Broken Instruments

Trevor Gallen [*] Ben Raymond

January 2023

Abstract

Repeated use of related instrumental variables by researchers can collectively invalidate these instruments. This paper examines two ways in which this can happen. First, when instruments sharing significant sources of variation are used to instrument multiple distinct covariates, it is increasingly likely the exclusion restriction was not satisfied in any individual specification from the outset. Second, when a variable is documented to affect many outcomes that are likely to be highly or even mildly persistent, using lagged values of that variable as an instrument is likely to violate the exclusion condition. This paper documents 959 instrumental variables papers from 1995-2019 in highly-ranked economics general interest and field journals. We find six groups of commonly-used instruments whose literatures, taken together, suggest they are likely to fail the strict exogeneity condition. These six instruments have been used in 83 top five publications and 235 well-ranked field or general interest journals outside the top five. We propose a new statistical test for suspect regressions and discuss its asymptotic properties. We then apply it to two IV papers.

JEL Classification: C13, C26, C36
Keywords: Instrumental variables, commonly-used instruments, exclusion restriction

[*] The views expressed herein are those of the authors and do not necessarily reflect the views of the Bureau of Labor Statistics. The authors would like to thank Natalie Bau, Tim Bond, Ian Fillmore, Dominic Smith, Jake Schild, Eliza Forscythe, Yana Gallen, Josh Gottleib, Peter Hull, Scott Imberman, Mohitosh Kejriwal, Steven Levitt, Daniel Millimet, Tim Moore, Casey Mulligan, Devin Pope, Jonathan Skinner, Chad Syverson, Chris Taber, and participants at Purdue University, University of Chicago, McGill University, University of Washington at St. Louis, the 2021 RAND Causal Inference Seminar, and the ASSA 2022 meetings for helpful comments and discussions. Correspondence: Purdue University Krannert School of Business, West Lafayette, IN 47906. Tel.: (765) 496-2458. Email: tgallen@purdue.edu. Web: www.tgallen.com

I Introduction

A strength of empirical economic research is its focus on causality and mechanisms that underlie economic phenomena. A crucial tool in determining causality has been instrumental variables (IV) regression, typically used to isolate a single causal channel. Unfortunately, a good instrument is hard to find because it must have a large impact on one variable, but no uncontrolled direct impact on related variables. Consequently, economists have exploited several promising and ubiquitous sources of natural variation—the weather, sibling structure, topological variation, religion, immigration, and language differences across regions—as instruments for hundreds of variables.

In this paper, we document all uses of instrumental variables in top five economics journals from 1995-2019. Within these hundreds of prominently published papers using IV strategies, we identify six repeatedly used instruments. These instruments are used 83 times in top five journals and 318 times in well-ranked field or general interest journals (including the top five). Our survey of publications is not restricted to one particular literature, and we find that the same or highly related instruments are often used across economic fields that do not cross-cite. We make the case that the causal channels established in a given paper frequently conflict with the exogeneity narrative of other papers using the same instrument.

The idea that overuse of an instrument in a single field may collectively invalidate it is not new. Bazzi and Clemens (2013) argues that growth instruments (legal origin and population) are overused in the macroeconomic growth literature. While the paper has hundreds of citations, we document that this issue has continued unabated both inside and outside of the growth literature. Consequently, we seek to further highlight this common issue and argue that in many cases, the use of an IV in one setting invalidates the use of the same IV in another setting. As a concrete example, Cutler and Glaeser (1997) use topology as an instrument for the segregation of cities which affects education, income, and single motherhood by race. In the macro literature on the effects of the 2008 housing bust, Mian and Sufi (2014) use a related topological measure as an instrument for housing supply elasticity to understand the effects of local house price variation on employment over the business cycle. But if those areas with a low housing supply elasticity due to topology also have high segregation for the same reason, and segregation affected employment across cities the 2008 Financial Crisis, then the exclusion restriction may be violated.

When does the repeated use of an instrument pose a threat to identification? We identify six commonly used instruments, but do not claim that all of the uses of these instruments are invalid. To determine whether the repeated use of an instrument poses a threat to identification, we propose a new, Hausman-like test for commonly used instruments. Our test compares two estimators. The first is the standard "single-paper" regression ignoring all other uses of a given IV. The second estimator includes as exogenous controls all other instrumented-for endogenous variables from other papers.[1] Under the null hypothesis that both estimators are consistent, the two coefficients will be approximately equal. Under the alternative hypothesis that either (or both) of the estimators are not consistent, our test would reject. The intuition behind our test is simple: when one or both of the estimators are inconsistent, they are unlikely to converge to the same value. Consequently, rejection of the null hypothesis verifies an issue with one (or both) of the estimators, while failure to reject gives confidence that the issues we discuss in this paper are not a concern. The difference between our test and the Hausman test is that ours does not assume efficiency for either estimator under the null, and so, the covariance between estimators may be larger than the smaller of the two estimator's variances, rather than equal as in Hausman's test. Our applications confirm the importance of this distinction in practice.

We demonstrate usefulness of our test, both in terms of power and in terms of ease of application. To demonstrate usefulness, we run Monte Carlo simulations in three environments. First, data are generated by a process in which not controlling for other potentially endogenous covariates is optimal, while controlling for them would yield an inconsistent estimator. In the second, controlling for endogenous covariates is optimal, while controlling yields an inconsistent estimator. In the last, the researcher has no viable consistent estimator: controlling and not controlling are inconsistent. We show that in all three environments, uncontrolled estimates that fail to be rejected by our test have good mean square error properties. This occurs because estimates are likely to fail to be rejected by our test only if the realized bias in a trial for both estimators is small, even if it is nonzero. In addition to demonstrating the test performs in theory, we apply our test to two highly cited papers. Our test casts doubt on the validity of bodies of water and elevation as a source of variation in housing prices unrelated to employment (Mian and Sufi, 2014) for reasons related to those described above. In contrast, our test does not reject the null for Rupert and Zanella (2018), who use firstborn gender as an instrument for age at grandparenthood, despite the fact that firstborn gender is also used as an instrument for single motherhood, family size, and a variety of other endogenous variables.

The six highly-repeated instrumental variables we identify are (1) bodies of water and elevation (2) sibling structure (3) ethnolinguistic fractionalization (4) religion (5) weather and (6) immigrant enclaves. The first five are similar, in that the literatures repeatedly use the same (or highly-related instrument) for multiple distinct endogenous covariates, as is famously the case for rainfall. For instance, Lee (2018) uses

[1] We emphasize that this approach is a test, and does not advocate automatically including "bad controls" to estimate the point coefficient.

rainfall as an instrument for crop yields, which affects manufacturing output. However, rainfall is also an instrument for hydroelectric energy production (Roberts and Schlenker, 2013; Allcott, Collard-Wexler, and O'Connell, 2016) and migration (Boustan, Fishback, and Kantor, 2010). Yet it seems plausible that migration and energy prices & quantities may affect crop yields. Our paper raises similar concerns about each of the first five instruments listed. The last instrument, immigrant enclaves, has a slight modification to this direct exogeneity violation, because the same instrument used on the same endogenous covariate for many outcomes is not necessarily a problem. However, when lagged values of a stock variable are used as instruments for current flows, which affect many outcomes, violations of exogeneity may arise, which we discuss at greater length in Section 3.

Our paper contributes to a growing literature criticizing the over-use of particular instruments. Both Bazzi and Clemens (2013) and Morck and Yeung (2011) discuss the potential for repeated use of the same instrumental variable to invalidate causal inference across papers. Bazzi and Clemens (2013) discusses the issue of collective invalidation of legal origins and population size instruments in the context of growth regressions and to our knowledge is the first to mathematically set down the idea of a literature's collective invalidation. We build on this work in three ways. We systematically classify the over-use of six additional instruments. Second, we focus on the cross-field connections between instruments, whereas Bazzi and Clemens (2013) focuses only on use population size and legal origins in the economic development literature.[2] Third, we formulate a new test which allows researchers a path forward to use common instruments and sets a framework for testing whether commonly used instruments are likely invalid in a particular setting. Another paper which discusses one of the six instruments we identify as commonly used is Sarsons (2015) which discusses why rainfall likely does not affect conflict solely through income shocks, and may be a bad instrument for conflict. We extend this work by noting more than fifteen other endogenous variables and more than thirty other outcome variables to which the instrument is applied. In addition to this extension, we highlight five other instruments and formally develop a test that can be applied to these commonly used instruments.[3] Dell, Jones, and Olken (2014) also review the literature on weather shocks, but avoid the concerns we raise focusing on the effect of weather on variables, rather than narrow causal chains that weather as an instrument is claimed to uncover.

In more recent work, Mellon (2021) examines the common use of weather as an instrument and,

[2] Bazzi and Clemens (2013) note in footnote 1 that ethnolinguistic fractionalization is sometimes used as an instrument, and the paper uses it as an exogenous control in a test for weak instruments, but do not criticize the instrument directly.

[3] Interestingly, the structure of Sarsons (2015) argument for the invalidity of rainfall as an instrument follows a similar structure to our test, which is general. Ironically, the Sarsons (2015) tests uses differences in the effect of rainfall by location relative to dams. The location of dams is one of the main endogenous regressors for the "topology" instruments described above.

similar to this paper, systematically documents its over-use across fields. Mellon (2021) additionally asks how large the exogeneity violation (from ignoring the endogenous regressor used in another paper) would have to be in order to render the effect found in a given paper insignificant. Again, this paper is the first to investigate the use of IV in general and identify all commonly used instruments, rather than a particular instrument. Beyond weather (Mellon (2021), Sarsons (2015)) and legal origins and population size (Bazzi and Clemens (2013)), we uncover repeated use of ethnolinguistic fractionalization, sibling structure, topological variation, religion, and historic immigration patterns as instruments. In addition, we provide a direct test of the validity of an instrument, while Mellon (2021) proposes a sensitivity analysis. We believe our approaches are complementary and could both improve future research.

Our investigation of commonly used instruments reveals the frequent use of a Card (2001)-style use of historic immigration patterns as an instrument for current immigration. As mentioned, this repeated use of an instrument is somewhat different than the other five we consider. In particular, because of the dynamic nature of the relationship between past and present immigration, potential violations of exogeneity proliferate when outcomes proliferate: a relationship between present immigration and a given outcome suggests a relationship between past immigration and lagged values of the same outcome. This paper is not the first to point out this theoretical issue with lagged instruments. Jaeger, Ruist, and Stuhler (2018) argue that because general equilibrium adjustments take time, and because immigration shocks are correlated over time, short- and long-run effects of immigration may be misstated. Our contribution is to argue the many documented outcomes increases the potential persistence of the system: if outcomes influence one another even weakly over time, persistence of a shock grows nonlinearly with the number of outcomes.

Our paper contributes to a broader literature on the repeated use of the same variation to uncover different causal relationships. Heath et al. (2019) makes the conceptually-related point that reusing natural experiments with different outcomes gives rise to multiple hypothesis testing issues. Heath et al. (2019) examine two in particular: the Regulation SHO pilot and business combination laws, which we do not touch on in this paper.

Other papers have noted that some bad instruments may be used in conjunction with one another. Kolesár et al. (2015) note that when some set of instruments Z_1, Z_2 for covariate of interest X_1 have a direct effect on the outcome of interest Y via some mediator X_2 that is not of interest (are "invalid" instruments), but that those Z's effects on Y through X_2 are independent of Z's effect on Y through X_1, a consistent estimator may still be constructed. As we discuss, because of the nature of these six instruments, we do not believe this is a reasonable assumption in the large majority of cases we document, and frequently only one IV is available. Finally, Young (2019) notes that a number of top instrumental variables papers rely on outliers, which plays a significant role in our small-sample Monte Carlo findings.

Section 2 of this paper describes the instances of repeated use of similar instruments. Section 3 discusses a framework for thinking about multiple-paper exogeneity violations. Section 4 offers a brief discussion of each of the six categories of potentially problematic instruments. Section 5 discusses the asymptotics of our Hausman-like test. Section 6 produces Monte Carlo evidence on IV estimator performance in a multiple-paper setting. Section 7 applies this test to Rupert and Zanella (2018) and Mian and Sufi (2014). Section 8 concludes.

II Trends in the use of common IVs

We create a database of approximately 960 instrumental variables papers from the top five economics journals by rank, and identify six strains of literature in which the exclusion restriction should attract particular attention.[4] First, "elevation and bodies of water," used to isolate exogenous components of housing supply, segregation, governance structure, dam location, infrastructure cost, and broadband provision among many other variables. Second, "sibling structure," used to isolate exogenous components of family size and fertility, father presence, parental wages, child schooling, age at grandparenthood, welfare receipt and geographical mobility, among other uses. Third, ethnicity/ethnolinguistic fractionalization, which is used to instrument for rule of law, corruption, democracy, income and investment, social trust, institutions, creditor protections, and welfare-state generosity. Fourth, religion, which is used to instrument for land regulation, social trust, national uncertainty aversion, the free press, private school share, bank regulation, and work ethic. Fifth, weather, which is used to instrument for agricultural and fishing productivity, economic growth, energy prices, commodity prices, pollution, population, migration, water quality, political changes, and managerial moods. Sixth, immigrant enclaves, which is used solely as an instrument for future flows of immigration, but which has been found to affect highly durable goods, such as physical capital stock, human capital stock, housing stock, and health. We stress that while not every correlation between two endogenous covariates with the same instrument invalidates an IV, the relationship should always be addressed, as a researcher's concern about an exogeneity violation should

[4] Many of these are not identical, but are typically highly correlated, which is why we adopt the phrase "potentially related" rather than "identical." Average hours of sunshine in an MSA and average cloud cover are highly related, as is average hours of sunshine and average temperature. Similarly, ethnic, linguistic,, and ethnolinguistic fractionalization are highly related (the lowest pairwise correlation is 0.70), and are each significantly correlated with religious fractionalization (Alesina et al., 2003).

rise when such correlations are found to be present.

Our data collection process contained two parts. First, from 1995-2019, all uses of articles with the phrase "instrumental variable" and derivative phrases in the American Economic Review, the Journal of Political Economy, the Quarterly Journal of Economics, Econometrica, and the Review of Economic Studies were catalogued. This resulted in approximately one thousand IV uses being examined. Many of these papers included common instruments, such as (1) dynamic panel instruments outlined in Arellano and Bond (1991) or Blundell and Bond (1998) (2) competing product characteristics or counts as in Berry, Levinsohn, and Pakes (1995) (3) Bartik (1991) instruments, and (4) Frankel and Romer (1999) gravity-based instruments. While these are extremely common instruments, we do not discuss them except in the case of immigrant enclaves.

We then examined this list and found the six potentially concerning instruments discussed above. After identifying six commonly-used instruments, we searched for uses of these instruments on economic journal-hosting websites, Elsiever, Online Wiley, JSTOR, and journal-specific websites. Our target journal was typically in the top sixty journals in RePeC's journal rankings. This garnered an additional 231 papers relevant to these six instruments. Importantly, the text below cannot possibly touch on all the uses of these instruments, and consequently we attempt to discuss only the most highly relevant papers, though we include other papers in our tables.

Figure 1 depicts the uses of these six instruments in all surveyed journals individually and jointly over time. While the use of one of these instruments is mostly limited to between two and five papers a year, the cumulative use of these instruments in economics journals reaches 317 papers by 2019.[5] Moreover, the total use of these instruments has been relatively steady since 2006 at approximately 13 papers per year and has not significantly declined for any of them over time. If anything, the data show the use of these IVs is at best leveling off after increasing precipitously from 2002-2013.[6]

III Framework

In this section, we outline potential pitfalls stemming from repeated use of the same instrument or the use of lagged values of a variable as an instrument for a covariate that affects many outcomes. We summarize three major issues with commonly-used instrumental variables. First, the obvious "direct"

[5] We include available categorized articles through September 2019.

[6] We are aware of a long literature on instrumental variables predating 1990. However, it is only in the 1990's that mechanical statements of simultaneous equations with excluded variables makes way for clear treatment and committed defense of instrument exogeneity in the papers we surveyed.

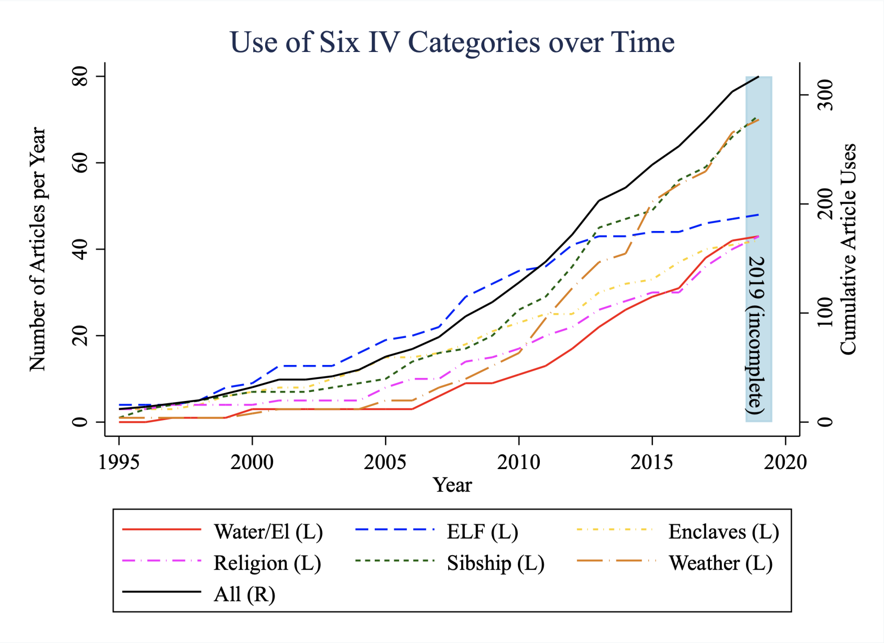

Figure 1: This figure depicts the use of six groups of potentially related instruments in well-ranked economics journals from 1995-2019: (i) elevation and bodies of water (ii) sibling structure (sibship) (iii) ethnicity/ethnic fractionalization (ELF) (iv) religion (v) weather and (vi) immigrant enclaves. Individual yearly uses for each variable and their sum are given by the left axis (L). Cumulative uses of all instruments are given by the right axis (R).

violation of exogeneity described by Morck and Yeung (2011) and Bazzi and Clemens (2013). When two separate papers use the same variable Z as an instrument for both X_1 and X_2's relationship with outcomes Y_1 and Y_2 respectively, we must be concerned with whether X_1 affects Y_2 or X_2 affects Y_1, a standard violation of exogeneity.[7] We term this a "direct" violation of exogeneity. Second, if covariates X_1 and X_2 satisfy the exogeneity condition, but also simultaneously determine one another, or if the error term of X_2 shares common variation with the confounder of X_1 and Y_1, then controlling for X_2 will reintroduce the confounder to the instrumented value of X_1 in the second stage, inducing an exogeneity violation where IV would have been valid in the absence of X_2 as a control. We term this an "induced" violation of exogeneity. Third, when a variable affects many outcomes that are potentially persistent. Even if the persistence of each individual outcome variable is relatively weak, the variables can generate substantial persistence jointly. This substantial joint persistence can induce significant serial correlation, invalidating the instrument. This last is a special case of direct exogeneity violation.

[7] It is theoretically permissible to use different transformations of the same instrument to identify both. For instance, if one endogenous covariate is affected by the presence of rainfall (a dummy) and another is continuously affected, then both can be identified in a joint estimation. However, in practice the collinearity between the two is likely to cause estimates to change, excluded F-statistics to fall, and standard errors to rise, and are likely to invalidate most of the finite-sample instruments we study.

We start with stating the true (generalized) data generating process:

$$X_1 = \gamma_1 Z + \theta_{21} X_2 + \eta_1 \tag{1a}$$

$$X_2 = \gamma_2 Z + \theta_{12} X_1 + \eta_2 \tag{1b}$$

$$Y_1 = \beta_{11} X_1 + \beta_{21} X_2 + \epsilon_1 \tag{2a}$$

$$Y_2 = \beta_{12} X_1 + \beta_{22} X_2 + \epsilon_2 \tag{2b}$$

Paper 1 and Paper 2 are interested in estimating β_{11} and β_{22} respectively. Paper $i \in \{1, 2\}$ estimates β_{ii} via instrumental variables because they believe the covariance between η_i and ϵ_i is nonzero due to a confounder, but that Z is orthogonal to the common variance in their error terms. However, we identify instances in which Paper i ignores X_j $(j \neq i)$ completely, equivalent to an assumption that $\beta_{ji} = 0$ (no direct exogeneity violation) or that $\gamma_j + \gamma_i \theta_{ij} = 0$, so that Z does not affect X_j after accounting for the effect of Z on X_j though X_i. We assume that both papers are correct in identifying confounders and that Z is orthogonal to the confounding variation in both papers, but potentially incorrect in ignoring the other paper's endogenous covariate. Our variance-covariance matrix of errors V is therefore:

$$V\left(\begin{bmatrix} Z \\ \eta_1 \\ \eta_2 \\ \epsilon_1 \\ \epsilon_2 \end{bmatrix}\right) = \begin{bmatrix} \sigma_Z^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\eta_1}^2 & \sigma_{\eta_1\eta_2} & \sigma_{\eta_1\epsilon_1} & \sigma_{\eta_1\epsilon_2} \\ 0 & \sigma_{\eta_1\eta_2} & \sigma_{\eta_2}^2 & \sigma_{\eta_2\epsilon_1} & \sigma_{\eta_2\epsilon_2} \\ 0 & \sigma_{\eta_1\epsilon_1} & \sigma_{\eta_2\epsilon_1} & \sigma_{\epsilon_1}^2 & \sigma_{\epsilon_1\epsilon_2} \\ 0 & \sigma_{\eta_1\epsilon_2} & \sigma_{\eta_2\epsilon_2} & \sigma_{\epsilon_1\epsilon_2} & \sigma_{\epsilon_2}^2 \end{bmatrix} \tag{3}$$

Importantly, neither paper places restrictions on $\sigma_{\eta_i\epsilon_j}$, $i \neq j$, so that, for instance, Paper 2's endogenous covariate X_2 may covary with the error term of Paper 1's outcome Y_1. For convenience, we provide a graphical representation of both Equations 1a-2b and the covariance matrix in Equation 3 in Figure 2, along with illustrative variables relevant to the sibship instrument.[8] If the confounder between schooling

[8] Usefully, the simultaneous equations (1a)-(2b) and 3 can be solved and rewritten as the reduced form:

$$X_1 = Z\frac{\gamma_1 + \theta_{21}\gamma_2}{1 - \theta_{21}\theta_{12}} + \frac{\eta_1 + \theta_{21}\eta_2}{1 - \theta_{21}\theta_{12}}, \quad X_2 = Z\frac{\gamma_2 + \theta_{12}\gamma_1}{1 - \theta_{12}\theta_{21}} + \frac{\eta_2 + \theta_{12}\eta_1}{1 - \theta_{12}\theta_{21}}, \quad \gamma_i^* = \frac{\gamma_i + \theta_{ji}\gamma_j}{1 - \theta_{ij}\theta_{ji}} \quad v_i = \frac{\eta_i + \theta_{ji}\eta_i}{1 - \theta_{ji}\theta_{ij}}$$

So that:

$$X_1 = Z\gamma_1^* + v_1 \quad X_2 = Z\gamma_2^* + v_2$$

and geographic mobility was skill, and skill also affected wages, then our "induced" violation channel would be present through the error terms. However, even if skill did not affect wages, so that η_2 is not correlated with either η_1 or ϵ_1, but schooling did affect wages (simultaneity between X_1 and X_2), then X_2 would still contain confounding information, and controlling for it would invalidate the estimator.

Possibility 1

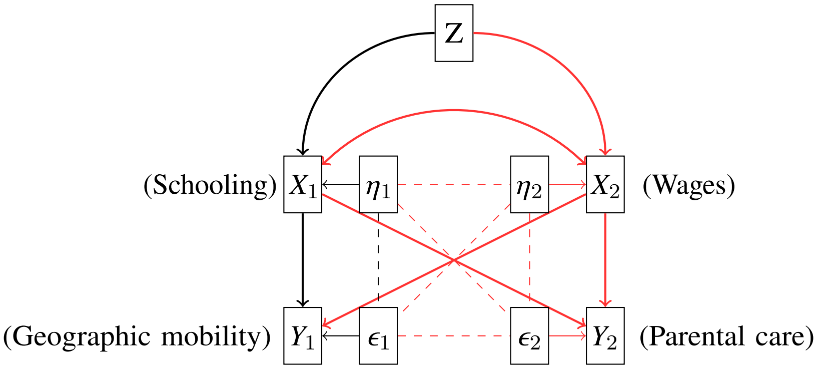

Figure 2: This figure depicts the structure of equations 1a-2b and the covariance structure implied by Equation 3. Potentially nonzero coefficients are denoted by solid arrows, and covariances by dashed lines. Instrument Z is orthogonal to all other errors. Paper 1 assumes only the black arrows and lines, which our full structure is given by both red and black lines. For concrete discussion, example variables are given in parentheses.

Given Equations 1-3, the asymptotic distributions of OLS and IV with and without X_2 as an exogenous control are shown in Table 1. We draw several lessons from Table 1. First, that OLS regressions with and without controls are biased for two distinct reasons. If the econometrician does not control for X_2, not only is the confounder an issue because $Cov(X_1, \epsilon_1) \neq 0$, but if $\beta_{21} \neq 0$, we have an additional omitted variable whose importance is larger the stronger the first stage of the second paper (γ_2^*), and the stronger an instrument Z is. If an econometrician does control for X_2 in OLS, then we remove the (potentially) omitted variable of X_2, but add another confounding term that scales with the amount of common variation X_2 has with Y_1's confounder. As discussed above, common variation may occur if either $\theta_{12} \neq 0$, so that X_1 introduces the confounder directly into X_2, or if the confounder was already present in both error terms.

Instrumental variables estimation does not solve the problem, as the second two rows of Table 1 make clear. Without controls, the omitted variable problem of X_2 is present. Paper 1's implicit assumption that $\beta_{21} = 0$ would imply that the naive regression recovers the true parameter of interest. However, the stronger the effect of Z on the second paper's covariate of interest, defined by (γ_2^*) holding constant the the full effect of Z on X_1 (γ_1^*), the more existing bias is exacerbated. The core observation of this paper is

Table 1 Asymptotic Distributions of Four Estimators

Estimator	Notation	Asymptotic Bias	$N\cdot$Asymptotic Variance
OLS, no controls	$\widehat{\beta_{11}}^{OLS,NC}$	$\beta_{21}\gamma_1^*\gamma_2^*\frac{Var(Z)}{Var(X_1)} + \frac{Cov(X_1,\epsilon_1)}{Var(X_1)}$	$\frac{\sigma^2_{Y\|X_1}}{\sigma^2_{X_1}}$
OLS, controls	$\widehat{\beta_{11}}^{OLS,C}$	$\frac{Cov(X_1,\epsilon_1)}{Var(X_1)-\frac{Cov(X_1,X_2)^2}{Var(X_2)}} + \frac{\frac{Cov(X_2,\epsilon_1)}{Var(X_2)}}{Var(X_1)-\frac{Cov(X_1,X_2)^2}{Var(X_2)}}$	$\frac{\sigma^2_{Y\|X_1,X_2}}{\sigma^2_{X_1}(1-r^2_{X_1,X_2})}$
IV, no controls	$\widehat{\beta_{11}}^{IV,NC}$	$\beta_{21}\frac{\gamma_2^*}{\gamma_1^*}$	$\frac{\sigma^2_{Y\|X_1,Z}}{\sigma^2_{X_1}r^2_{ZX_1}}$
IV, controls	$\widehat{\beta_{11}}^{IV,C}$	$-\frac{\gamma_2^*Cov(\eta_2^*,\epsilon_1)}{\gamma_1^*Var(\eta_2^*)-\gamma_2^*Cov(\eta_1^*,\eta_2^*)}$	$\frac{\sigma^2_{Y\|X_1,X_2,Z}(1-r^2_{ZX_2})}{\sigma^2_{X_1}(r^2_{ZX_1}-r^2_{ZX_2}r^2_{X_2X_1})^2}$

Table 1: Table 1 displays the asymptotic means and variances of four estimators for β_{11} in the system of equations given by Equations 1-3 and displayed graphically in Figure 2. "OLS" denotes simple ordinary least squares regression, and the OLS moments of the asymptotic distributions are included for comparison purposes. "IV" denotes the use of Z as an instrument for X_1. "Controls" denotes inclusion of X_2 as an exogenous control. X_1^* and ξ_1^* denote the residual of X_1 and ξ_1 after being regressed on X_2. $\gamma_1^* = \frac{\gamma_1+\theta_{21}\gamma_2}{1-\theta_{21}\theta_{12}}$, and $\gamma_2^* = \frac{\gamma_2+\theta_{12}\gamma_1}{1-\theta_{21}\theta_{12}}$.

that if a researcher is uncertain about whether or not β_{21} is zero, publication of the second paper is likely to increase our belief that γ_2^* was always large and therefore increase concern. However, controls in an IV regression are no panacea. If X_2 shares unexplained variation with the confounder between X_1 and Y_1 that necessitated IV to begin with, bias is introduced, as we discussed was the case with OLS regression. If X_1 and X_2 share significant amounts of variation, or γ_2^* is large, the instrumental variables regression with controls can be more biased than any of the other regressions we have studied.

The final implication of Table 1 is that the source of bias for IV with and without controls is different. The first is driven by the causal effect β_2, and by the relative strengths of the two first stages. The second is driven by the non-causal shared variation of X_2 with Y_1's, residual of X_1.

A special case of direct violation occurs when past X's are used as instrument's for current X's. A major example of this is Card (2001) and derivative papers. Suppose we wish to estimate the effect of immigrants on local education or local land values. However, identification is threatened by the tendency of immigrants to immigrate to places where wages are high, housing prices are low, and education is high quality per unit cost. This produces a correlation between η_1 and ϵ_1, threatening identification. To obtain accurate estimates, it is thus necessary to find an instrument for current immigration flows that does not affect local education or local land values, or to control for all downstream effects, which seems impossible in the context of immigration.

A traditional solution to this particular problem is to use past immigration as an instrument for current immigration. Bartel (1989) finds immigrants tend to immigrate to places where there are already many immigrants. He argues this is due to "supply side" factors such as culture, linguistic familiarity, or "weak ties" that may be unrelated to "demand side" factors, such as city-level productivity. This reasoning implies current immigrant levels determine a city's exposure to national immigration trends. Thus, if immigration from Mexico increases by 10%, it is plausible that cities with preexisting Mexican immigrant populations will absorb proportionally more of that immigration flow than states with little preexisting population simply because of ethnic enclaves shifting benefits of a location for reasons unrelated to for instance, native wages, producing a valid instrument that increases immigration for reasons other than productivity.

However, the proposed relationship between current immigration flows and past immigration flows is subject to criticism. As outcomes affected by immigration proliferate, so too do the potential X's the econometrician must be concerned about. If we believe that ΔX_t affects Y_t^1 and Y_t^2, by assumption it is also the case that ΔX_{t-1} affected Y_{t-1}^1 and Y_{t-1}^2, which themselves potentially affect Y_t. For a concrete example, consider the case in which immigration is shown to affect highly durable and dynamic outcomes such as housing stock, human capital stock, firm capital stock, and health stock of an urban area, as has been established by a multitude of immigration papers. A researcher might use the stock of Mexican immigrants in Los Angeles interacted with national flows of immigrants as an instrument for Mexican immigrants in 1990. However, because the stock of immigrants in 1980 was itself a product of both a stock and a flow, and immigrants in 1980 affected slow-moving variables such as housing stock and human capital investment, it is reasonable to suspect that those effects are ongoing even fifteen or thirty years later, violating exogeneity.[9]

IV Six Categories of Related Instruments

Because of the sheer magnitude of uses and relevant literature, we provide a condensed discussion of the most concerning or significant re-uses of instrumental variables in this section. A thorough list can be found in Appendix A.

[9] A similar argument is made in Jaeger, Ruist, and Stuhler (2018). Our contribution is to document the sheer weight of papers finding potential long-run effects, and to note that controlling for these outcomes is not sufficient for identification if there is simultaneity between outcomes or they share confounders, both of which appear very likely in the case we outlined.

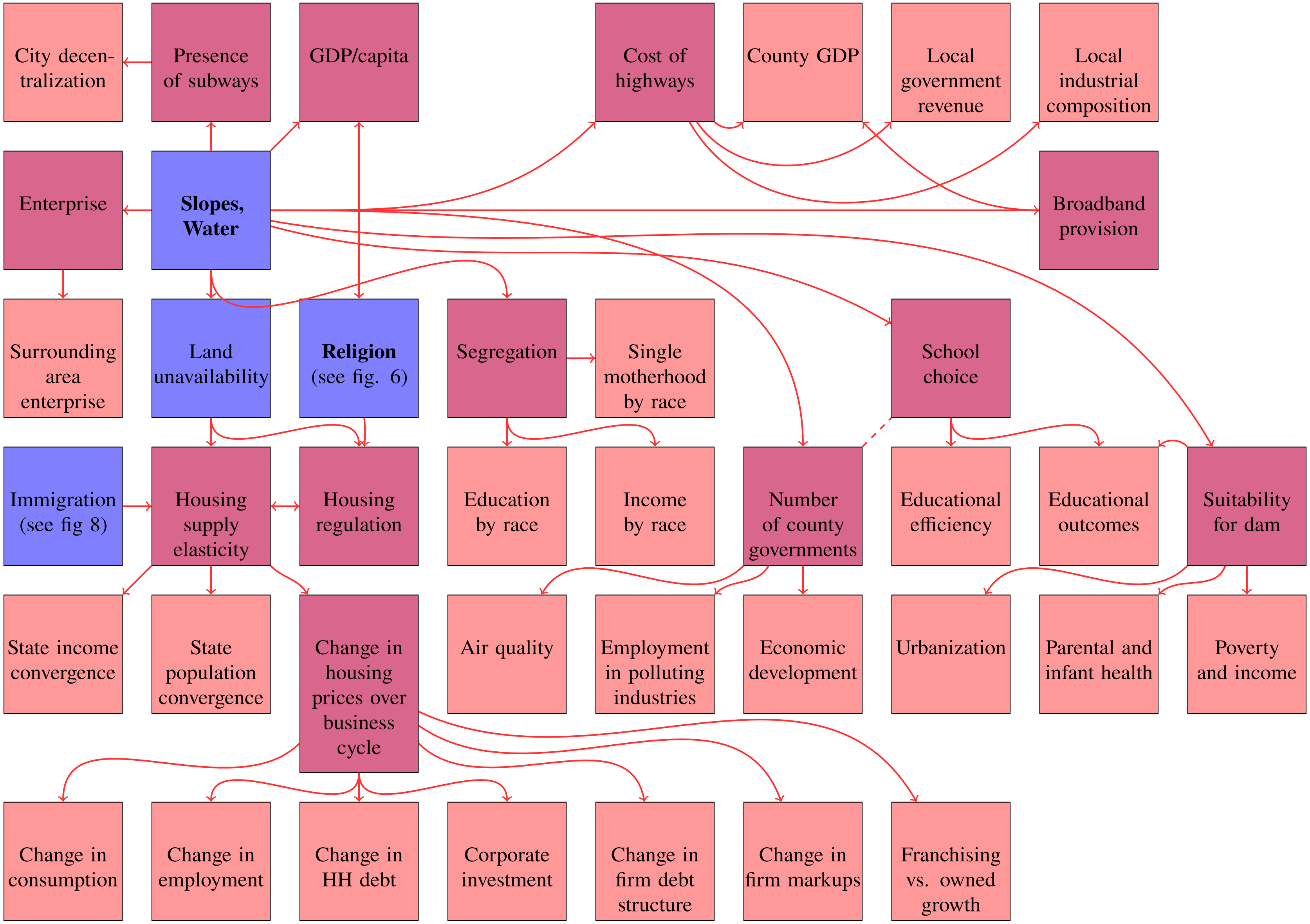

Figure 3: This figure summarizes selected research using elevation and bodies of water as an instrument. Blue are the instruments or measurements created from the presence of elevation changes and bodies of water. In purple are the instrumented-for endogenous covariates. In red are outcomes. A more complete list of 55 paper-instrument-outcomes can be found in Appendix A. Some variables, such as “city decentralization” and “urbanization” or “economic development” “GDP/capita” and “poverty and income” are listed in multiple places for legibility.

Elevation and bodies of water

Changes in elevation and the presence of bodies of water (including rivers and streams) are used as an instrument 18 times in top five papers, and 23 times in other top field or general interest journals. They are used, either implicitly or explicitly, to instrument for approximately fifteen outcomes: (1) change in housing prices (2) change in housing stock (3) city density (4) farming (including income) outcomes (5) enterprise (6) segregation (7) school governance structure (8) number of county governments (9) presence of dams (10) cost of highways (11) broadband provision (12) share of developed land (13) access to international markets (14) access to domestic market center and (15) presence of piped water. These fifteen instrumented covariates are then used to estimate more than thirty-five unique outcomes, from household health, longevity, and fertility, to firm earnings, worker wages, household education decisions, and investment outcomes, to air quality, industrial composition, trade openness, and educational efficiency.

Perhaps the most important recent paper in this literature is Saiz (2010), which connects the presence of elevation and bodies of water to the amount of available land. Using this information, Saiz (2010) produces city-specific estimates of the elasticity of the housing supply. Relevantly for this paper, Saiz also uses religion and climate as instruments in the spirit of a test of overidentification, finding his estimates little changed. A number of top journal articles have used these measures of housing supply elasticities to instrument for actual housing prices' effect on other economic variables such as debt growth (Mian and Sufi, 2011), employment growth (Mian, Rao, and Sufi, 2013), and change in consumption (Mian and Sufi, 2014). In addition to these studies, Saiz also connects local religion with regulation, but local religion has also been linked to education, welfare disability, and marriage and divorce rates (Gruber, 2005). Elevation and bodies of water also influence the construction of dams (Duflo and Pande, 2007), which are not only vital to the world's energy supply but also have been found to be related to local longevity, income, education, infant mortality, and other human capital measures (Lipscomb, Mobarak, and Barilam, 2013). Finally, growth in enterprise has also been found to be more prevalent in areas closer to rivers and waterways and in the immediate surrounding areas (Felkner and Townsend, 2011). Figure 3 shows how the discussed economic variables (as well as selected others omitted for brevity) relate to each other. The sheer volume of interconnected variables likely at the center of a complex causal web seems to be cause for concern.

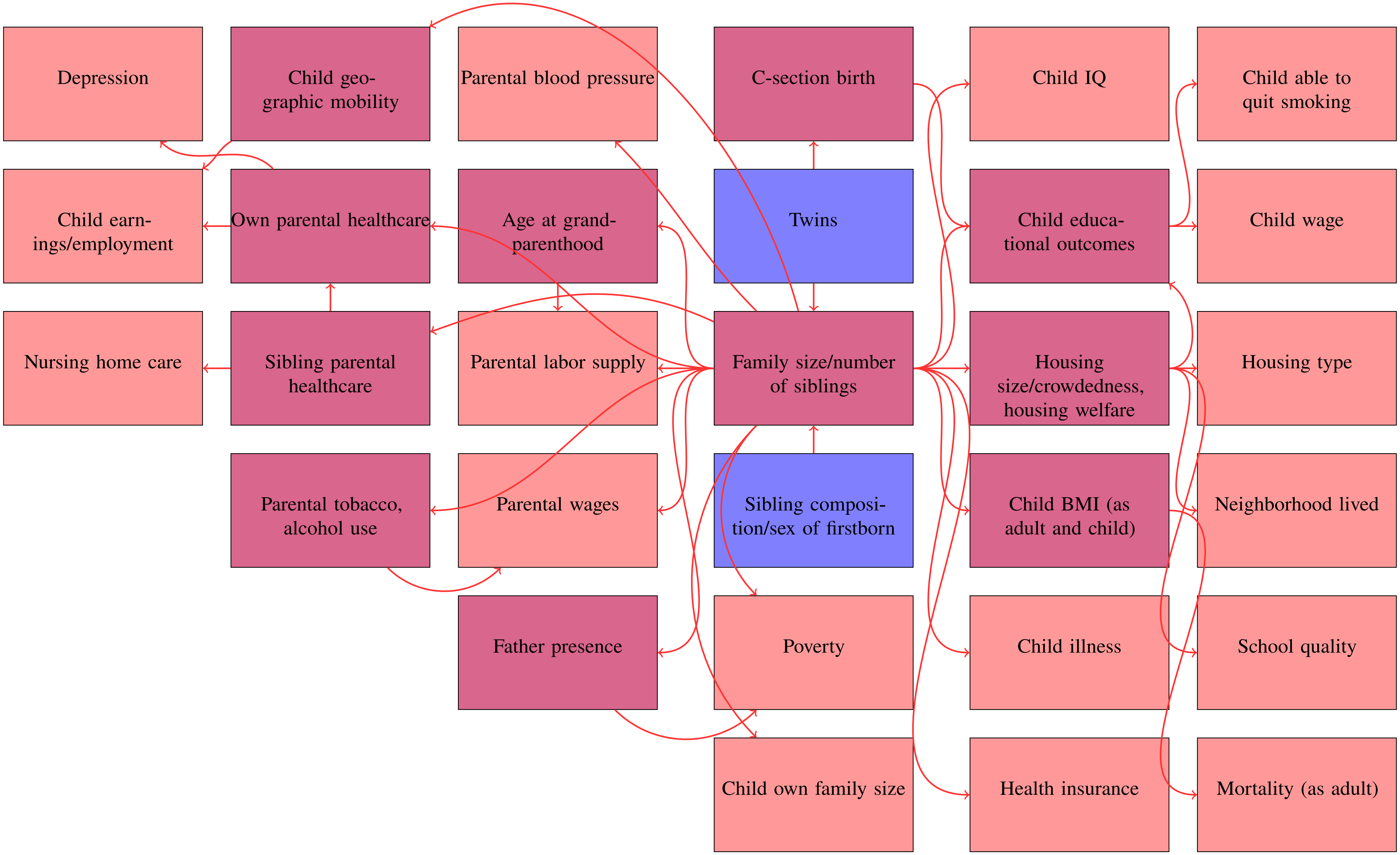

Figure 4: This figure summarizes selected research using sibling structure. Blue are the instruments or measurements created from sibling age and sex structure. In purple are the instrumented-for endogenous covariates. In red are outcomes. Some papers use family size and sibling structure directly to instrument for other covariates of interest, so it is also linked to endogenous covariates.

Sibling structure

Frequently available in any dataset on the family, the age and gender composition of an individual's siblings is a popular instrument, and is used 12 times in top 5 papers and 36 times other well-ranked journals. Gender mix in particular is used to instrument for the effect of women's education on earnings as well as for the effect of family size on academic performance (Conley and Glauber, 2006), income, health, insurance, blood pressure, obesity (Cáceres-Delpiano and Simonsen, 2012), and child BMI and illness (Palloni, 2017). Moreover, gender mix is also used to instrument for welfare generosity which also impacts the neighborhood a child lives in, their corresponding school quality, and whether or not a grade is repeated (Currie and Yelowitz, 2000). Gender mix is also used to instrument for parent's marital status which can affect a multitude of economic outcomes (Ananat and Michaels (2008); Dahl and Moretti (2008)).

The number of siblings in a family unit is used as an instrument for education's effect on wages (Levin and Plug, 1999; Taber, 2001; Korpi and Tåhlin, 2009), for schooling's impact on the decision to quit smoking (Sander, 1995), and for the effect of migration on earnings (Ziliak and Kniesner, 1999). The presence of twins is also used as an instrument to estimate a variety of effects such as: family size on divorce rates (Jacobsen, Pearce, and Rosenbloom, 2001), family size on a child's personality (Fletcher and Kim, 2019), family size on a child's own family size (Kolk, 2015), and more. These effects and selected others are shown in Figure 4. Many of these uses, such as the likelihood siblings will be present to take care of an elderly parent, child geographic mobility, and parent's age at grand-parenthood are likely to affect decision-making such as education earlier in the lifecycle via dynamic optimization. Moreover, while a household member may know which sibling is likely to live near parents when older and provide care, that information is far less likely to be available to researchers. These facts raise questions regarding the use of variables related to sibling structure as instruments.

Ethnolinguistic fractionalization and language

Ethnolinguistic fractionalization (ELF) and similar concepts, including ethnic fractionalization, linguistic fractionalization, ethnolinguistic polarization, or fraction speaking a European language, are frequently used to examine the effect of "institutions" or "governance" on relevant economic outcomes, most importantly growth.[10] ELF and related concepts are used 4 times in top 5 papers and 37 times in other,

[10] While fraction speaking English or another European language is potentially different in concept, the correlation between speaking a European language and ethnolinguistic fractionalization is -0.29 (taken from Hall and Jones (1999) and Bazzi and

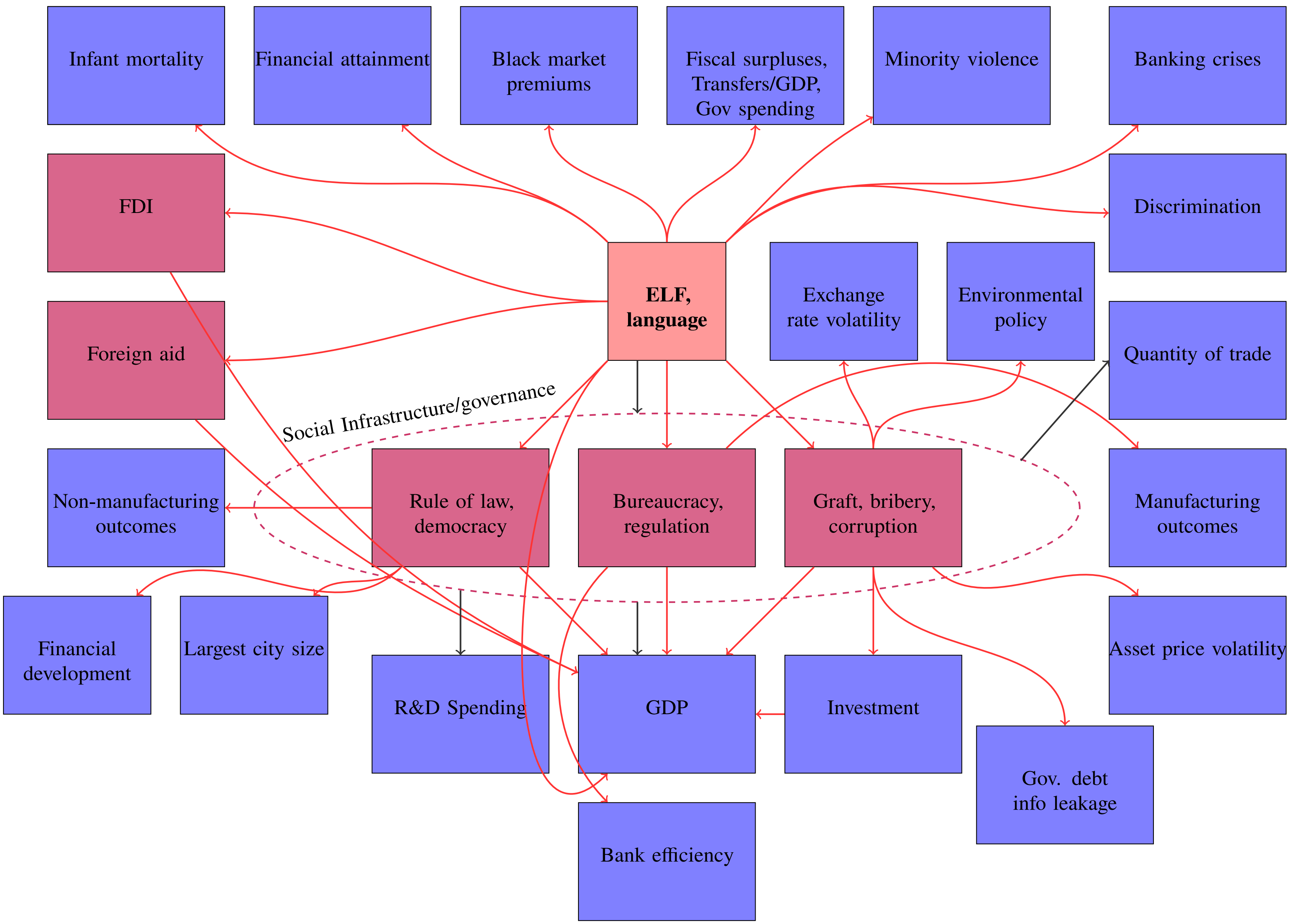

Figure 5: This figure summarizes selected research using ethnolinguistic fractionalization as an instrument. Blue are the instruments or measurements created from ELF. In purple are the instrumented-for endogenous covariates. In red are outcomes. Because ELF is used for both agglomerations of (1) rule of law/democracy, (2) bureaucracy and (3) regulation and graft, bribery and corruption, as well as each individually, we depict with black lines generating from a dashed circle outcomes that use agglomerations (“social infrastructure/governance”).

well-regarded journals. However, governance is a broad concept, and these ethno-linguistic measures are used to discuss multiple distinct but interrelated concepts. Usefully, Kaufmann, Kraay, and Zoido-Lobatón (1999) define three distinct components of governance: (1) the rule of law, (2) bureaucratic efficiency/effectiveness, and (3) graft/bribery.

Mauro (1995) uses ELF to instrument, in different regressions, for (1) corruption's effect on investment and growth and (2) for bureaucratic efficiency's effect on investment and growth, where bureaucratic efficiency includes judiciary efficiency, red tape, and corruption. As an instrument for corruption alone, ELF has been used by Michaelides et al. (2015) and Michaelides, Milidonis, and Nishiotis (2019) to instrument for a country's Transparency International's "Corruption Perceptions Index," finding higher corruption is linked to information leakages before government debt downgrades and high volatility of asset prices and exchange rates. LaPorta et al. (1999) find ELF is negatively associated with a wide range of government performance inferiority, from property rights and regulation to corruption, delays, tax compliance, public goods, and government intervention. Two less causally-focused papers relate ELF to an enormous slew of variables including school attainment, financial attainment, black market premiums, fiscal surpluses, infrastructure spending, discrimination, and minority violence (Easterly and Levine, 1997), and also banking crises, property rights, business regulation, transfers and subsidies as a fraction of GDP, democracy and political rights, and infant mortality (Alesina et al., 2003). Rodrik, Subramanian, and Trebbi (2004) use fraction speaking English, among other variables, to instrument for both the rule of law and for market integration, finding per-capita GDP is primarily determined by institutions rather than integration. Ades and Glaeser (1995) use ELF, along with other variables, as an instrument for a country being under both dictatorship and its trade policies, which affects main city size. Ultimately, ELF's correlations with many interconnected variables may give a researcher pause in interpreting causal channels. Figure 5 shows the relationships between ELF and the variables discussed above as well as additional relationships listed in Appendix A.

Religion

Like ethnolinguistic fractionalization, national, local, or personal religion is a frequently-measured and unambiguously important factor for numerous economic outcomes. Because religion shapes culture, it is frequently used both as an instrument for many endogenous covariates and as a control. For instance, local

Clemens (2013) data). In a regression of fraction speaking a European language on ethnolinguistic fractionalization, for every 1% more fractionalized a society becomes, it is -0.4% less likely to speak a European language.

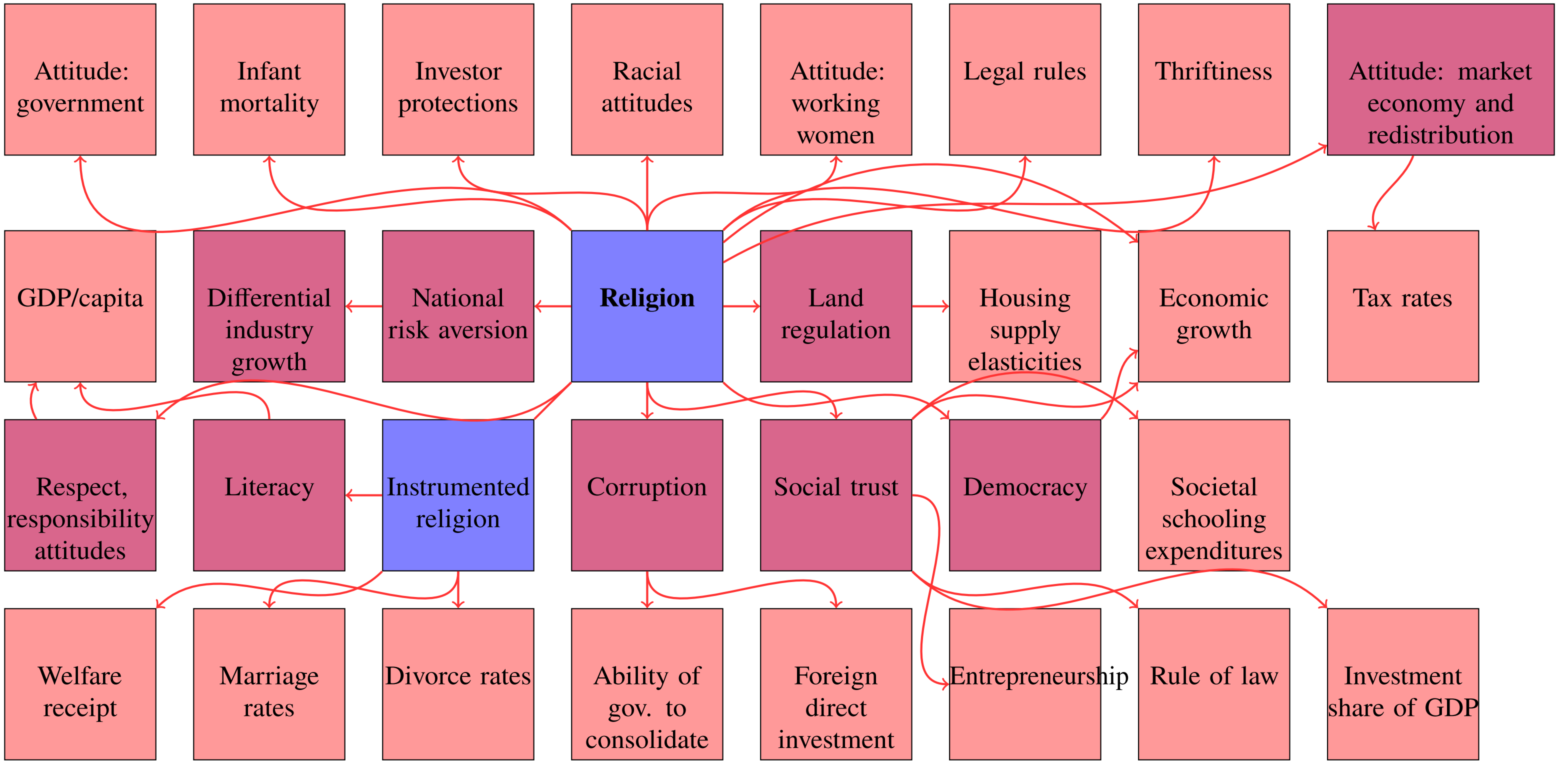

Figure 6: This figure summarizes selected research using religion as an instrument. Blue are the instruments or measurements created from religion. In purple are the instrumented-for endogenous covariates. In red are outcomes. Many papers link it correlationally to outcomes, which provide potential causal tests of IV papers. Many micro papers instrument for individual religion, and find it correlated with a number of important outcomes. While these papers may not be threatened by other papers in this graph, they may threaten papers and are included for that reason.

religious background is used to instrument for regulation in Saiz 2010's commonly-used supply elasticity estimates. More generally, it appears in the literature seven times in top 5 papers and 37 times in other well-regarded journals. Though papers historically have been relatively careful about causality, religion has been suggestively and intuitively linked to a number of outcomes. Guiso, Sapienza, and Zingales (2003), while "well aware of the difficulty in interpreting the observed correlation as causal effects," find religion is linked to an enormous number of outcomes: attitudes toward a variety of topics, such as cooperation, government, working women, the market economy, and racism, but also legal rules and societal thriftiness. Even when religion's causal effects are not a focus, it is linked as a control to a variety of important outcomes such as: property rights, regulation, tax rates, corruption, bureaucratic delays, the size of government, public rights, illiteracy, schooling, and democracy (LaPorta et al., 1999), investor productions (Stulz and Williamson, 2003), economic growth (McCleary and Barro, 2006), infant mortality (Brainerd and Menon, 2014), and more. While our point does not threaten the causal identification of these papers, their use of religion as a control suggests covariance between religion and their outcomes of interest, potentially threatening the use of religion as an instrument in other papers. Religion is also used to instrument for the effect of social trust on schooling expenditures (Bjørnskov, 2012) , the effect of societal respect on GDP per-capita (Mobarak, 2005), the effect of national uncertainty aversion on differential industry growth (Huang, 2008), and more.

To further complicate matters, religion is also often *instrumented for*. To name just a few, Gruber (2005) instruments local religiosity with area ancestry, Becker and Woessmann (2009, 2008, 2018) uses distance to Wittenberg as an instrument for Protestantism, and Waldinger (2017) uses the initial missionary treks into Mexico to instrument for religious (mission) presence. When religion is instrumented for validly, this increases the concern for other non-instrumented uses, and so we include these papers. Figure 6 shows the complex interconnected framework implied by the literature for this instrument for the variables discussed as well as additional variables and relationships further listed in Appendix A. The variety and number of documented relationships using this group of variables suggests a single causal channel is unlikely.

Weather

Weather, as distinct from climate, is frequently used as an instrument for a plethora of endogenous covariates, appearing in 30 top five publications and 40 other well-regarded publications. Mellon (2021) documents its further use 176 times in economics and political science journals, many of which were not considered here. Perhaps most prominently, rainfall is used as an instrument for income, which affects the likelihood of conflict (Miguel, Satyanath, and Sergenti, 2004; Sarsons, 2015). It is also used as an instrument for growth or local income shocks, which affects local witch killings (Miguel, 2005), land invasions (Hidalgo et al., 2010), democratic change (Burke and Leigh, 2010; Brückner and Ciccone, 2011), consumption (Kazianga and Udry, 2006), remittances (Arezki and Brückner, 2012), sale of durable investment goods (Fafchamps, Udry, and Czukas, 1998), trade balance (Brückner and Gradstein, 2013), urbanization (Brückner, 2012), manufacturing output, employment and capital investment (Lee, 2018) and the rate of time preference of households (Tanaka, Camerer, and Nguyen, 2010; Di Falco et al., 2019).[11]

Other weather phenomenon besides rainfall are used as instruments as well such as: sky cover as an instrument for the effect of managerial expansion beliefs (moods) on hiring and capital investment (Chhaochharia et al., 2018), lagged weather as an instrument for the effect of dry/cold conditions on virus transmissions (Adda, 2016), and rainfall variance as an instrument for the effect of land concentration on banks per capita (Rajan and Ramcharan, 2011). With so many potential avenues for affecting households and nations, it is perhaps no surprise that rainfall is correlated with a number of short-run and long-run effects. We extend the concern of Sarsons (2015) about rainfall's use as an instrumental variable for income by documenting fifteen other uses of rainfall as an instrument that may help explain her results, as well as noting more than thirty other outcomes that may also be affected by rainfall. Figure 7 shows the myriad of weather-related empirical relationships, which at best muddles casual interpretation.

Immigrant enclaves

Immigrant enclaves are used as instruments in 12 top five papers and 58 other well-regarded articles. Its use as an instrument begins with Bartel (1989)'s documentation that immigrants appear to migrate to locations where other immigrants have already located. Altonji and Card (1991) use historical immigrant share by city as an instrument for new immigrant shares by city, which affect wages. The instrument took on its more classical form in Card (2001), which used historical immigration patterns to predict (instrument for) immigrant inflows, which affect native outflows, employment/population ratios for natives and immigrants, and wages.

Importantly, the instrument is not valid under serial correlation: if, for instance, a city underwent a permanent and unmitigated productivity shock that brought both past immigrants and current immigrants

[11] Sarsons (2015) finds rainfall does not affect the agricultural production of districts in India that were downstream of dams, but that conflict in these districts measured via ethnic riots persisted, suggesting conflict was caused by another channel other than income shocks.

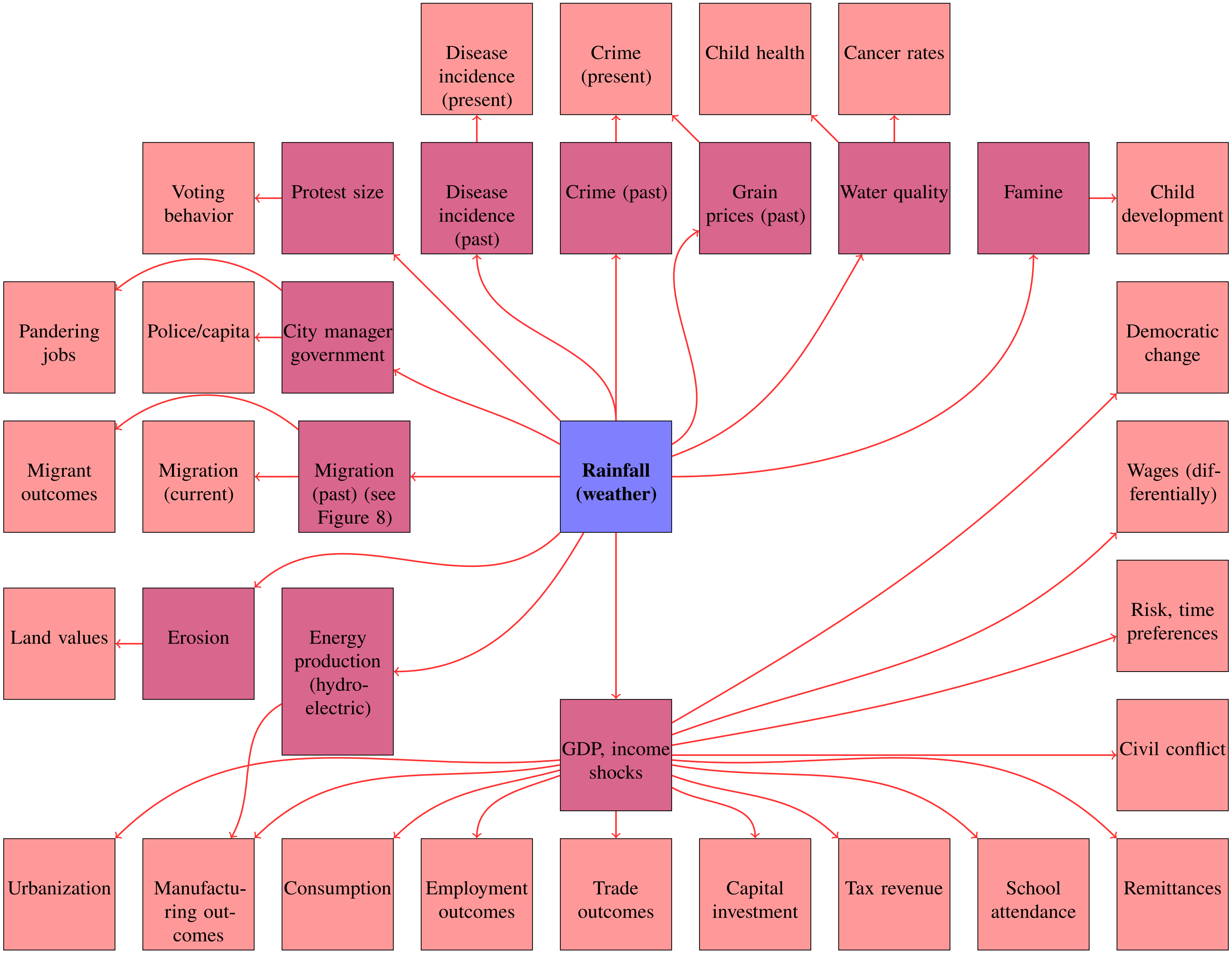

Figure 7: This figure summarizes selected research using weather as an instrument. Blue are the instruments or measurements created from weather. In purple are the instrumented-for endogenous covariates. In red are outcomes.

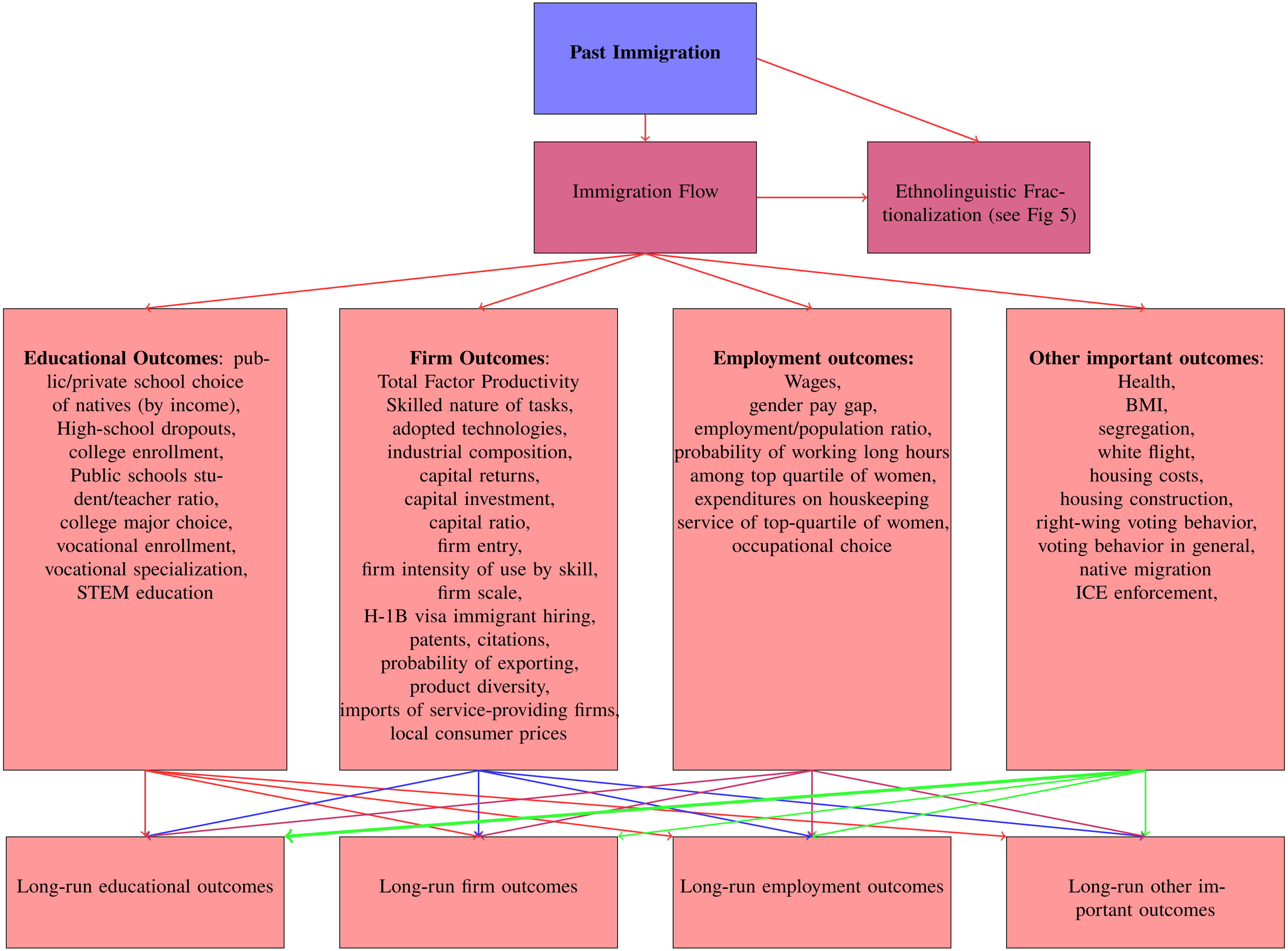
Past Immigration
Immigration Flow
Ethnolinguistic Fractionalization (see Fig 5)
Educational Outcomes: public/private school choice of natives (by income), High-school dropouts, college enrollment, Public schools student/teacher ratio, college major choice, vocational enrollment, vocational specialization, STEM education
Firm Outcomes: Total Factor Productivity Skilled nature of tasks, adopted technologies, industrial composition, capital returns, capital investment, capital ratio, firm entry, firm intensity of use by skill, firm scale, H-1B visa immigrant hiring, patents, citations, probability of exporting, product diversity, imports of service-providing firms, local consumer prices
Employment outcomes: Wages, gender pay gap, employment/population ratio, probability of working long hours among top quartile of women, expenditures on houskeeping service of top-quartile of women, occupational choice
Other important outcomes: Health, BMI, segregation, white flight, housing costs, housing construction, right-wing voting behavior, voting behavior in general, native migration ICE enforcement,
Long-run educational outcomes
Long-run firm outcomes
Long-run employment outcomes
Long-run other important outcomes

Figure 8: This figure summarizes selected research using immigrant enclaves as an instrument.

to a city while simultaneously increasing wages of natives, then the instrumented effect of immigrants on native wages would be misstated. However, as durable and simultaneous outcomes proliferate, serial correlation becomes likely. Unlike other instruments in this paper, the past presence of immigrants is typically only used to instrument for the flow (or new stock) of immigrants today. However, it is found to be connected with a number of highly durable outcomes including: education (Card and Lewis, 2007; Farre, Ortega, and Tanaka, 2018; Orrenius and Zavodny, 2015; Røed and Schøne, 2016; Shih, 2017), immigrant women's labor force participation (Adda, Dustmann, and Stevens, 2017), long-run child outcomes (Havnes and Mogstad, 2015), exports (Parrotta, Pozzoli, and Sala, 2016; Ottaviano, Peri, and Wright, 2018), capital investment by firms (Baum-Snow, Freedman, and Pavan, 2018; Lafortune, Lewis, and Tessada, 2019), low skilled wages (see Card (2001)), and more. Figure 8 shows these relationships as well as other relationships listed in Appendix A. The staggering number of persistent outcomes associated with this instrument raises concerns regarding its validity, as these durable channels likely make serial correlation a genuine threat to identification. Moreover, if immigration has causal effects on ethnolinguistic fractionalization or religion, those areas of research become important to consider.

V Testing for Invalid Instruments

Much of the evidence we discuss in the preceding section is descriptive, and relies on sheer volume of causal connections established in the literature to raise concern about an instrument's use. But some uses may be valid, even if others are not. This section builds on the literature review and offers a way to use it not only qualitatively but quantitatively by proposing a new Hausman-like test for instrumental validity, which we apply in the next section.

Table 1 showed there were two distinct sources of bias for the uncontrolled and controlled instrumental variables regressions. When not controlling, the bias term was given by $\beta_{21}\frac{\gamma_2^*}{\gamma_1^*}$, the direct violation weighted by the strength of the two first stages. With controls, it was given by $-\frac{\gamma_2^* Cov(\eta_2^*,\epsilon_1)}{\gamma_1^* Var(\eta_2^*)-\gamma_2^* Cov(\eta_2^*,\epsilon_1)}$, which is affected by simultaneity between endogenous covariates which loads η_2^* with variation contained in ϵ_1 or initial covariance in the error terms of our outcome of interest and proposed controls. Our test takes advantage of the fact that these two are "independent," in the sense that they are causally and mathematically distinct, so that a precise failure to reject suggests both biases are small or zero.

Formally, we propose a test similar to a Durbin-Wu-Hausman test (Hausman, 1978), testing whether or not the coefficients of the “uncontrolled” and “controlled” regressions are the same.[12] One important distinction between the two: under our test’s null hypothesis, both estimators are consistent (with no assumption on efficiency). Under the alternative hypothesis, at least one is not consistent (potentially both), so that:

$$H_0 : \beta_{11}^{IV,NC} = \beta_{11}^{IV,C}$$

$$H_a : \beta_{11}^{IV,NC} \neq \beta_{11}^{IV,C}$$

The idea behind our test is intuitive. As can be seen from in the third column of Table 1, uncontrolled and controlled biases are driven by different covariances. Biases in the controlled case are driven by direct violations of exogeneity, X_2 affecting Y_1. Biases in the uncontrolled case are driven by “indirect" violations of exogeneity, such as when X_1 and X_2 affect one another, or X_2 shares confounding variation with X_1 and Y_1. Because the two biases are different, if one or the other is large, then our test will reject. Even if both are large, they are unlikely to be the same asymptotically. The contrapositive is also true: if our test fails to reject, then both biases were likely small.

Our test is asymmetrically useful for a researcher. A failure to reject the null is very informative for a researcher, and their path forward is clear. However, a rejection may be due to many reasons: either one, or both of the estimators is not consistent, but it does not specify which. This property is shared by many statistical tests, including the ubiquitous Leamer (1983)-style robustness checks, in which a researcher includes and does not include various coefficients as exogenous controls. In such cases, failure to reject leads to no further action, while rejection requires substantive argumentation on why the two regressions give different coefficients. Similarly, for the Sargan-Hansen test, in which a failure to reject is interpreted to mean the excluded instruments are exogenous, while a rejection may indict either the excluded or the included instruments (or both). In our test, Leamer-style robustness checks, or Sargan-Hansen tests, a rejection of the null, while less useful than a failure to reject, highlights the need for increased awareness and discussion of the problem: if one estimate is believed to be valid, a clear defense of it against the alternative estimate is necessary. We note that even in the case of rejection, the magnitude of the test statistic, and not just its p-value, is important in this substantive discussion: coefficients whose difference is statistically significant but economically unimportant may be less concerning than an economically meaningful difference.

[12] While our literature review provides a list of controls a researcher should use, some may be inappropriate. For instance, a second paper may use a first paper’s instrument to instrument for the first paper’s outcome, rather than endogenous covariate. In such a case it would obviously be inappropriate to control for the outcome. Nevertheless, we believe highlighting why a potential control is excluded is a useful disciplining exercise.

While researchers routinely compare estimates with different sets of controls informally, our contribution to this practice has three components. First, this paper's literature review and Appendix Tables explicitly pre-specify a set of relevant controls that have not been adopted by the literature. Second, in practice robustness checks are joint tests of coefficient equality and multicollinearity, in that an estimator is likely to be informally rejected due to the inclusion of controls if either the coefficient changes or standard errors grow to preclude significance. Our test justifies only the first rejection, but not the second. Finally, there is no explicit model other than a concern for omitted variables in the standard robustness check, while our test discusses distinct sources of bias, which allows for easier interpretation when rejected.

While our estimator takes the form of a Hausman-like test, we cannot adopt the assumption that either of our estimators is efficient, which is needed to show that the covariance between estimators is equal to the variance of one of the estimators in the Hausman test. Instead, we derive the covariance of estimators in our framework and find that it is likely to be large, particularly in small samples. This mirrors our applications later in this paper. The estimated efficient asymptotic variance-covariance matrix of $\widehat{\beta}^{IV,NC}$ and $\widehat{\beta}^{IV,C}$ can be estimated in a standard stacked/multi-equation GMM framework, and is given by equation 4:

$$\widehat{Cov}(\widehat{\beta}^{IV,stacked}) = \left(\begin{bmatrix} \Sigma_{XZ,C} & 0 \\ 0 & \Sigma_{XZ,NC} \end{bmatrix}' \begin{bmatrix} Z_C'\hat{\epsilon}_1\hat{\epsilon}_1{}'Z_C & Z_C'\hat{\epsilon}_1\hat{\epsilon}_2{}'Z_{NC} \\ Z_{NC}'\hat{\epsilon}_2\hat{\epsilon}_1{}'Z_C & Z_{NC}'\hat{\epsilon}_2\hat{\epsilon}_2{}'Z_{NC} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{XZ,C} & 0 \\ 0 & \Sigma_{XZ,NC} \end{bmatrix} \right)^{-1} \tag{4}$$

Where $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ are the estimated error terms from the controlled and not-controlled regressions, $\Sigma_{XZ,i}$ $i \in \{NC, C\}$ is the variance-covariance matrix of X and Z for the relevant set of controls. Examining the estimated covariance of $Cov(\widehat{\beta}_1^{IV,C}, \widehat{\beta}_1^{IV,NC})$, in the two-endogenous outcome case, we have:

$$\widehat{Cov}(\widehat{\beta}_1^{IV,C}, \widehat{\beta}_1^{IV,NC}) = Cov(\hat{\epsilon}_1, \hat{\epsilon}_2)\frac{Cov(Z, X_2)^2 - Var(X_2)Var(Z)}{Cov(Z, X_1)(Cov(X_1, X_2)Cov(Z, X_2) - Cov(Z, X_1)Var(X_2))} \tag{5}$$

The covariance of $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ is likely to be substantial and positive. First, any true noise term ϵ in Y_1 is unlikely to be significantly partialled out by either regression. Second, residual covariance between η and ϵ may remain after the IV procedure, which may happen in the presence of correlated error terms ($\epsilon_1, \eta_1, \eta_2$ all correlated) or due to a direct violation of exogeneity, (both $\hat{\epsilon}$'s contain common variation from η_1).

With this covariance in hand, we can write the distribution of the test statistic:

$$\left(\widehat{\beta}^{IV,NC} - \widehat{\beta}^{IV,C}\right) \sim \mathcal{N}\left(\beta_{21}\frac{\gamma_2^*}{\gamma_1^*} + \frac{\gamma_2^* Cov(\eta_2^*, \epsilon_1)}{\gamma_1^* Var(\eta_2^*) - \gamma_2^* Cov(\eta_2^*, \epsilon)}, Var(\widehat{\beta}^{IV,NC}) + Var(\widehat{\beta}^{IV,C}) - 2Cov(\widehat{\beta}^{IV,NC}, \widehat{\beta}^{IV,C})\right) \tag{6}$$

where $Var(\widehat{\beta}^{IV,NC})$ and $Var(\widehat{\beta}^{IV,C})$ are given in Table 1 and the covariance is given in Equation 5. Equation 6 highlights that in order for the mean of the difference between the estimators to be zero under the alternative hypothesis, there would have to be a significant coincidence of biases between covariance of Z with X_2 on one hand, and Z and the residual variation of the first paper's confounder that is orthogonal to X_2 in the other. This, combined with Equation 5 also highlights the difference between the covariance and either of the variances in Table 1.

For the reasons discussed in Hausman (1978) (pp. 1255), allowing for a nontrivial covariance between the two estimators leads to a situation in which power functions cannot easily be calculated in closed form. The strongest practical limitations to our test are likely to come from regressions with large standard errors relative to effect size. Young (2019) documents that, in top published IV work, IV standard errors are typically nearly five times larger than those of OLS, and partially as a consequence, the 95 percent confidence intervals of IV include the OLS point estimate roughly eighty percent of the time. Fortunately, the variance of the difference between estimators is typically one-fifth to one tenth the size of the variance of the estimator itself in our two examples, suggesting power is less likely to be a concern than univariate standard errors would suggest. Regressions driven entirely by unexplained outliers will be unchanged by controls, so variance of the difference will be small even when the variance of each estimator is large.

Another concern is that our test could be used in a model selection procedure.[13] That is, it may be the case that author(s) decide to adopt the uncontrolled estimator if our test fails to reject that there is a difference between the uncontrolled and controlled estimators, but to adopt the controlled estimator and argue their model is valid conditional on controls if our test suggests the two are likely different. Consequently, as is the case for any robustness check used for model selection, our test would then be subject to the critique of Leeb and Pötscher (2005). Leeb and Pötscher (2005) argue the omission of a data-driven model selection procedure will have an ambiguous effect on the mean square error and will depend on the specific data generating process. It should be emphasized, however, that this difficulty is not unique to our test used in this way; any model selection procedure will encounter similar tradeoffs. We note that the model selection procedure described above will be particularly vulnerable to this when the uncontrolled estimator is correct, the controlled estimator is not, and standard errors are small.

While we have assumed treatment effect homogeneity in our baseline, the asymmetry of our test's interpretation allows it to remains useful in the presence of heterogeneous treatment effects. Under treatment effect heterogeneity, binary IV without controls will estimate the local average treatment effect

[13] We thank an anonymous referee for bringing this to our attention

(LATE) under standard assumptions (SUTVA, independence, exclusion restriction, and monotonicity). IV with controls under similar conditional assumptions may estimate a variety of outcomes, depending on whether the instrument is interacted with the control, is “saturated" and “weighted" (Angrist and Imbens, 1995), or Abadie (2003) kappa weighting is used. If the two are able to estimate the same weighted LATE, then our test works as before. Even if IV with controls is only used to produce a differently-weighted average of treatment effects (as in standard 2SLS, for instance), a failure to reject would confirm that not only are biases small, but heterogeneity is unimportant relative to standard errors. While failure to reject is now more informative to a researcher, rejection is now less informative. One additional possibility is that both estimators are valid, but are estimating differently-weighted treatment effects. While Abadie kappa weighting may alleviate this concern, we admit that rejection is even less definitive than under treatment effect homogeneity.

Another concern may be that not all of the potentially endogenous covariates are available for the test. Fortunately, this too does not pose a problem, again due to the asymmetric nature of failure to reject and rejection in our test. For instance, if we add a new potentially endogenous covariate X_3 symmetrically into the system of equations given by (1a)-(3), but exclude it from our controls, then the controlled and uncontrolled estimators each have a single new term, but otherwise do not change:

$$\widehat{\beta}^{IV,NC} = \beta_1 + \frac{\gamma_2^*}{\gamma_1^*}\beta_2 + \frac{\gamma_3^*}{\gamma_1^*}\beta_3 \quad \text{and} \quad \widehat{\beta}^{IV,C} = \beta_1 - \frac{\gamma_2^* cov(\eta_2^*, \epsilon_1)}{\gamma_1^*\sigma_{\eta_2}^2 - \gamma_2^*\sigma_{\eta_1,\eta_2}^2} + \frac{\gamma_3^*\beta_3 - \gamma_2^*\beta_3(\gamma_2^*\gamma_3^*\sigma_Z^2 + \sigma_{\eta_1,\eta_2}^2)}{\gamma_1^*\sigma_{\eta_2}^2 - \gamma_2^*\sigma_{\eta_1,\eta_2}^2}$$

In such a case, failure to reject therefore remains informative, while a rejection may be due to either X_2 or the hidden effects of X_3 that are only partially controlled for by X_2, with their joint dependence on Z likely to generate nonzero covariance.

VI Monte Carlo Exercise

We conduct Monte Carlo tests of the multiple-use instrument problem. Consider the system given by equations 1-3, but allow for any number of new papers with endogenous covariate of interest X_j, $j \in \{2..m\}$, rather than $j = 2$ only, so that:

$$X = Z\Gamma + X\Theta + \eta$$

$$Y = X\beta + \epsilon$$

Where X, η, Y, and ϵ are nxm matricies, Z is $nx1$, Θ and β are mxm, and Γ is 1xm, where n is the number of observations, common for all regressions for convenience.

Recall that the "direct" violation for the first variable of interest is controlled by the first row of β, while the "indirect" violation is controlled through the first row of Θ^{-1} and the observations of the variance-covariance matrix corresponding to the covariances between ϵ_1 and η_j, $j \neq 1$. For clarity, we run three exercises. In the first case, after producing all coefficients, we assume that the matrix of β's is diagonal so that there is no direct exogeneity violation, meaning instrumenting for X using Z ignoring other X_j's is appropriate. In the second case, after producing all coefficients, we assume the variance-covariance matrix for errors in equation 3 is zero except for $\sigma^2_{\eta_i}$, and $\sigma_{\eta_i \epsilon_i}$, and that $\Theta = 0$ for all entries, so that controlling for X_j, $j \neq 1$ is appropriate and will recover the coefficient of interest. In the third case, we report our full calibration, with no sparsity imposed on β or the variance-covariance matrix of errors, save for Z's orthogonality. In such a case no consistent estimator is available to a researcher.

We therefore need to calibrate the parameters in Γ, Θ, β, as well as the $2mx2m$ covariance matrix for η and ϵ.[14] Our choice of parameterization is driven by several goals. First, there is a large confounder $cov(\eta_i, \epsilon_i)$ that significantly biases OLS estimates of β_{ii}. Second, the instrument Z displays significant variation and has a large effect of X, so that the first-stage excluded instrument F-statistic is likely to be far above ten. These first two are consistent with assumptions in all proposed standalone papers. Third, the direct exogeneity violation caused by β_{j1} is mean zero but displays significant variation. Third, the induced exogeneity violation caused by $\sigma_{\eta_j \epsilon_1}$, $j \neq 1$ and θ_{ij} is likely small. Our Monte Carlo distributions and coefficients are given in Table 2.

We assume that Z is distributed normally with mean zero and variance one. γ, which gives the effect of Z on X is distributed normally, with mean one and variance two. Combined with our other assumptions, this yields a distribution of excluded F-statistics between 400 and 500 with 1000 observations. The off-diagonal elements of Θ which control simultaneity in X, are assumed to be distributed normally and independently with mean zero and standard deviation of 0.1. The diagonal elements are zero.

$\beta_{1,1}$, the coefficient of interest, is fixed at unity. Off-diagonal elements of β, controlling the "direct" violation, are distributed normally and independently, with mean zero and standard deviation 0.1, bounding the contribution of any single direct violation to be relatively small compared to the coefficient of interest.

Finally, the covariance matrix for η and ϵ is assumed to be distributed Wishart with degrees of freedom 2000. This is chosen so the covariance matrix Σ is positive semidefinite. Σ is normalized by 2000, do

[14] We assume, as in Equation 3, that the covariance of Z with all other elements is zero, and therefore do not include it for notational convenience.

that its entries are the expected variance-covariance matrix, while maintaining a positive-semidefinite covariance matrix and also allowing degrees of freedom to inject noise. The diagonal of Σ are chosen so that the expected variance of η_i is 10, and the expected variance of ϵ_i is 40. The off-diagonals are chosen so that the expected covariance between ϵ_i and η_i is five. All other off-diagonal entries of Σ are zero, reflecting significant independence between X's.

Table 2: Distribution of Monte Carlo Parameters

Description	Variable	Distribution
Coefficients		
Main instrument strength	γ_j	$\mathcal{N}(0,2)$
Effect of X's on one another	$\theta_{ij}, i \neq j$	$\mathcal{N}(0,0.05)$
Effect of X's on itself	θ_{ii}	0
Main effect of interest, X_1 on Y_1	β_{11}	1
Magnitude of X's on Y_1's	$\beta_{1j}, i \neq j$	$\mathcal{N}(0,0.03)$
Error Terms and Z		
Distribution of instrument	Z	$\mathcal{N}\left(0,1\right)$
Distribution of errors	$\begin{bmatrix} \eta \\ \epsilon \end{bmatrix}$	$\mathcal{W}\left(\Sigma, 2500\right)/2500$
Entries of Σ		
Diagonals of η entries in Σ	$\sigma^{\eta}_{i,i}$	10
Diagonals of ϵ entries in Σ	$\sigma^{\epsilon}_{i,i}$	40
Diagonals of η, ϵ entries in Σ	$\sigma^{\eta,\epsilon}_{i,i}$	5
Off-diagonal entries in Σ	$\sigma_{ij}, i \neq j$	0

Table 2: This tables displays the parameters governing the Monte Carlo distribution corresponding to an N-dimensional version of equations (1)-(3). Signs of γ_i and β_{ii} $i \neq 1$ are flipped negative with 50% probability. The sign of all coefficients but β_{11} is taken from a Bernoulli distribution with probability 0.5.

We limit our interest to estimated coefficients that are between -8 and 10 (the true coefficient is one), simulating an estimation selection process in which a researcher discards an IV if it generates nonsensical results.[15] Our proposed filter for estimators is to use the uncontrolled IV estimators only when that estimator is not statistically significant from the controlled estimator at the 10% level. We then compare the estimates that pass our filter against the distribution of OLS without controls, as well as IV with and without controls. We show that our test contains useful information: when an estimate passes our filter, it is likely to have low mean-square error. Perhaps more importantly, if either of the two

[15] Dropping unusual estimates is certainly done in empirical research. For instance, Cho and Rust (2017) report considering using interest-free loan installment offers as an instrument for interest rates on consumer demand, but rejected the instrument when it yielded an upward-sloping demand curve. Failure to do this for the IV estimator in our trials yields ridiculous results. For instance, while the lowest 1st percentile of IV regression coefficient is -1.86, significantly far from the true value of 1, the lowest 0.1th percentile is -28, and the 0.01th percentile is -293. The performance of the IV estimator is severely hampered by outliers. For instance, the kurtosis of the OLS estimator is approximately 1.7 in our Monte Carlo sample, it is 187,000 for the controlled IV estimator and 269,000 for the uncontrolled IV estimator.

estimators (uncontrolled or controlled) is correct while the other is not, any not-rejected parameters have low mean-squared error, consistent with theory. Even in a world in which neither estimator is correct, and there may be no viable means of generating a consistent estimator, not-rejected parameters have lower mean-squared error than either estimator separately, as large errors are detected and rejected.

Our test's logic remains the same even if a researcher only has access to a fraction of potentially endogenous covariates, though it is predictably less effective. Figure 9 depicts the mean square error of the five estimators (OLS without controls, IV with and without controls, and our filtered estimator with access to all N potentially endogenous covariates, or only one other covariate) for $\widehat{\beta^{1,1}}$ in several scenarios as a function of distinct potential IV paper uses. The first column depicts a sample given by the calibration in Table 2, but with off-diagonals for β of zero, so that the uncontrolled IV is consistent. The second column instead sets the off-diagonals of θ to zero and sets the covariance of cross-paper error terms to zero, so that the controlled IV is consistent. The third column is given by our baseline calibration: no consistent estimator is available. The top panels give our "small sample" (1000 observation) results, while the bottom panels give our "large sample" (1000000 observation) results. For the no-consistent-estimator panels, our filter with all covariates available removes approximately 12-14% of estimated coefficients of the in small samples, and approximately 95% of estimators in large sample.

For low-levels of potential instrumental variable uses, OLS is inferior to IV in both small- and large-sample. Importantly, however, the bias of OLS does not greatly differ depending on the number of potential IV uses, while our other estimators do. When $N = 1000$, small-sample correlations between errors dominate. Because of this, IV with controls consistently performs worse, even when it is the asymptotically correct estimator (first row, column two). Our filter performs well, with included coefficients slightly better than IV with no controls. When $N = 100,000$ the usefulness of our test becomes clearer. When not including controls is optimal (second row, first column), our filter only chooses those estimates that are (near) correct, with a MSE near zero. When including controls is optimal (second row, second column) our filter again passes only those estimates that are correct. Even in the third column, where neither estimator is asymptotically correct, and no correct estimator is available to the econometrician, our filter delivers estimates with lower MSE than OLS, even when IV with and without controls are worse.

The benefits of having access to only a single, randomly-chosen potentially endogenous covariate out of N is also apparent. In small samples, it performs nearly as well as our full-information estimator. In larger samples, particularly when all available estimators are biased, it reduces the MSE of not-rejected estimators remarkably. It does not perform as well as our full-information estimator when controls are

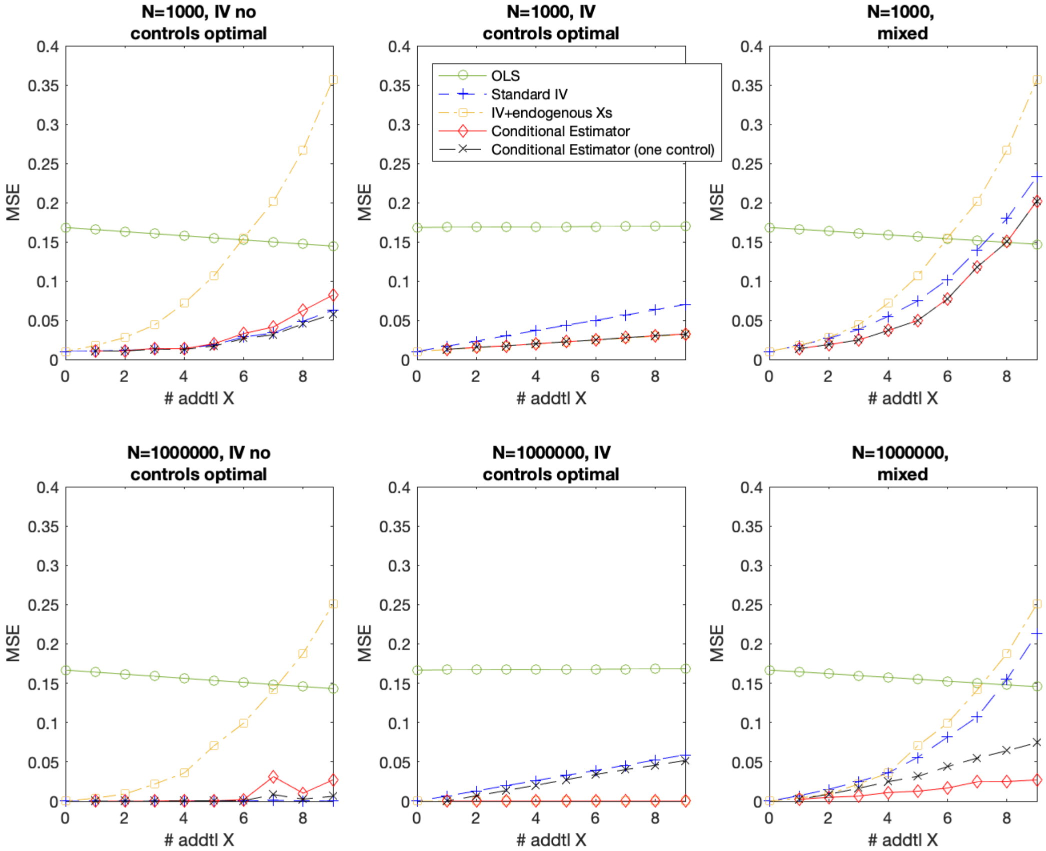

Figure 9: This figure depicts the mean-squared error of four estimators, generated from 100,000 Monte Carlo simulations of the system of equations calibrated in Table 2. The first estimator is an OLS regression of y^1 on x^1. The second is the "standard" IV estimator, which instruments for x^1 using z. The third augments the second by including other paper's endogeneous variables as controls. The fourth is the IV inverse variance weighted combination of controlled and uncontrolled IV, if and only if the estimators are statistically indistinguishable at the 10% level (pass our test).

optimal, but still improves on the accuracy of the uncontrolled IV from which it gets its values.

Figure 10 illustrates the power of our test. Because all estimators are biased in our main "mixed" setup, we define power as the fraction of times our test correctly rejects IV estimation when it is farther in absolute value from the true causal effect than OLS. For instance, with six potential IV uses and a sample size of 100,000, our filter catches between 87-93% of IV estimators whose point estimate is worse than OLS, depending on the correlation between X's. With small samples, IV in our calibration has reasonably large standard errors, so we detect fewer bad point estimates: between 10-20%.

We conclude that controlling for other paper's endogenous covariates and comparing the estimator to our proposed single paper estimator yields a useful test. When the two coefficients are statistically indistinguishable, the IV regression without controls is likely to have low mean-squared error. While it

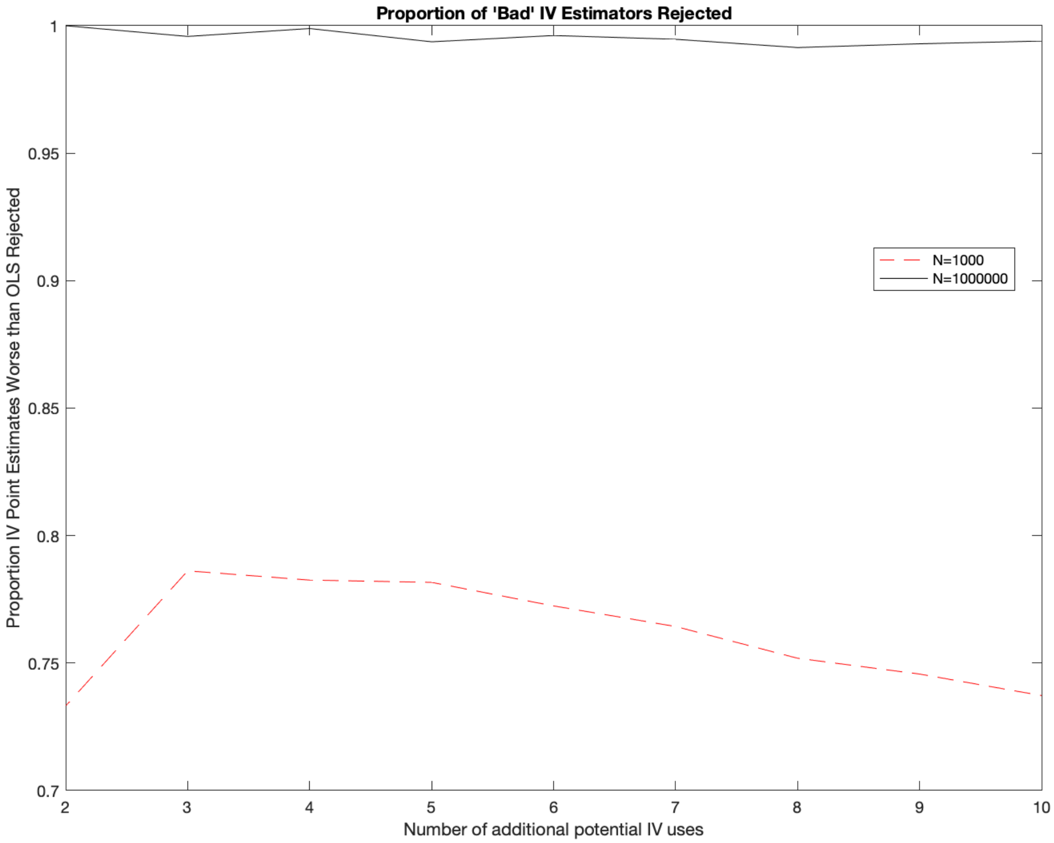

Figure 10: This figure depicts the fraction of "bad" estimates kept by our conditional estimator in the "mixed" data generating process. We define a "bad" IV estimate as one in which the absolute distance of $\widehat{\beta^{IV,NC}}$ from the true β is greater than $\widehat{\beta^{OLS,NC}}$, so that OLS would be preferred.

is true that researchers already routinely informally compare the estimates in different columns of table based on different sets of controls, it is often unclear how to interpret differences between coefficients. This test seeks to formalize the interpretation of these differences in the context of an IV framework, and provide researchers with a helpful test to incorporate into their empirical toolbox.[16]

VII Applications

We apply our proposed tests to two papers in different literatures. First, we examine Rupert and Zanella (2018), which establishes a strong relationship between grandchild birth and grandmother's labor supply. To get an exogenous source of variation for age at grandparenthood, the authors use whether or not a grandmother's first child was a daughter or a son (sibship). In line with this paper's concerns, the authors

[16] Oster (2019) seeks to provide similar formal guidance but not in the context of an instrumental variable regression.

explicitly recognize the potential issues with this multiple-use instrument, focusing on the role of divorce. They argue that because firstborn girls may induce divorce, which increases women's labor supply, they may be finding lower bounds. Second, we examine Mian and Sufi (2014), who use the housing supply elasticities estimated by Saiz (2010) to produce an instrument for housing price changes over the business cycle, which in turn affects employment at the MSA level. Because Saiz (2010) uses elevation and bodies of water (and in the primary regression, religion) as instruments, there may be concern that housing supply elasticities (caused by elevation changes and bodies of water) may also affect segregation, road quality, presence of dams, and size of city government, which may in turn affect an MSA's employment drop in 2008-2009.

VII.1 Rupert and Zanella 2018

To apply our Hausman-like procedure to Rupert and Zanella (2018), we gather several additional variables in the PSID shown to be related to the instrument used to produce their main results concerning grandmother labor supply elasticities. Using the variables from Figure 4, we include (i) housing crowdedness, (ii) child education, (iii) child BMI, (iv) child mobility, (v) parent alcohol use, and (vi) parent tobacco use. For housing crowdedness, we use the person per room (PPR) measure. For child education, we use indicator values denoting the highest level of education achieved by a child. For child BMI, we construct the measure using the standard formula BMI $= \frac{kg}{m^2}$ and create indicator variables for whether or not a parent ever had an obese or underweight child. For child mobility, we construct an only-child indicator variable.[17] For parent alcohol use, we create an indicator variable that takes a value of one if an individual ever reports being a frequent drinker, defined as having a drink at least "several times a week." Similarly, tobacco use is controlled for with an indicator variable taking a value of one if an individual ever reports smoking more than nine cigarettes per day. The data used in Rupert and Zanella (2018) as well as the data for our additional controls are all publicly available from the Panel Survey of Income Dynamics (PSID). We use all available data from the PSID core sample to date, covering the years 1968-2017.[18]

In the first column in the top panel of Table 3, we report replicated results for females from Rupert and Zanella (2018)'s Table 7, column 10, which shows the effect of being a grandparent for seniors using whether or not a senior's firstborn child was female as an instrument.[19] In column (2) of Table 3, we run

[17] Rainer and Siedler (2009) show only children are less likely to move far away from home than individuals with siblings. The authors suggest this result is born out of a sense of duty to care for aging parents. Since individuals with siblings can possibly have a sibling living near their parents, this increases the likelihood they move elsewhere.

[18] This differs slightly from Rupert and Zanella (2018) who use data from 1968-2015. Our point estimates are little changed.

[19] We would like to thank Peter Rupert and Giulio Zanella for generously sharing data and code with us that greatly assisted in

Table 3 Rupert and Zanella (2018) Replication

	Log conditional hours		
	(1)	(2)	(3)
Grandparent	-0.366*	-0.661*	-0.549*
	(0.174)	(0.316)	(0.266)
Endogenous Controls	No	No	Yes
Conditional Data	No	Yes	Yes
N	56374	25316	25316
P-value of Hausman-like test			0.127

Standard errors in parentheses, * $p<0.05$, ** $p<0.01$

Variance-Covariance Matrix

	β_1	β_2	β_3
β_1	(.)	(.)	(.)
β_2	(.)	0.0707	0.0825
β_3	(.)		0.0997

Table 3: Column (1) of the top panel displays our replication for Rupert and Zanella (2018)'s instrumental variables estimates of the effect of grandchildren on the labor supply of grandmothers. Column (2) shows results using the same specification as (1), but conditioning on the availability of data for additional potentially endogenous covariates. That is, we restrict (2) to only include observations that have data for the additional covariates we intend to use to test the robustness of this IV but use the same specification as in (1). Column (3) shows the results using the conditional data after incorporating additional potentially endogenous controls into the specification. The second panel shows the variance-covariance matrix obtained from our estimation of column (2) and column (3) in the first panel. Note, though the difference between β_2 and β_3 is small relative to each βs variance, the covariance is quite large relative to each βs variance. Consequently, the standard error of the difference small, leading to a p-value of 0.127.

the same regressions while conditioning on the data being available in both specifications, which we refer to as conditional data. The results are quantitatively much larger, and are statistically significantly different because of the conditional data restriction. To be consistent, we compare the author's strengthened results in column (2) against the same sample with potentially endogenous controls. We use the sample covariance of individual moment conditions with and without controls to estimate covariance of estimated coefficients in a joint GMM estimation.

Our comparison focuses on the results in columns (2) and (3) of Table 3, top panel. A t-test reveals this difference is marginally insignificant (p=0.127): our test fails to formally reject Rupert and Zanella (2018). The variance-covariance matrix in the bottom panel of Table 3, highlights an important difference between our test and the Hausman test. Under the (incorrect in this setup) Hausman assumptions, the variance of the difference of estimators would be calculated as 0.029, whereas we find it to be 0.005,

the replication and extension.

because the covariance is significantly larger than the smaller of the two variances (the efficient variance in the Hausman test's null). Thus, even though the difference between the two coefficients is small relative to each βs variance, the relatively large covariance makes the standard error of the difference small, resulting in a small (but marginally insignificant) p-value. Due to the asymmetric nature of our test, this failure to reject with relatively tight standard errors around the difference is highly informative for researchers, and our Monte Carlo exercise suggests that Rupert and Zanella (2018)'s baseline estimator is likely to have good mean-square error properties, and that for the subsample we have controls for, the covariate-weighted LATE and the uncontrolled LATE are quite similar, suggesting little room for treatment effect heterogeneity on this sample.

VII.2 Mian and Sufi 2014

To apply our Hausman-like test to Mian and Sufi (2014), we compare three of their specifications to those same specifications with potentially endogenous variables added as exogenous controls. The authors seek to understand how decline in household net worth may affect employment during the Great Recession. In one of their main instrumental variable specifications, the authors regress the change non-tradable employment, defined as either (i) restaurant and retail store employment or (ii) geographically concentrated industries, on the change in housing net worth at the county level. To get an exogenous change in housing prices, the authors instrument the change in housing net worth using the housing supply elasticity measured by Saiz (2010). We focus on Table 3 in their paper. The authors include two important sets of controls, both of which include variables this paper cautions may be endogenous. The first is industry-level controls, (both housing supply elasticity and slopes/bodies of water may affect industrial composition) and the second is demographic controls such as fraction white, median household income, and the poverty level. We compare these main instrumental variables specifications against the same specifications with included endogenous controls.[20]

As additional exogenous controls we include measures by MSA of (i) segregation (ii) density (iii) local government fractionalization (iv) dams (v) roads per capita (vi) broadband provision (vii) rail length and (viii) population growth. For segregation, we include the index of dissimilarity at the tract level within the county and its square. For density we include area and population in levels and logs. For local government fractionalization we proxy using number of independent school authorities divided by the MSA area. For

[20] Because Broomfield County, Colorado only became an independent county in 2001, but our segregation and population controls use census data before that, we drop Broomfield, resulting in only 539 observations, not 540 as Mian and Sufi have. Their results are little changed by the exclusion.

dams, we use counts of the number of major dams in the MSA. To control for roads, we include a measure of miles of roads per person per area. For broadband, we use the average of broadband availability by zip code in the early 2000's. For population growth, we include log population growth from 1990 to 2000. We also include the potential endogenous controls for demographics Mian and Sufi acknowledge in their paper within our set of potentially endogenous variables.

Specifications (1), (3), and (5) in Table 4 replicate Mian and Sufi (2014)'s Table 3 specifications (5), (6), and (7). Column (1) examines restaurant and retail employment using instrumented housing net worth and using two-digit industries as controls. Column (3) replaces restaurant and retail employment with geographically concentrated employment. And specification (5) adds county-level demographic controls to the first. Adjacent specifications (2), (4), and (6) add in both our new potentially endogenous covariates as well as Mian and Sufi's demographic controls. While Mian and Sufi report both spatially-adjusted standard errors and clustered standard errors, they note that clustered standard errors are larger. To make our test conservative in terms of rejection, we take clustered standard errors as well. We calculate the covariance between regression coefficients of different regressions by estimating the two jointly in a GMM framework and allowing arbitrary correlation within clusters across regressions. This means for instance that the observation-level error between specification (1) for San Bernardino County in California and the error in specification (2) (the same regression but with additional controls) for San Diego County in California can be arbitrarily correlated, because they are in the same cluster.

We apply our test between the three specifications of interest and display the results at the bottom of the first panel. The first two reject cross-specification coefficient equality at the 10% and the 5% level, respectively. This constitutes grounds for rejection of the null in our test. As was also the case in our test of Rupert and Zanella (2018), observation-level errors display a high positive covariance, driving us to estimate a high covariance between estimators, with similarly high covariances occurring if we instead bootstrapped the covariances. For instance, comparing specifications (4) and (5), while the point estimates with and without our controls have variances of 0.007 and 0.016 respectively, their covariance is a comparably large 0.005. As with our Rupert and Zanella (2018) results, this example again highlights the importance of our deviation from the Durbin-Wu-Hausman test.

Our test's asymmetric usefulness means that it does not definitively invalidate Mian and Sufi (2014)'s findings. That said, it does raise an issue which merits further thought and discussion. Mian and Sufi (2014)'s original estimates may be valid if, for instance, segregation does not affect an MSA's economic recovery directly, and segregation only shares confounding variation with housing net worth changes and

Table 4 Mian and Sufi (2014) Replication

	(1) Rest. & Retail	(2) Rest. & Retail	(4) Geog. Concen.	(5) Geog. Concen.	(3) Rest. & Retail	(6) Rest. & Retail
Δ Housing Net Worth	0.374**	0.610**	0.208*	0.466**	0.489**	0.610**
	(0.062)	(0.153)	(0.086)	(0.125)	(0.127)	(0.153)
Industry Controls	Yes	Yes	Yes	Yes	Yes	Yes
Possibly Endogenous Dem. ctls.	No	Yes	No	Yes	Yes	Yes
Other Endogenous variable as ctls.	No	Yes	No	Yes	No	Yes
P-value of Hausman-like test		0.070		0.017		0.159
N	539	539	539	539	539	539

Standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

Variance-Covariance Matricies

	$\beta^{IV,NC}$	$\beta^{IV,C}$	$\beta^{IV,NC}$	$\beta^{IV,C}$	$\beta^{IV,NC}$	$\beta^{IV,C}$
$\beta^{IV,NC}$	0.017	0.011	0.007	0.005	0.016	0.016
$\beta^{IV,C}$	(.)	0.023	(.)	0.016	(.)	0.023

Table 4: The top panel of Table 4 displays our replication for Mian and Sufi (2014)'s instrumental variables estimates of the effect of a change in housing net worth on employment at the MSA level. Column (1), (3), and (5) replicate Mian and Sufi's instrumental variables results from their Table 3, columns 5, 6, and 7. Adjacent columns (2), (4), and (5) add in possibly endogenous demographic controls (which are included in Mian and Sufi's Column (5) as well as other possibly endogenous controls outlined above. We report the p-value from our Hausman-like test of the difference between an instrumental variable estimator and the same estimator with potentially endogenous covariates added as controls at the bottom of the first panel. All standard errors are calculated by clustering at the state level. The second panel reports the estimated variance-covariance matrix of the estimators, highlighting high covariance between estimates.

economic recovery, e.g. that the controlled estimator is invalid. Alternatively, it could be argued that both estimators are valid, and controlled 2SLS is only reporting a differently-weighted average of treatment effects. In either case, our test has provided an important flagging mechanism for further discussion, if not outright rejection. However, following our Monte Carlo tests, it raises the possibility that OLS may be less biased that IV in this case.

VIII Conclusion

This paper has discussed some of the issues that arise with "popular" instruments and has discussed six categories of potentially problematic instruments. The use of these potentially troubling instruments is not rare: by our count, which puts only a lower bound on the problem, 318 papers used these instruments, and 83 top five papers. Nor has the use of these instruments declined over time. As we discuss, many of these instruments are also likely related: immigration is likely to affect ethnolinguistic fractionalization and local religion, bringing entire literatures into the discussion of exogeneity. Bodies of water and elevation are likely to affect what cities immigrants immigrate to via the ease of segregation, affect on housing values. Climate and religion affect housing regulation, and so on. We propose that these relationships are not a coincidence: instruments that are strong enough to satisfy many IV first stages are less likely to satisfy many second stages for the same reason they pass the first: ubiquity and economic importance.

We stress that we do not condemn instrumental variables as typically practiced. Some of the examples in this paper are surely well-identified. Moreover, the vast majority of IV papers do not use these instruments, but use instruments that are idiosyncratic to their application, or that are less likely to cause concern. We have not focused on these papers. Instead, we caution against the use of accepted instruments for new purposes.

To better understand when multiple unique uses of an instrument is valid, we propose a new test related to the Hausman test: running a "single paper" IV regression ignoring the other potentially endogenous covariates, and comparing the regression coefficient of interest to an IV regression that includes all those potentially endogenous variables as exogenous controls. We show that because the two potential sources of bias in these two regressions differ, statistical equality between coefficients suggests that either their biases are likely to be small. We differ from the Hausman test by estimating the covariance between coefficients of interest.

We then apply our test to two high-quality instrumental variables papers: Rupert and Zanella (2018), which uses firstborn girls as an instrument for age at which one becomes a grandparent, which affects labor supply, and Mian and Sufi 2014, which uses Saiz (2010)'s elevation and bodies of water derived elasticities as an instrument for housing price changes, which affect non-tradable employment. In the first case, our test formally fails to reject differences between estimated coefficients. In the second case, we find tentative reason for concern, depending on whether or not Mian and Sufi are correct to control for demographics.

We provide two clear positive messages going forward: first, more awareness should be paid to the notion that literatures, or sets of literatures, can "collectively invalidate" an instrument. Second, instrumental variables estimates that are robust to the inclusion of other endogenous controls are likely to be good estimators.

We also submit that a surprising number of papers have used the phrase "while this instrument has been used in [another paper], we are the first to use it in this context." While this is typically used to describe a contribution, it is instead a warning. Just because an instrument has "passed what might be called the American Economic Review (AER)-test," in Rodrik, Subramanian, and Trebbi (2004)'s colorful phrasing, does not mean it is a good instrument for a new paper.

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113 (2):231–263. URL `https://ideas.repec.org/a/eee/econom/v113y2003i2p231-263.html`.

Adda, Jerome. 2016. "Economic Activity and the Spread of Viral Diseases: Evidence from High Frequency Data." *The Quarterly Journal of Economics* 131 (2):891–941.

Adda, Jerome, Christian Dustmann, and Katrien Stevens. 2017. "The Career Costs of Children." *Journal of Political Economy* 125 (2):293 – 337.

Ades, Alberto F. and Edward L. Glaeser. 1995. "Trade and Circuses: Explaining Urban Giants." *The Quarterly Journal of Economics* 110 (1):195–227.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. 2003. "Fractionalization." *Journal of Economic Growth* 8 (2):155–194.

Allcott, Hunt, Allan Collard-Wexler, and Stephen D. O'Connell. 2016. "How Do Electricity Shortages Affect Industry? Evidence from India." *The American Economic Review* 106 (3):587–624.

Altonji, Joseph and David Card. 1991. "The Effects of Immigration on the Labor Market Outcomes of Less-skilled Natives." In *Immigration, Trade, and the Labor Market*. National Bureau of Economic Research, Inc, 201–234.

Ananat, Elizabeth O. and Guy Michaels. 2008. "The Effect of Marital Breakup on the Income Distribution of Women with Children." *The Journal of Human Resources* 43 (3):611–629.

Angrist, Joshua D. and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90 (430):431–442.

Arellano, Manuel and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58 (2):277–297.

Arezki, Rabah and Markus Brückner. 2012. "Rainfall, financial development, and remittances: Evidence from Sub-Saharan Africa." *Journal of International Economics* 87 (2):377 – 385.

Bartel, Ann P. 1989. "Where Do the New U.S. Immigrants Live?" *Journal of Labor Economics* 7 (4):371–391.

Bartik, Timothy J. 1991. *Who Benefits from State and Local Economic Development Policies?* No. wbsle in Books from Upjohn Press. W.E. Upjohn Institute for Employment Research.

Baum-Snow, Nathaniel, Matthew Freedman, and Ronni Pavan. 2018. "Why Has Urban Inequality Increased?" *American Economic Journal: Applied Economics* 10 (4):1–42.

Bazzi, Samuel and Michael A. Clemens. 2013. "Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth." *American Economic Journal: Macroeconomics* 5 (2):152–186.

Becker, Sascha O. and Ludger Woessmann. 2008. "Luther and the Girls: Religious Denomination and the Female Education Gap in Nineteenth-Century Prussia." *The Scandinavian Journal of Economics* 110 (4):777–805.

———. 2009. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *The Quarterly Journal of Economics* 124 (2):531–596.

———. 2018. "Social Cohesion, Religious Beliefs, and the Effect of Protestantism on Suicide." *The Review of Economics and Statistics* 100 (3):377–391.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4):841–890.

Bjørnskov, Christian. 2012. "How Does Social Trust Affect Economic Growth?" *Southern Economic Journal* 78 (4):1346–1368.

Blundell, Richard and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics* 87 (1):115–143.

Boustan, Leah Platt, Price V. Fishback, and Shawn Kantor. 2010. "The Effect of Internal Migration on Local Labor Markets: American Cities during the Great Depression." *Journal of Labor Economics* 28 (4):719–746.

Brainerd, Elizabeth and Nidhiya Menon. 2014. "Seasonal effects of water quality: The hidden costs of the Green Revolution to infant and child health in India." *Journal of Development Economics* 107 (C):49–64.

Brückner, Markus. 2012. "Economic growth, size of the agricultural sector, and urbanization in Africa." *Journal of Urban Economics* 71 (1):26 – 36.

Brückner, Markus and Antonio Ciccone. 2011. "Rain and the Democratic Window of Opportunity." *Econometrica* 79 (3):923–947.

Brückner, Markus and Mark Gradstein. 2013. "Effects of transitory shocks to aggregate output on consumption in poor countries." *Journal of International Economics* 91 (2):343 – 357.

Burke, Paul J. and Andrew Leigh. 2010. "Do Output Contractions Trigger Democratic Change?" *American Economic Journal: Macroeconomics* 2 (4):124–157.

Cáceres-Delpiano, Julio and Marianne Simonsen. 2012. "The toll of fertility on mothers' wellbeing." *Journal of Health Economics* 31 (5):752 – 766.

Card, David. 2001. "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration." *Journal of Labor Economics* 19 (1):22–64.

Card, David and Ethan G. Lewis. 2007. "The Diffusion of Mexican Immigrants During the 1990s: Explanations and Impacts." In *Mexican Immigration to the United States*, NBER Chapters. National Bureau of Economic Research, Inc, 193–228.

Chhaochharia, Vidhi, Dasol Kim, George M. Korniotis, and Alok Kumar. 2018. "Mood, firm behavior, and aggregate economic outcomes." *Journal of Financial Economics* .

Cho, SungJin and John Rust. 2017. "Precommitments for Financial Self-Control? Micro Evidence from the 2003 Korean Credit Crisis." *Journal of Political Economy* 125 (5):1413–1464.

Conley, Dalton and Rebecca Glauber. 2006. "Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variation in Fertility." *The Journal of Human Resources* 41 (4):722–737.

Currie, Janet and Aaron Yelowitz. 2000. "Are public housing projects good for kids?" *Journal of Public Economics* 75 (1):99 – 124.

Cutler, David M. and Edward L. Glaeser. 1997. “Are Ghettos Good or Bad?” *The Quarterly Journal of Economics* 112 (3):827–872.

Dahl, Gordon B. and Enrico Moretti. 2008. “The Demand for Sons.” *The Review of Economic Studies* 75 (4):1085–1120.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2014. “What Do We Learn from the Weather? The New Climate-Economy Literature.” *Journal of Economic Literature* 52 (3):740–798.

Di Falco, Salvatore, Peter Berck, Mintewab Bezabih, and Gunnar Köhlin. 2019. “Rain and impatience: Evidence from rural Ethiopia.” *Journal of Economic Behavior and Organization* 160:40 – 51.

Duflo, Esther and Rohini Pande. 2007. “Dams.” *The Quarterly Journal of Economics* 122 (2):601–646.

Easterly, William and Ross Levine. 1997. “Africa’s Growth Tragedy: Policies and Ethnic Divisions.” *The Quarterly Journal of Economics* 112 (4):1203–1250.

Fafchamps, Marcel, Christopher Udry, and Katherine Czukas. 1998. “Drought and saving in West Africa: are livestock a buffer stock?” *Journal of Development Economics* 55 (2):273 – 305.

Farre, Lidia, Francesc Ortega, and Ryuichi Tanaka. 2018. “Immigration and the public-private school choice.” *Labour Economics* 51:184 – 201.

Felkner, John S. and Robert M. Townsend. 2011. “The Geographic Concentration of Enterprise in Developing Countries.” *The Quarterly Journal of Economics* 126 (4):2005–2061.

Fletcher, Jason M. and Jinho Kim. 2019. “The effect of sibship size on non-cognitive Skills: Evidence from natural experiments.” *Labour Economics* 56:36 – 43.

Frankel, Jeffrey A. and David H. Romer. 1999. “Does Trade Cause Growth?” *American Economic Review* 89 (3):379–399.

Gruber, Jonathan. 2005. “Religious Market Structure, Religious Participation, and Outcomes: Is Religion Good for You?” *The B.E. Journal of Economic Analysis and Policy* 5 (1).

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2003. “People’s opium? Religion and economic attitudes.” *Journal of Monetary Economics* 50 (1):225–282.

Hall, Robert E. and Charles I. Jones. 1999. "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" *The Quarterly Journal of Economics* 114 (1):83–116.

Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (6):1251–1271.

Havnes, Tarjei and Magne Mogstad. 2015. "Is universal child care leveling the playing field?" *Journal of Public Economics* 127:100 – 114. The Nordic Model.

Heath, Davidson, Matthew Ringgenberg, Mehrdad Samadi, and Ingrid M. Werner. 2019. "Reusing Natural Experiments." Working Paper 2019-03-021.

Hidalgo, F. Daniel, Suresh Naidu, Simeon Nichter, and Neal Richardson. 2010. "Economic determinants of land invasions." *The Review of Economics and Statistics* 92 (3):505–523.

Huang, Rocco R. 2008. "Tolerance for uncertainty and the growth of informationally opaque industries." *Journal of Development Economics* 87 (2):333 – 353.

Jacobsen, Joyce P., James Wishart Pearce, and Joshua L. Rosenbloom. 2001. "The effects of childbearing on women's marital status: using twin births as a natural experiment." *Economics Letters* 70 (1):133–138.

Jaeger, David A, Joakim Ruist, and Jan Stuhler. 2018. "Shift-Share Instruments and the Impact of Immigration." Working Paper 24285, National Bureau of Economic Research.

Kaufmann, Daniel, Aart Kraay, and Pablo Zoido-Lobatón. 1999. "Aggregating governance indicators." Policy Research Working Paper Series 2195, The World Bank.

Kazianga, Harounan and Christopher Udry. 2006. "Consumption smoothing? Livestock, insurance and drought in rural Burkina Faso." *Journal of Development Economics* 79 (2):413 – 446. Special Issue in honor of Pranab Bardhan.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens. 2015. "Identification and Inference With Many Invalid Instruments." *Journal of Business & Economic Statistics* 33 (4):474–484.

Kolk, Martin. 2015. "The causal effect of an additional sibling on completed fertility: An estimation of intergenerational fertility correlations by looking at siblings of twins." *Demographic Research* 32 (51):1409–1420.

Korpi, Tomas and Michael Tåhlin. 2009. “Educational mismatch, wages, and wage growth: Overeducation in Sweden, 1974-2000.” *Labour Economics* 16 (2):183 – 193.

Lafortune, Jeanne, Ethan Lewis, and Jose Tessada. 2019. “People and Machines: A Look at the Evolving Relationship between Capital and Skill in Manufacturing, 1860-1930, Using Immigration Shocks.” *The Review of Economics and Statistics* 101 (1):30–43.

LaPorta, Rafael, Florencio Lopez de Silanes, Andrei Shleifer, and Robert Vishny. 1999. “The Quality of Government.” *Journal of Law, Economics and Organization* 15 (1):222–279.

Leamer, Edward E. 1983. “Let’s Take the Con Out of Econometrics.” *The American Economic Review* 73 (1):31–43.

Lee, Iona Hyojung. 2018. “Industrial output fluctuations in developing countries: General equilibrium consequences of agricultural productivity shocks.” *European Economic Review* 102:240 – 279.

Leeb, Hannes and Benedikt M. Pötscher. 2005. “Model Selection and Inference: Facts and Fiction.” *Econometric Theory* 21 (1):21–59. URL `http://www.jstor.org/stable/3533623`.

Levin, Jesse and Erik J.S Plug. 1999. “Instrumenting education and the returns to schooling in the Netherlands.” *Labour Economics* 6 (4):521 – 534.

Lipscomb, Molly, A. Mushfiq Mobarak, and Tania Barilam. 2013. “Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil.” *American Economic Journal: Applied Economics* 5 (2):200–231.

Mauro, Paolo. 1995. “Corruption and Growth.” *The Quarterly Journal of Economics* 110 (3):681–712.

McCleary, Rachel M. and Robert J. Barro. 2006. “Religion and Economy.” *The Journal of Economic Perspectives* 20 (2):49–72.

Mellon, Jonathan. 2021. “Rain, Rain, Go Away: 176 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable.” Working paper.

Mian, Atif, Kamalesh Rao, and Amir Sufi. 2013. “Household Balance Sheets, Consumption, and the Economic Slump.” *The Quarterly Journal of Economics* 128 (4):1687–1726.

Mian, Atif and Amir Sufi. 2011. “House Prices, Home Equity-Based Borrowing, and the US Household Leverage Crisis.” *American Economic Review* 101 (5):2132–56.

———. 2014. “What Explains the 2007-2009 Drop in Employment?” *Econometrica* 82 (6):2197–2223.

Michaelides, Alexander, Andreas Milidonis, and George P. Nishiotis. 2019. “Private information in currency markets.” *Journal of Financial Economics* 131 (3):643 – 665.

Michaelides, Alexander, Andreas Milidonis, George P. Nishiotis, and Panayiotis Papakyriakou. 2015. “The adverse effects of systematic leakage ahead of official sovereign debt rating announcements.” *Journal of Financial Economics* 116 (3):526 – 547.

Miguel, Edward. 2005. “Poverty and Witch Killing.” *The Review of Economic Studies* 72 (4):1153–1172.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. “Economic Shocks and Civil Conflict: An Instrumental Variables Approach.” *Journal of Political Economy* 112 (4):725–753.

Mobarak, Ahmed. 2005. “Democracy, Volatility, and Economic Development.” *The Review of Economics and Statistics* 87 (2):348–361.

Morck, Randall and Bernard Yeung. 2011. “Economics, History, and Causation.” *Business History Review* 85 (1):39–63.

Orrenius, Pia M. and Madeline Zavodny. 2015. “Does Immigration Affect Whether US Natives Major in Science and Engineering?” *Journal of Labor Economics* 33 (S1):S79–S108.

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

Ottaviano, Gianmarco I.P., Giovanni Peri, and Greg C. Wright. 2018. “Immigration, trade and productivity in services: Evidence from U.K. firms.” *Journal of International Economics* 112 (C):88–108.

Palloni, Giordano. 2017. “Childhood health and the wantedness of male and female children.” *Journal of Development Economics* 126:19 – 32.

Parrotta, Pierpaolo, Dario Pozzoli, and Davide Sala. 2016. “Ethnic diversity and firms’ export behavior.” *European Economic Review* 89:248 – 263.

Rainer, Helmut and Thomas Siedler. 2009. “O brother, where art thou? The effects of having a sibling on geographic mobility and labour market outcomes.” *Economica* 76 (303):528–556.

Rajan, Raghuram G. and Rodney Ramcharan. 2011. “Land and Credit: A Study of the Political Economy of Banking in the United States in the Early 20th Century.” *The Journal of Finance* 66 (6):1895–1931.

Roberts, Michael J. and Wolfram Schlenker. 2013. “Identifying Supply and Demand Elasticities of Agricultural Commodities: Implications for the US Ethanol Mandate.” *The American Economic Review* 103 (6):2265–2295.

Rodrik, Dani, Arvind Subramanian, and Francesco Trebbi. 2004. “Institutions Rule: The Primacy of Institutions Over Geography and Integration in Economic Development.” *Journal of Economic Growth* 9 (2):131–165.

Røed, Marianne and Pål Schøne. 2016. “Impact of Immigration on Inhabitants’ Educational Investments.” *The Scandinavian Journal of Economics* 118 (3):433–462.

Rupert, Peter and Giulio Zanella. 2018. “Grandchildren and their grandparents’ labor supply.” *Journal of Public Economics* 159:89 – 103.

Saiz, Albert. 2010. “The Geographic Determinants of Housing Supply.” *The Quarterly Journal of Economics* 125 (3):1253–1296.

Sander, William. 1995. “Schooling and smoking.” *Economics of Education Review* 14 (1):23 – 33.

Sarsons, Heather. 2015. “Rainfall and Conflict: A Cautionary Tale.” *Journal of Development Economics* 115 (July):62–72.

Shih, Kevin. 2017. “Do international students crowd-out or cross-subsidize Americans in higher education?” *Journal of Public Economics* 156:170 – 184.

Stulz, Rene M. and Rohan Williamson. 2003. “Culture, openness, and finance.” *Journal of Financial Economics* 70 (3):313 – 349.

Taber, Christopher R. 2001. “The Rising College Premium in the Eighties: Return to College or Return to Unobserved Ability?” *The Review of Economic Studies* 68 (3):665–691.

Tanaka, Tomomi, Colin F. Camerer, and Quang Nguyen. 2010. “Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam.” *American Economic Review* 100 (1):557–71.

Waldinger, Maria. 2017. “The long-run effects of missionary orders in Mexico.” *Journal of Development Economics* 127:355 – 378.

Young, Alwyn. 2019. “Consistency without Inference: Instrumental Variables in Practical Application.” Working papers, London School of Economics.

Ziliak, James P. and Thomas J. Kniesner. 1999. “Estimating Life Cycle labor Supply Tax Effects.” *Journal of Political Economy* 107 (2):326–359.