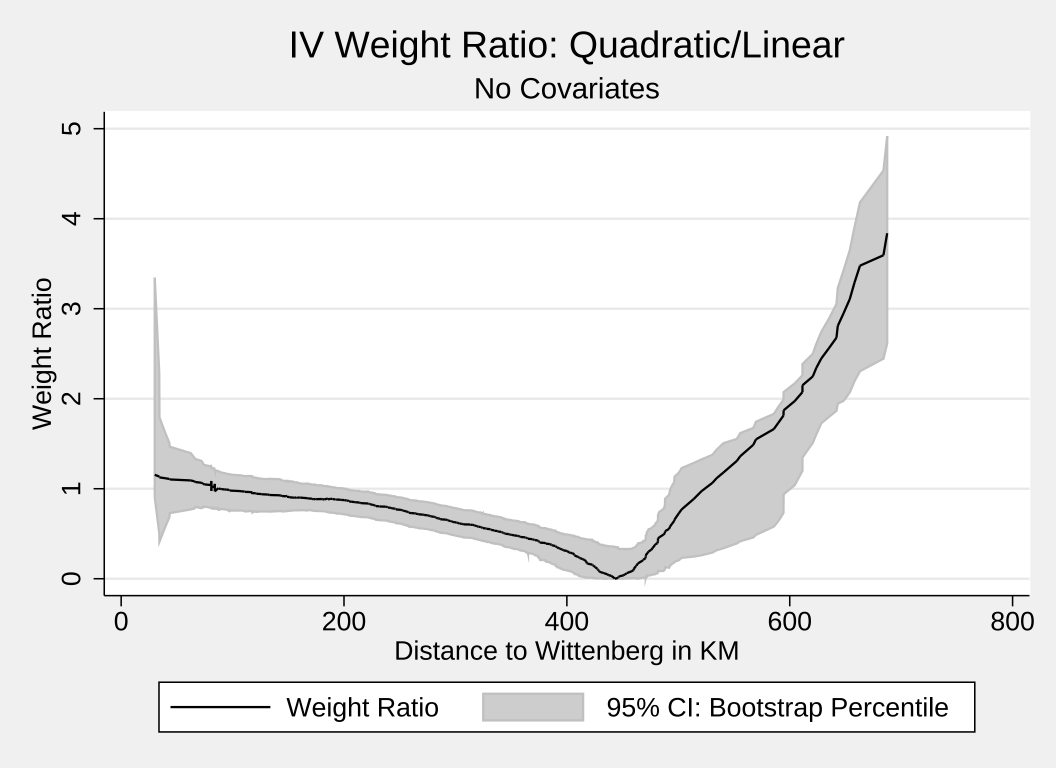
For the numerator, we can start by noting that by demeaning x and z, we have that $E\left[g(z_i)\right] = 0$, simplifying the expression slightly. Next, Angrist, Graddy, and Imbens (2000) derive their result as the limiting case of a multi-valued discrete instrument where the discrete values of z are ordered by the implied value of x— that is they are ordered by $g(z)$. Since 2SLS is equivalent to using the first stage fitted values as the instrument in an IV estimation, we can rewrite the integral as going over values of $g(z)$. Then we can rewrite the integral as the product of a conditional expectation and a probability based on the fact that the truncated density— conditional on being larger than some value a— for a random variable X can be written as $f(X)/Pr(X > a)$. For a given value of z, say z^*, the conditional expectation is simply the expected value of the first stage fitted values, given by $g(z)$, conditional on having a value of $g(z)$ greater than $g(z^*)$. The probability is simply the probability that the value of $g(z)$ is greater than $g(z^*)$. It follows that we can write the weight as:

$$\lambda_g(z) = \frac{\frac{\partial x}{\partial z}(z) \cdot E\left[g(\zeta) | g(\zeta) > g(z)\right] Pr\left(g(\zeta) > g(z)\right)}{Cov(x_i, g(z_i))}$$

Appendix B.2. Weight Ratio with Confidence Intervals

Figure B.1 displays the same estimated weight ratios as in figure 4.1, but with bootstrapped confidence intervals. For each value of z, the CI is based on a separate (specific to the value of z) 1000 replication bootstrap procedure of estimating the two first stages and constructing the weight ratios. The confidence intervals presented are based on the percentiles of the weight ratio estimate distribution across the bootstrapped samples. We use the percentiles rather than a normal approximation to respect the fact that the weight ratio must be nonnegative.

Figure B.1: Becker and Woessmann (2009) IV Weight Ratio with Bootstraped 95% Confidence Intervals



We see that the estimated weight ratio is much more precise at intermediate values. The precision depends on two main factors: the precision of the estimated parameters of $g_1(z)$ and $g_2(z)$ and how sensitive the ranking within the fitted value distribution is for a particular value of z across different draws. The precision of the estimated parameters affects the weight ratio estimates at all values of z similarly, while the placement in the ranking will differ by values of z. The relative precision at intermediate values is due to the fact that the distribution of z is denser at intermediate values. In particular, this implies that the relative ranking in the $g(z)$ distribution is less sensitive across different bootstrap replications where the density is higher and therefore the conditional expectations and probabilities are also less sensitive.

Appendix B.3. Weight Ratio Decomposed

As argued previously, the appeal of analyzing the weight ratios instead of the weights is the fact that the component due to the true underlying relationship between x and z, $\partial x/\partial z$, cancels out in the ratio. However, in order to provide more insight into why the weight ratio evolves the way it does, we can decompose it into the contribution due to the quadratic and linear first stages. That is, we can consider the following components of the weight ratio:

$$\tilde{\lambda}_g(z) = \frac{E\left[g(\zeta)|g(\zeta) > g(z)\right] Pr\left(g(\zeta) > g(z)\right)}{Cov(x_i, g(z_i))}$$

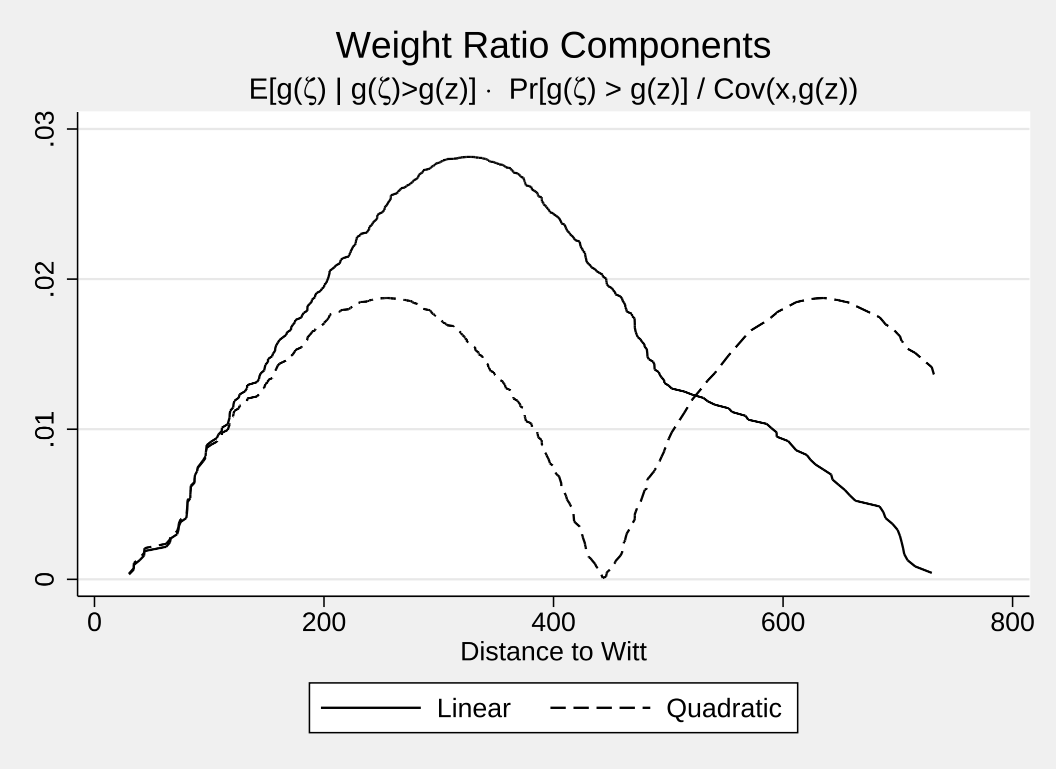
$\tilde{\lambda}_1(z)$ and $\tilde{\lambda}_2(z)$ are simply the denominator and numerator of the weight ratio— or they can be thought of as the weights for the linear- and quadratic-in-z 2SLS estimators if $\partial x/\partial z$ were constant (i.e. if all units complied with the instrument in the same way). Plotting these components separately helps provide more intuition on how the two estimators— using a linear-in-z or quadratic-in-z first stage— differ.

Figure B.2 displays the weight ratio components for the Distance to Wittenberg example. Loosely speaking, the quadratic-in-z estimator shifts some weight to counties further from Wittenberg that have similar first stage fitted values to those at intermediate distance. This is directly related to the idea of ordering the counties by the values of $g(z)$ when calculating the weight ratios and reflects the quadratic fit presented in Figure 3.1. Once again, the importance of these differences for the final estimates is driven by the level of compliance ($\partial x/\partial z$) at each value of the instrument. The patterns in Figure B.2 also confirm that the difference in weighting is not due to shifting all the weight to high values of z, rather it reflects placing similar weight on distances that imply similar first stage fitted values.

Appendix B.4. Weight Ratios with Covariates

Since the argument is often made that an instrument is "as good as randomly assigned" once other factors are controlled for, it is necessary to introduce other covariates into the model. Here, we will denote the other covariates by w. Following Angrist, Graddy, and Imbens (2000), we assume that the additional covariates enter additively and linearly:

Figure B.2: Becker and Woessmann (2009) IV Weight Ratio Components



Estimated weight ratio components based on the sample analogue of equation (4.5). For each observed value of z, we use the fitted values for the linear and quadratic first stages— $\hat{g}_1(z)$ and $\hat{g}_2(z)$— to estimate the sample mean, probabilities, and covariances.

(A5) Linear and Additive Covariates

$$y(z|w) = y_0(z) + \theta w$$

$$x(z|w) = x_0(z) + \kappa w$$

This assumption requires that the partial effect of z on either x or y does not depend on w.

With the addition of (A5), Angrist, Graddy, and Imbens (2000) show that the IV estimator based on the ratio of the coefficient on z from the second stage reduced form to the coefficient on z from the first stage does not depend on w. The formal proof is in the appendix of Angrist, Graddy, and Imbens (2000) (Lemma 2), however intuitively we are interested in the effect of a change in z on either y or x *holding w fixed*. With w fixed at w^*, the difference between $y(z|w^*)$ and $y(z'|w^*)$ will not depend on θw^* under (A5): $y(z|w^*) - y(z'|w^*) = y_0(z) + \theta w^* - y_0(z') - \theta w^* = y_0(z) - y_0(z')$. The same argument can be made for $x(z, w^*)$. What is important for us is that the addition of other covariates does not change the basic setup of the problem. Certainly, $g(\cdot)$ is now also a function of w, but the conditional expectation is still based on the fitted values that take into account the relationship between z and w in the sample:

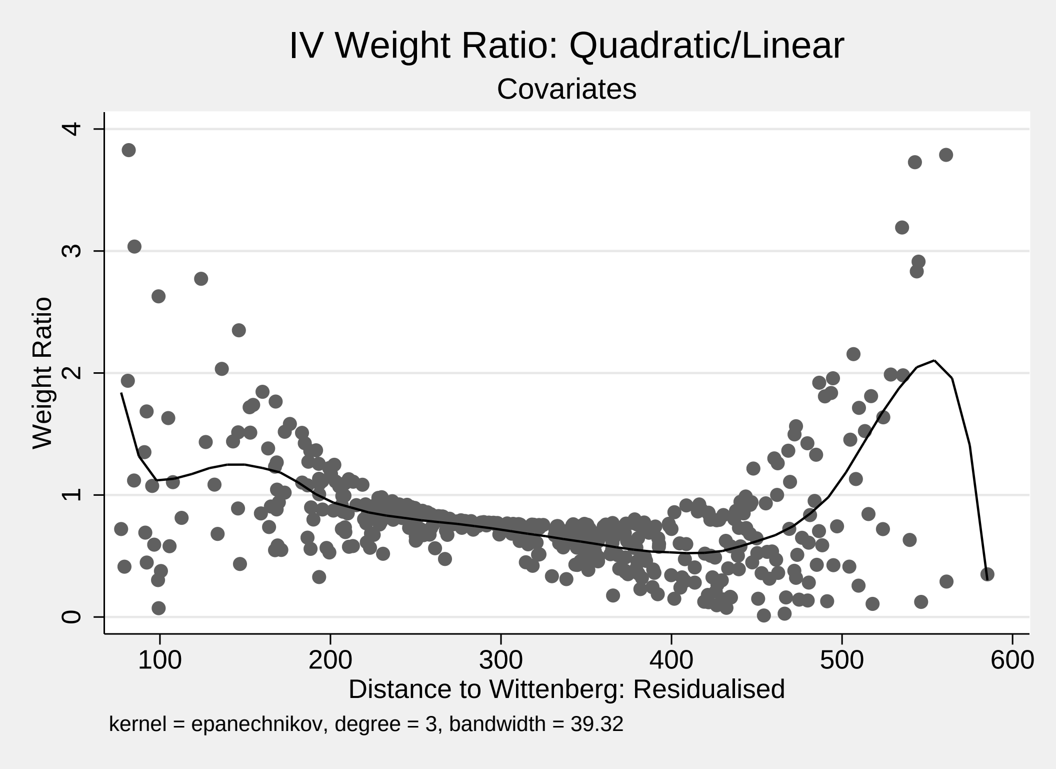
$$\begin{aligned}\frac{\lambda_2(z,w)}{\lambda_1(z,w)} =& A\left[\frac{E\left[g_2(\zeta,w)|g_2(\zeta,w) > g_2(z,w)\right]\cdot Pr(g_2(\zeta,w) > g_2(z,w))}{E\left[g_1(\zeta,w)|g_1(\zeta,w) > g_1(z,w)\right]\cdot Pr(g_1(\zeta,w) > g_1(z,w))}\right]\\ &\text{where } A = \frac{Cov\left(x_i, g_1(z_i,w_i)\right)}{Cov\left(x_i, g_2(z_i,w_i)\right)}\end{aligned}$$

We could estimate weight ratios for different values of w. Instead, we simply use the conditional expectations of $g(z, w)$ using the realized values of w that are associated with each observation. This is the natural extension of viewing 2SLS as IV using the first stage fitted values as the instrument. In practice we use the residual from a regression of z on w as the instrument. Empirically this introduces the complication that there is not a simple mapping from our weight ratio to the original z.

Now we return to BW's main result for the Literacy outcome when including other covariates. Recall from Table 3.1, BW get a statistically significant estimate of $\hat{\beta}_1 = 0.1885$, but when we include the square of the distance from Wittenberg as an additional instrument, the point estimate falls to $\hat{\beta}_2 = 0.0932$. To account for covariates, we regress each instrument on the other covariates and use the residual as the instrument. This does not change the 2SLS estimate, but is helpful for focusing on the the role the instruments play once the other covariates have been partialled out. In Figure B.3, we plot the weight ratio with respect to the residualized distance to Wittenberg where we have added the mean back in. Here, it is helpful to consider the residualized distance as an "effective" distance. That is, there are other factors (mostly demographic in this case) that are correlated with distance whose effect can be cast in terms of an equivalent change in the distance from Wittenberg. Figure B.3 also plots a local polynomial smoothing line to help summarize the results.

The quadratic tends to place more weight on counties with a residualized distance from Wittenberg that is either at the low (<200km) or high end (>500km) while distances near the middle are given less weight on average. In order to explain why $\hat{\beta}_2 < \hat{\beta}_1$, the partial effects need to be smaller on average for counties when they are, in effective distance, either

Figure B.3: Becker and Woessmann (2009) IV Weight Ratio with Covariates



the closest to or farthest from Wittenberg.

Appendix C. Albouy (2012) (2012) Critique of AJR

As was mentioned earlier, the original results from AJR have been called into question due to data and methodological concerns. Most notably, these concerns appear in Albouy's 2012 comment on the original AJR paper. Albouy (2012) makes several critiques, however three particularly poignant concerns are aimed at the quality of the data collected, how the standard errors and confidence intervals are determined, and potential issues due to not distinguishing data collected based on soldiers on campaign or slave laborers. In terms of data quality, Albouy notes that many of the country level observations are derived from historical records for other countries with various adjustments made. Albouy refers to such cases as having "Conjectured Mortality Rates" and estimates separate coefficients using the subsample of countries without conjectured rates.

Albouy raises two issues with the implied precision of the estimates. The first is partially related to the previous data concern. Namely, he chooses to cluster the standard errors so that all the countries with a mortality rate derived from the same source are included in the same cluster. Second, he proposes the use of Anderson-Rubin Confidence Intervals (AR CI) for inference. While traditional confidence intervals based on the Wald statistic can be susceptible to problems with weak instruments, the AR CI are more robust to the presence of weak instruments (see Moreira (2009) for a discussion of the AR test statistic).

Table C.1 provides a partial replication of Albouy's analysis. We choose the same baseline samples and specifications as for our AJR replication and extension. Note that Albouy does not present results for the subsample excluding NeoEuropes when including the latitude control or for the Non-African subsample. For comparability to our AJR table, we present those results using Albouy's data and methodology. Panel A reflects Albouy's replication of AJR, Panel B checks the sensitivity to including indicators for data originating from soldiers on campaign or slave laborers, Panel C removes all data points based on conjectured mortality rates, and Panel D includes the campaign and slave labor dummies when using the Non-Conjectured subsample.

Starting with Panel A, we see linear-in-z estimates very similar to the original AJR results in Table 3.2. Note that the AR CI for the first three rows are not symmetric around the point estimate, but do exclude zero. For the base sample when including continent indicators, the AR CI is even more unorthodox consisting of two asymmetric and unbounded sets, a potential outcome when using AR CI. As Albouy notes, since the IV estimate can be cast as the ratio of the the second stage reduced form coefficient on z to the first stage coefficient on z, it makes sense that if zero cannot be ruled out for the first stage coefficient, infinity can not be ruled out for the IV estimate. These unbounded and disjoint AR CI reflect the underlying uncertainty for the estimates and represent one of the key criticisms of AJR found in Albouy. Note that the analysis of the quadratic-in-z estimates follows that from our AJR extension. However, the AR CIs provide another point of comparison. Namely, the AR CI tend to be smaller, in some cases considerably so, when we include the squared mortality rate as an additional instrument. For instance, the AR CI for the base sample shrinks from an implied effect of a one standard deviation increase in protection between 160% and 3500% to between 135% and 315%. Perhaps more important, the AR CI for the Continent Indicators specification becomes bounded (i.e. no longer includes positive and negative infinity) and excludes zero. This marks an important improvement in inference due to increased efficiency in the first stage. Abstracting from the change in the coefficient estimate, this suggests that part of Albouy's criticism may by ameliorated by choosing a better fitting first stage function.

When controlling for campaign and slave labor dummies in Panel B, we see the linear-in-z point estimates change from Panel A. However, as before the Non-African subsample is somewhat robust to the inclusion of these additional regressors. The overidentification results are more mixed, with a rejection in only one case, but with practically different estimates for two cases. For the base sample, the quadratic first stage delivers a bounded AR CI that excludes zero, while the linear first stage did not.

By removing observations with conjectured mortality rates in Panels C and D, we see more instances where the AR CI are improved considerably by changing the first stage function to include a quadratic in the instrument. Despite some very large differences in point estimates between the linear- and quadratic-in-z cases, there are no cases were the overidentification test rejects. This is likely due to the small sample size (between 13 and 28 countries) and resulting lack of power. Taken as a whole, the results of our Albouy replication and extension suggest a role for instrument polynomials in the first stage to greatly improve the precision of estimates. Coupled with our initial motivation to explore the validity of the instrument, this suggests a broader role for considering higher order polynomials of continuous instruments. The fact that this approach of adding the squared instrument is simple to implement makes it appealing as a common sensitivity analysis to be undertaken.

Table C.1

Albouy (2012) Replication and Extension

Sample &		**Linear**	**Quadratic**	
Specification	**Statistic**	*2SLS*	*2SLS*	*LIML*
Panel A: AJR Replication				
Base	$\hat{\beta}$	0.9620***	0.7282***	0.8343***
	s.e.	(0.2655)	(0.1228)	(0.1876)
	AR CI	[0.64, 2.39]	[0.57, 0.95]	
	First Stage F	7.30	15.74	
	Overid p-value		0.0726	
	Z^2 p-value		0.0217	
Excluding	$\hat{\beta}$	1.1647***	0.9737***	1.0905***
NeoEuropes	s.e.	(0.4219)	(0.2345)	(0.3280)
	AR CI	[0.70, 7.26]	[0.61, 1.77]	
	First Stage F	4.56	5.03	
	Overid p-value		0.2577	
	Z^2 p-value		0.1916	
Excluding	$\hat{\beta}$	0.5994***	0.5796***	0.5830***
Africa	s.e.	(0.1055)	(0.0883)	(0.0902)
	AR CI	[0.40, 0.89]	[0.36, 0.82]	
	First Stage F	37.89	25.89	
	Overid p-value		0.6458	
	Z^2 p-value		0.1630	
Base w/	$\hat{\beta}$	1.0739*	0.7273***	0.8139***
Continent	s.e.	(0.5330)	(0.2049)	(0.2558)
Indicators	AR CI	(-∞, -3.08]U[0.41, +∞)	[0.18, 1.86]	
	First Stage F	2.72	5.70	
	Overid p-value		0.2500	
	Z^2 p-value		0.0649	

Table C.1

Albouy (2012) Replication and Extension				
Sample &		**Linear**	**Quadratic**	
Specification	**Statistic**	*2SLS*	*2SLS*	*LIML*
Panel B: Add Campaign and Slave Labor Indicators				
Base	$\hat{\beta}$	1.1536**	0.7390***	0.9969**
	s.e.	(0.5078)	(0.2104)	(0.4903)
	AR CI	$(-\infty, -17.59]\text{U}[0.60, +\infty)$	[0.54, 1.60]	
	First Stage F	3.67	6.49	
	Overid p-value		0.0690	
	Z^2 p-value		0.0696	
Excluding	$\hat{\beta}$	1.3043**	1.0814***	1.2673**
NeoEuropes	s.e.	(0.6288)	(0.3905)	(0.5782)
	AR CI	$(-\infty, -5.80]\text{U}[0.64, +\infty)$	$(-\infty, -1.92]\text{U}[0.54, +\infty)$	
	First Stage F	3.14	2.80	
	Overid p-value		0.3178	
	Z^2 p-value		0.3621	
Excluding	$\hat{\beta}$	0.6589***	0.6330***	0.6383***
Africa	s.e.	(0.1490)	(0.1310)	(0.1354)
	AR CI	[0.40, 1.41]	[0.34, 1.51]	
	First Stage F	17.30	9.47	
	Overid p-value		0.6984	
	Z^2 p-value		0.2396	
Base w/	$\hat{\beta}$	1.1866	0.7477***	0.8882**
Continent	s.e.	(0.7311)	(0.2478)	(0.3639)
Indicators	AR CI	$(-\infty, -0.67]\text{U}[0.29, +\infty)$	$(-\infty, -4.85]\text{U}[-0.67, +\infty)$	
	First Stage F	1.73	3.44	
	Overid p-value		0.2271	
	Z^2 p-value		0.1219	

Table C.1

Albouy (2012) Replication and Extension				
Sample &		**Linear**	**Quadratic**	
Specification	**Statistic**	*2SLS*	*2SLS*	*LIML*
Panel C: Remove Conjectured Mortality				
Base	$\hat{\beta}$	0.8203**	0.7300***	0.7595***
	s.e.	(0.3382)	(0.2219)	(0.2419)
	AR CI	$(-\infty, -7.92]\text{U}[0.38, +\infty)$	[0.32, 1.89]	
	First Stage F	3.62	4.29	
	Overid p-value		0.4897	
	Z^2 p-value		0.2255	
Excluding	$\hat{\beta}$	0.9464*	0.9461**	0.9461**
NeoEuropes	s.e.	(0.4996)	(0.4500)	(0.4500)
	AR CI	$(-\infty, -1.38]\text{U}[0.35, +\infty)$	$(-\infty, -0.23]\text{U}[0.31, +\infty)$	
	First Stage F	2.37	1.83	
	Overid p-value		0.9964	
	Z^2 p-value		0.7399	
Excluding	$\hat{\beta}$	0.9545***	0.7490***	0.8362**
Africa	s.e.	(0.2890)	(0.2328)	(0.3068)
	AR CI	[0.56, 3.55]	[0.52, 3.95]	
	First Stage F	6.67	4.16	
	Overid p-value		0.3445	
	Z^2 p-value		0.2040	
Base w/	$\hat{\beta}$	1.2451	0.6586*	0.8840*
Continent	s.e.	(1.1815)	(0.3538)	(0.5115)
Indicators	AR CI	$(-\infty, +\infty)$	[-1.58, 2.19]	
	First Stage F	0.91	4.37	
	Overid p-value		0.2001	
	Z^2 p-value		0.1582	

Table C.1

Albouy (2012) Replication and Extension				
Sample &		**Linear**	**Quadratic**	
Specification	**Statistic**	*2SLS*	*2SLS*	*LIML*
Panel D: Remove Conjectured Mortality & Add Campaign and Slave Labor Indicators				
Base	$\hat{\beta}$	0.8997	0.6845	0.7780
	s.e.	(0.9326)	(0.4899)	(0.6300)
	AR CI	$(-\infty, +\infty)$	$(-\infty, +\infty)$	
	First Stage F	0.67	1.35	
	Overid p-value		0.5674	
	Z^2 p-value		0.4614	
Excluding	$\hat{\beta}$	0.8259	0.7779	0.8335
NeoEuropes	s.e.	(0.8117)	(0.7899)	(0.9472)
	AR CI	$(-\infty, +\infty)$	$(-\infty, +\infty)$	
	First Stage F	0.65	0.35	
	Overid p-value		0.6044	
	Z^2 p-value		0.8207	
Excluding	$\hat{\beta}$	1.0331*	0.5963	0.6922
Africa	s.e.	(0.4885)	(0.4118)	(0.5812)
	AR CI	[0.34, 6.61]	$(-\infty, -47.25]$U$[0.09, +\infty)$	
	First Stage F	5.79	3.36	
	Overid p-value		0.1231	
	Z^2 p-value		0.1871	
Base w/	$\hat{\beta}$	1.4376	0.5134	0.8013
Continent	s.e.	(2.6149)	(0.4478)	(0.8753)
Indicators	AR CI	$(-\infty, +\infty)$	$(-\infty, +\infty)$	
	First Stage F	0.29	2.88	
	Overid p-value		0.2306	
	Z^2 p-value		0.2309	