

Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth[†]

By SAMUEL BAZZI AND MICHAEL A. CLEMENS*

Concern has intensified in recent years that many instrumental variables used in widely-cited growth regressions may be invalid, weak, or both. Attempts to remedy this general problem remain inadequate. We show how a range of published studies can offer more evidence that their results are not spurious. Key steps include: grounding growth regressions in more generalized theoretical models, deployment of new methods for estimating sensitivity to violations of exclusion restrictions, opening the “black box” of GMM with supportive evidence of instrument strength, and utilization of weak-instrument robust tests and estimators. (JEL C52, E23, F35, O41, O47)

One of the great projects of economic research is to establish the causes of growth. Separating causes from correlates, however, is difficult. Many researchers have recently addressed this difficulty by deploying instrumental variables in cross-country datasets. This can help to identify causes of growth if the instruments do not materially affect growth through channels other than the variable of interest (the instruments are “valid”) and if the instruments correlate well with the variable of interest (the instruments are “strong”). Unfortunately, for reasons not always transparent in published studies, these instruments can be invalid, weak, or both.

In this paper, we examine problems of instrument validity and strength in several growth papers recently published in general-interest and top field journals—not to single out those papers, but to concretely illustrate a general phenomenon that goes well beyond them. First, we discuss how an instrument that is plausibly valid when used in a single setting can be shown invalid by its use in additional settings. Second, we offer evidence that unacknowledged weak instruments may generate spurious findings in important applications, especially those using the popular Generalized Method

*Bazzi: University of California, San Diego, 9500 Gilman Dr. #0534, La Jolla, CA 92093-0534, (e-mail: sbazzi@ucsd.edu). Clemens: Center for Global Development, 1800 Massachusetts Ave. NW, Third Floor, Washington, DC 20036, (e-mail: mclemens@cgdev.org). We thank William Easterly, Aart Kraay, Paul Niehaus, Valerie Ramey, James Rauch, and David Roodman for helpful discussions. We are grateful to three anonymous referees for extremely helpful suggestions. We thank the following for graciously providing access to data and programs: Thorsten Beck, William Hauk, Jason Hwang, Marla Ripoll, Arvind Subramanian, and Sarah Voitchovsky. We gratefully acknowledge support from the Hewlett Foundation. Any errors that remain are exclusively ours. Previous versions of the paper circulated under the titles “Blunt Instruments: A Cautionary Note on Establishing the Causes of Economic Growth,” and “Blunt Instruments: On Establishing the Causes of Economic Growth.” Nothing herein necessarily represents the views of the Center for Global Development, its board, or its funders.

[†]To comment on this article in the online discussion forum, or to view additional materials and an author disclosure statement(s), visit the article page at <http://dx.doi.org/10.1257/mic.5.2.152>.

of Moments (GMM) dynamic panel estimators. This evidence consists of simulation exercises and simple diagnostics on the data underlying published studies.

Our contribution is to show that these problems can have important consequences in real published work, and to suggest remedies. We advocate four ways that growth researchers can surmount these difficulties: by basing instrumental variable regressions on theory sufficiently general to comprise other published results with the same instrument, by using the latest methods to probe sensitivity to violations of the exclusion restriction, by opening the “black box” of GMM with complementary methods to assess instrument strength, and by deploying weak instrument robust testing procedures and estimators. We discuss each in detail below.

I. Instrumentation and Its Discontents

The wave of international growth empirics begun by Baumol (1986) and advanced by Barro (1991) inspired early skepticism even from its own contributors:

“Using these regressions to decide how to foster growth is... most likely a hopeless task. Simultaneity, multicollinearity, and limited degrees of freedom are important practical problems for anyone trying to draw inferences from international data. Policymakers who want to promote growth would not go far wrong ignoring most of the vast literature reporting growth regressions.” (Mankiw 1995).

Researchers thereafter began to address many of these problems. They became more assiduous in checking the robustness of results to the choice of regression specification (Fernández, Ley, and Steel 2001; Sala-i-Martin 1997; Sala-i-Martin, Doppelhofer, and Miller 2004). They explored concerns about parameter heterogeneity, measurement error, and influential observations (Temple 1999; Hauk and Wacziarg 2009). They expanded their samples as the succession of years and improvements in information technology have brought a flood of new data (Bosworth and Collins 2003; Easterly, Levine, and Roodman 2004).

Beyond this, researchers have taken greater care in identifying the causal portion of the relationships they observe across countries. Architects of growth regressions published in top journals have used cross-country instrumental variables for governance quality,¹ trade,² and foreign aid,³ among several other growth determinants. Advances in econometrics have assisted this search for better identification—especially the advent of sophisticated dynamic panel Generalized Method of Moments (GMM) estimators, which entered the growth literature with Caselli, Esquivel, and Lefort (1996).

¹These include cross-country instrumental variables based on exogenous deaths of national leaders while in office (Jones and Olken 2005), colonial-era settler mortality (Acemoglu, Johnson, and Robinson 2001), a Soviet-era survey of ethnolinguistic fractionalization (Mauro 1995), distance from the equator (Hall and Jones 1999), and Pacific-basin wind patterns (Feyrer and Sacerdote 2009).

²These include cross-country instruments based on geography (Frankel and Romer 1999; Frankel and Rose 2002).

³These include cross-country instruments based on political ties, economic policies, and country size (Burnside and Dollar 2000; Angeles and Neanidis 2009). Boone (1996) also uses instruments based on political ties and country size in related work examining the impact of aid on a range of macroeconomic and development outcomes.

But in parallel with these welcome efforts, the economics literature in general has showed increasing concern with the strength and validity of instrumental variables in practice (surveyed by Murray 2006). Close investigations have suggested that many cross-country instruments may be weak, invalid, or both, in widely-cited studies on the growth effects of governance or trade (e.g. Rodríguez and Rodrik 2001; Brock and Durlauf 2001; Dollar and Kraay 2003; Glaeser et al. 2004; Albouy forthcoming; Kraay 2008). Notwithstanding the popularity of instrumental variables in recent growth empirics, Durlauf, Johnson, and Temple (2005) conclude that “the belief that it is easy to identify valid instrumental variables in the growth context is deeply mistaken. We regard many applications of instrumental variable procedures in the empirical growth literature to be undermined by the failure to address properly the question of whether these instruments are valid”. Acemoğlu (2010) decries the widespread use of “instruments without theory,” and Hauk and Wacziarg (2009) see “unjustified claims of causality” as a prominent feature of growth empirics.

This paper extends a growing body of research aimed at identifying econometric best practice in growth empirics. First, we provide concrete evidence on ways in which published studies can collectively invalidate the instruments used in each study separately. Second, building on Bun and Windmeijer (2010) and Hauk and Wacziarg (2009), we indicate and suggest remedies for different sources of bias in the most popular estimator deployed in panel data growth econometrics, the system GMM estimator of Blundell and Bond (1998). Through a Monte Carlo simulation and simple diagnostic tests, we demonstrate the ways in which plausibly valid instruments can mask important weak instrument biases. We conclude with a discussion of solutions that applied researchers can deploy when faced with these identification challenges.

II. When Strong Instruments Are Invalid

To pass a rigorous peer review, each growth study employing an instrumental variable offers theoretical and empirical reasons to believe that the instrument is not substantially correlated with the regression’s error term.

It is well known that this is difficult to establish. There can be a multiplicity of theoretical arguments for and against any given exclusion restriction. The true error term is unobserved in all applied settings, and empirical tests of overidentifying restrictions, which often have low power, hinge on the untestable assumption that at least one instrument is valid—among other reasons.

What is not as well known is that *collectively* the literature establishes the invalidity of some instruments that growth econometricians now use widely, calling into question broad classes of their findings. Suppose that growth is determined by

$$(1) \quad g = \beta_0 + \sum_{j=1}^k \beta_j x_j + \varepsilon,$$

where g is growth; the x_j are a set of k potentially endogenous determinants of growth; the β are parameters to be estimated; and ε is an error term with

mean zero. Suppose we have an instrumental variable z such that $E[z\varepsilon] = 0$ but $\text{Cov}(z, x_j) \neq 0 \ \forall j$. We now try to estimate k separate regressions

$$(2) \quad g = \beta_j^0 + \beta_j x_j + \varepsilon_j, \quad j = 1, \dots, k$$

in each case instrumenting for x_j with z , where $\varepsilon_j \equiv \sum_{\ell \neq j} \beta_\ell x_\ell + \varepsilon$. But unless for every j it is the case that $\beta_\ell = 0$ (or more implausibly $x_j \approx x_\ell$) for all $\ell \neq j$, we have $\text{Cov}(z, \varepsilon_j) = \sum_{\ell \neq j} \beta_\ell \text{Cov}(z, x_\ell) \neq 0 \ \forall j$, and the instrument z is invalid in *every* regression (2). In other words, if existing research has shown that z is a strong instrument for a variable x_ℓ not included in a regression of the form (2) and $\beta_\ell \neq 0$, then z need not be a valid instrument for x_j . Any estimate $\hat{\beta}_j$ will be biased to an unknown degree in an unknown direction, throwing into question the credibility of all results from the regressions (2). As Durlauf, Johnson, and Temple (2005, 635) point out, “Since growth theories are mutually compatible, the validity of an instrument requires a positive argument that it cannot be a direct growth determinant or correlated with an omitted growth determinant.”

And the story gets worse. We might think that including some of the omitted $x_{\ell \neq j}$ in the regression (2) could help, but that brings a new problem. For each $x_{\ell \neq j}$ included in (2), an additional instrument \tilde{z} is required—one that is valid ($E[\tilde{z}\varepsilon] = 0$) and remains strong when used with the other (i.e., $\text{Corr}(z, x_j | \tilde{z}) \neq 0$ and $\text{Corr}(\tilde{z}, x_{\ell \neq j} | z) \neq 0$). This is a high bar.⁴ Setting aside the difficulty of finding multiple valid instruments, Dollar and Kraay (2003) describe a case where each of two instruments appears strong in isolation but is so highly correlated with the other that both are weak when used together. We return to problems of instrument weakness in Section III.

A. Original Sins

These systematic problems with instrument validity arise prominently in the widespread use of “legal origins” in growth regressions, a practice that has become the subject of frequent grumbling at conference coffee breaks. A flotilla of recent cross-country growth regressions has employed an indicator of the origin of a country’s legal system (British, French, Scandinavian, and so on) as an instrument in a variety of regression specifications—each one of which suggests that the instrument is invalid in all of the other specifications. Many have passed the rigors of peer review at general interest journals and top field journals.

Friedman et al. (2000) use legal origin as an instrument for five separate measures of “the quality of economic institutions” (corruption, tax rates, over-regulation, etc.) in regressions with the size of the unofficial economy as the dependent variable, which could directly affect growth. Djankov et al. (2003) use legal origin as an instrument for “the degree of formalism of the legal procedure,” which they argue causes a decline in the quality of the legal system (its honesty, impartiality, ability to enforce contracts, and so on) that could be a major determinant of growth. Lundberg and Squire (2003) use legal origin as an instrument for inflation, the inequality of

⁴In fact, nonzero partial correlation is not enough. If the instruments z and \tilde{z} are weak, then even a small degree of endogeneity in the instruments could lead IV estimates to be more biased than OLS (Hahn and Hausman 2005).

land ownership, and several other variables that they argue directly affect growth. If any two of these studies are correct, growth is determined by a form of equation (1) that renders instrumentation in the IV regressions (2) invalid.

It does not stop there. Alfaro et al. (2004) use legal origin as an instrument for private sector credit, bank credit, and stock market capitalization, which they argue condition the effect of foreign direct investment (FDI) on growth. Levine, Loayza, and Beck (2000) similarly use legal origin to instrument for three separate proxies for financial intermediation, all of which they argue cause economic growth. Glaeser et al. (2004) use legal origin as an instrument for “executive constraints” and average years of schooling in the population, with the level of income per capita as the dependent variable. Beck, Demirgüç-Kunt, and Levine (2005) use legal origin as an instrument for “the relative size of the small and medium enterprise sector,” which could be associated with growth. There are other examples.

If two or more of the above endogenous variables sufficiently affect growth, then instrumentation can be valid in at most one of these studies, and at worst none.

B. Size Matters—Through Various Channels

We turn to another instrument in widespread use, and dwell on it at greater length because its problems are less broadly recognized. Several recent cross-country studies published in general interest journals and top field journals rest their identification strategies on the correlation of population size with some endogenous variable. In each case, the authors give plausible reasons why population size is not only a strong instrument but uncorrelated with their regressions’ error terms: growth regressions do not typically find population scale effects (Rose 2006; Easterly 2009).

However, when viewed collectively, these studies exhibit a problem that undermines their careful arguments in support of instrument validity. Given that none of these studies include the other studies’ endogenous variables as regressors, if population size is a strong and valid instrument in even one of these studies, then it is invalid in *all* of the others.⁵ In other words, the conjecture in Deaton (2010) that measures of country size can affect growth through multiple channels has empirical support.

This pattern emerges in several recent and prominently published regressions. Some investigators use population size (among other geographic characteristics) as an instrument for trade as a determinant of the level of income per capita (Frankel and Romer 1999; Frankel and Rose 2002) or its growth (Spolaore and Wacziarg 2005). Others regress growth not on the level of trade but on an indicator of the mix of goods exported, instrumented by population size (Hausmann, Hwang, and Rodrik 2007), without controlling for the level of trade. Still others use population size as an instrument to identify the effect of foreign aid on democracy (Djankov, Montalvo, and Reynal-Querol 2008), which many studies find to correlate with growth in some fashion.⁶ Another approach uses country size—measured by area and level of GDP, but

⁵Even if these studies included one or more of the endogenous variables in other studies, the authors would face precisely the problem discussed earlier, in the paragraph prior to Subsection IIA.

⁶For investigations of the effect of democracy on growth, see Barro (1996), Tavares and Wacziarg (2001), Giavazzi and Tabellini (2005), Rodrik and Wacziarg (2005), Persson and Tabellini (2006, 2007), and Papaioannou and Siourounis (2008).

strongly correlated with population—to instrument for receipts of FDI as a determinant of growth (Borensztein, De Gregorio, and Lee 1998).

The exclusion restriction necessary for population size to be a valid instrument for each of these endogenous variables is violated to a greater degree, to the extent that the causal pathway identified in any of the other studies is correct. Regardless of any theoretical and empirical case for instrument validity made by each paper in the group, population size can only be a strictly valid instrument in one of them at best, and none of them at worst. The degree to which each estimate is thereby biased could be small or large, but should not be ignored.

C. Strength in Numbers, but Not Validity

The problem extends further than this, however, in a way that is not generally recognized. Many studies resort to multiple instruments, responding to criticism by pointing out that allegations of invalidity or weakness only apply to some of the instruments. It is common to gloss over the problem that the most valid instruments in the basket could be the weakest, and that the strongest could be the least valid.

Building on the above discussion of the population size instrument, it is possible for a study in which the identification strategy appears to rely on multiple instruments to rely entirely on population size. Rajan and Subramanian (2008) execute cross-section regressions of growth on foreign aid receipts, with aid instrumented by a variable constructed (in an auxiliary or zero-stage regression) from aid-recipient population size, aid-donor population size, colonial relationships, and language traits (see Appendix A). Rajan and Subramanian (2008, footnote 16) write, “Our instrument... contains information that is not just based on recipient size.”⁷ But the instrument contains, in fact, almost no information beyond the size of the recipient’s population. In Rajan and Subramanian’s (2008) data, for the period 1970–2000, the in-sample correlation of log population and the constructed instrument is -0.93 . In the periods 1980–2000 and 1990–2000, this correlation is -0.95 . In effect, Rajan and Subramanian (2008) are instrumenting for aid with population alone, though they recognize the problem with using population size as an instrument.⁸

This problem deserves additional discussion, since it is common in applied work to rest identification on a group of instruments without making explicit which of them bears the burden of identification and, therefore, the key burden of validity. Frankel and Romer (1999) demonstrate that their gravity-based instrument—also constructed in an auxiliary regression—contains information beyond country size by treating log population and log area as exogenous and, hence, including them

⁷They justify this claim (in their table 5, panel C) by using one measure of country size (population) as an excluded instrument in the construction of their generated instrument (\bar{a}_i) and, in a robustness check, showing that \bar{a}_i retains strength when a different measure of country size (land area) is used as an additional excluded instrument in the first stage. But the only way to accurately assess whether or not \bar{a}_i contains information beyond population size is to test whether or not it retains significance when population itself is included as a separate instrument, as we do here.

⁸“While a measure of country size could in itself be a plausible instrument, the reason not to make it the preferred one is that there is uncertainty whether it can satisfy the exclusion restriction; that is, a number of reasons can be advanced as to why a recipient’s size would have an independent effect on growth.” (Rajan and Subramanian 2008, footnote 16).

in both the first-stage and the second-stage.⁹ Taking this minimalist approach, we explore, in Tables 1 and 2, the role of population as an instrument using the original data of Rajan and Subramanian (2008).

For each specification, we test for underidentification and for weak instruments. To test for underidentification, we report p -values for a test of the null hypothesis that the structural equation is underidentified based on a Lagrange-Multiplier (LM) test using the rank-based rk statistic due to Kleibergen and Paap (2006). A rejection of the null indicates that the smallest canonical correlation between the endogenous variables and the instruments is nonzero. However, nonzero correlations are not sufficient for strong identification. We therefore also report first-stage F -statistics—Wald statistics based on Cragg and Donald (1993) and the Kleibergen and Paap (2006) generalization to non-independently and -identically distributed errors—and associated p -values for weak-instruments hypothesis tests.

Following the diagnostic approach developed in Stock and Yogo (2005) and implemented in Yogo (2004), we report p -values for the null hypotheses that the bias in the point estimate(s) on the endogenous variable(s) is greater than 10 percent or 30 percent of the OLS bias, or that the the actual size of the t -test that the point estimate(s) on the endogenous variable(s) equal zero at the 5 percent significance level is greater than 10 or 25 percent.¹⁰ While it has become common practice in the empirical growth literature to report first-stage F -statistics, the inferential implications often go unstated. By reporting p -values, we offer a probabilistic lens into the weak-instruments problem.

Table 1 shows that essentially all instrumentation power in the primary Rajan and Subramanian specification comes from the population instrument. Column 1 exactly reproduces a representative cross-section regression (Rajan and Subramanian 2008, table 4, column 2). Instrumentation is very strong, as indicated by the tests for underidentification and weak instruments. Column 2 of Table 1 includes log population in the second stage, and instrument strength collapses. We fail to reject the null hypothesis that the structural equation is underidentified. Applying the conditional likelihood ratio (CLR) test of Moreira (2003), which is robust to weak instruments, we obtain an uninformative confidence interval on aid/GDP comprising the entire real line. Column 3 discards Rajan and Subramanian's (2008) constructed instrument altogether and uses log population alone as an instrument for aid, giving results nearly identical to those in column 1. Column 4 re-estimates the constructed instrument without the population size terms, and instrument strength

⁹However, upon more rigorous examination of the exclusion restrictions implicit in this instrument, Frankel and Rose (2002) conclude that among the six plausibly exogenous geographic determinants of trade flows used to construct their predicted trade instrument, log population is the only one that violates the implicit overidentifying restrictions used in constructing the instrument. See footnote 15 of Frankel and Rose (2002). This result supports our claims in this section about the nonexcludability of size. Debate over other aspects of the Frankel and Romer specification can be found in Rodríguez and Rodrik (2001) and Noguer and Siscart (2005).

¹⁰These p -values are based on comparing the appropriately scaled large-sample versions of the Cragg-Donald and Kleibergen-Paap statistics to the critical values in Stock and Yogo (2005). Critical values have not been tabulated for the Kleibergen-Paap rk statistic since the specific thresholds depend on the type of violation of the independently and identically distributed assumption, which differ across applications. We follow others in the literature and apply the critical values tabulated for the Cragg-Donald statistic to the Kleibergen-Paap results (see Baum, Schaffer, and Stillman 2007). See online Appendix A for further details.

TABLE 1—RAJAN AND SUBRAMANIAN (2008) CROSS-SECTION REGRESSIONS, 1970–2000

	(1)	(2)	(3)	(4)
Point estimate: Aid/GDP	0.096 (0.070)	0.911 (4.083)	0.078 (0.066)	−15.944 (633.474)
CLR confidence set: Aid/GDP ^a	[−0.027, 0.291]	(−∞, ∞)	[0.039, 0.252]	(−∞, ∞)
Initial log population		1.604 (7.923)		
Initial log GDP/capita	1.409 (0.435)	1.061 (12.782)	1.438 (0.403)	25.491 (953.378)
<i>Other parameter estimates omitted</i>				
Excluded instrument	\bar{a}_r	\bar{a}_r	$\ln(\text{population})$	$\bar{a}_r \text{ sans population}$
Observations	78	78	78	78
Kleibergen-Paap LM test (p -value) ^b	0.0004	0.772	0.0001	0.978
Cragg-Donald Wald stat ^c	31.63	0.133	36.30	0.001
H_0 : t -test size > 10 percent (p -value)	< 0.001	0.982	< 0.001	0.999
H_0 : t -test size > 25 percent (p -value)	< 0.001	0.774	< 0.001	0.980
H_0 : relative OLS bias > 10 percent (p -value)	< 0.001	0.952	< 0.001	0.996
H_0 : relative OLS bias > 30 percent (p -value)	< 0.001	0.852	< 0.001	0.987
Kleibergen-Paap Wald stat ^c	36.12	0.073	32.14	0.001
H_0 : t -test size > 10 percent (p -value)	< 0.001	0.987	< 0.001	0.999
H_0 : t -test size > 25 percent (p -value)	< 0.001	0.831	< 0.001	0.984
H_0 : relative OLS bias > 10 percent (p -value)	< 0.001	0.965	< 0.001	0.996
H_0 : relative OLS bias > 30 percent (p -value)	< 0.001	0.852	< 0.001	0.987

Notes: The dependent variable in all specifications is average annual growth in GDP per capita over the period. \bar{a}_r is the generated instrument for foreign aid receipts/GDP (see Appendix A). Heteroskedasticity-robust standard errors in parentheses. Column 1 exactly replicates the baseline result from Rajan and Subramanian (2008, table 4, column 2) for the 1970–2000 cross-section. Column 2 includes log population in the second stage. Column 3 replaces estimated aid/GDP \bar{a}_r with log population as the sole excluded instrument. Column 4 removes donor and recipient population terms from the zero-th stage specification used to estimate the predicted aid/GDP instrument \bar{a}_r , retaining only the colonial ties indicators.

^aThe CLR confidence set corresponds to the weak-instrument robust confidence set obtained using the conditional likelihood ratio test in Moreira (2003).

^bThe null hypothesis of the Kleibergen-Paap LM test is that the structural equation is underidentified (i.e., the rank condition fails). The test uses a procedure from Kleibergen and Paap (2006).

^cIn this special case of a single endogenous regressor, the Cragg-Donald and Kleibergen-Paap Wald statistics reduce, respectively, to the standard nonrobust and heteroskedasticity-robust first-stage F -statistics. Below each, we report the p -values from tests of whether (i) the actual size of the t -test that $\beta_{aid} = 0$ at the 5 percent significance level is greater than 10 or 25 percent, and (ii) the bias of the IV estimates of β_{aid} reported in the table are greater than 10 or 30 percent of the OLS bias. In both cases, the critical values are obtained from Stock and Yogo (2005). Although critical values do not exist for the Kleibergen-Paap statistic, we follow the approach suggested in Baum, Schaffer and Stillman (2007) and apply the Stock and Yogo critical values initially tabulated for the Cragg-Donald statistic. The critical values for (ii) are (less conservatively) based on three instruments since one cannot calculate critical values in the (finite-sample) bias tests for the case of one endogenous variable and fewer than three instruments.

is abysmally low. Table 2 shows only the first-stage F -statistics from the Rajan and Subramanian (2008) cross-section regressions for 1970–2000 and 1980–2000 (results are similar for 1990–2000), first in exact replication of their results, then with population terms deleted from the construction of their instrument, then with the instrument constructed based only on population and its interactions.

In all cases, we cannot reject that the structural equation is underidentified and aid is weakly instrumented when information about population is absent from the

TABLE 2—INSTRUMENTATION STRENGTH IN RAJAN AND SUBRAMANIAN (2008) CROSS-SECTION REGRESSIONS

Period	1970–2000 ($N = 78$)			1980–2000 ($N = 75$)		
	Replication	Colonial vars. only	Population vars. only	Replication	Colonial vars. only	Population vars. only
“Zero-Stage” Specification:	(1)	(2)	(3)	(4)	(5)	(6)
Point estimate: Aid/GDP	0.096 (0.070)	−15.944 (633.474)	0.078 (0.067)	−0.004 (0.095)	−0.308 (0.389)	−0.028 (0.084)
CLR confidence set: Aid/GDP	[−0.027, 0.292]	(−∞, ∞)	[−0.039, 0.254]	[−0.186, 0.232]	(−∞, ∞)	[−0.194, 0.170]
Kleibergen-Paap LM test (p -value)	0.0004	0.978	0.0001	0.0002	0.282	0.0001
Cragg-Donald Wald stat	31.63	0.001	35.90	29.37	1.41	40.54
H_0 : t -test size > 10 percent (p -value)	< 0.001	0.999	< 0.001	0.001	0.888	< 0.001
H_0 : t -test size > 25 percent (p -value)	< 0.001	0.980	< 0.001	< 0.001	0.341	< 0.001
H_0 : relative OLS bias > 10 percent (p -value)	< 0.001	0.996	< 0.001	< 0.001	0.772	< 0.001
H_0 : relative OLS bias > 30 percent (p -value)	< 0.001	0.987	< 0.001	< 0.001	0.503	< 0.001
Kleibergen-Paap Wald stat	36.12	0.001	31.62	31.26	1.41	39.65
H_0 : t -test size > 10 percent (p -value)	< 0.001	0.999	< 0.001	0.001	0.888	< 0.001
H_0 : t -test size > 25 percent (p -value)	< 0.001	0.984	< 0.001	< 0.001	0.340	< 0.001
H_0 : relative OLS bias > 10 percent (p -value)	< 0.001	0.997	< 0.001	< 0.001	0.770	< 0.001
H_0 : relative OLS bias > 30 percent (p -value)	< 0.001	0.990	< 0.001	< 0.001	0.502	< 0.001

Notes: In all specifications, the instrumental variable is aid/GDP predicted from the zero-stage regression. The dependent variable in all specifications is average annual growth in GDP per capita over the period. Heteroskedasticity-robust standard errors in parentheses. Following the original paper, we retain the degrees-of-freedom adjustment to the Kleibergen-Paap F and LM statistics based on robust standard errors. For each of the three periods, the first column is based on an exact replication of the baseline result in Rajan and Subramanian (2008, table 4). The second column removes donor and recipient population terms from the zero-th stage specification used to estimate the predicted aid/GDP instrument \bar{a}_i , retaining only the colonial ties indicators. The third column retains only the population terms in the zero-th stage. See the notes to Table 1 for more details on the CLR confidence set as well as the Kleibergen-Paap and Cragg-Donald tests. Results for the period 1990–2000 are similar and can be found in online Appendix D.

constructed instrument, and strongly instrumented when (only) those variables containing information about population are present. Moreover, using the CLR method of inference robust to weak instruments, it is not possible to rule out extremely large *or* extremely small negative *or* positive effects of aid on growth.

The Rajan and Subramanian (2008) cross-section method is indistinguishable from instrumenting exclusively with aid-recipient population. The subsequent discussion of the validity of any other variable in the instrument matrix, then, is not informative about the causal relationship between aid and growth. What matters is the validity of the instrument that strongly identifies causation. Since that is only country size, the Rajan and Subramanian (2008) analysis shares the same problem faced by the other papers resting on the population instrument—all of the aforementioned papers that use the population instrument invalidate its usage in these studies, since the regressions there do not control for the level of trade, the mix of goods exported, FDI, or democracy. And the Rajan and Subramanian (2008) exercise does not resolve important questions about the validity of the population instrument in all of the other papers that use it because those papers do not control for aid receipts.

Beyond Generated Instruments.—This problem is, in fact, much more general than the use of instruments generated from auxiliary regressions. As an example, we consider two other prominent studies in the aid and growth literature: Burnside and Dollar (2000) and its highly-cited antecedent in Boone (1996). Table 3 examines the interplay of instrument strength and validity in each of these studies, which employ country size alongside several other instruments in a pooled 2SLS specification.¹¹ Again, we test for underidentification (Kleibergen-Paap LM test) and weak instruments (Cragg-Donald and Kleibergen-Paap Wald stats). We also show Hansen's J tests of the null hypothesis that, roughly speaking, the instruments are valid.

Unsurprisingly, instrumentation is strong in columns 1 and 3, which replicate the studies' baseline specifications including the size instruments. However, when relaxing the excludability of log population in column 2, instrument strength collapses as Boone's political instruments identifying prominent donor-recipient relationships are weakly correlated with aid/GDP. Meanwhile, relaxing the excludability of the size instruments in Burnside and Dollar (2000), we find in column 4 that the remaining policy instruments (see notes below the table) still explain some of the variation in aid/GDP as we reject the null of underidentification. Yet, the point estimate on aid/GDP has 10–30 percent of the OLS bias. Further relaxing the excludability of the policy instruments, instrument strength drops considerably in column 5 as the remaining political instruments identifying prominent donor-recipient relationships again prove to be weakly correlated with aid/GDP. We can thus conclude that the seminal aid and growth studies due to Boone (1996) and Burnside and Dollar (2000) suffer from the same identification challenges as their most recent successor, Rajan and Subramanian (2008).

In Table 3, we also deploy several tests of overidentifying restrictions aimed at characterizing instrument (in)validity. In column 1, the p -value of 0.12 for the Hansen (1982) test provides evidence against the null hypothesis that the full set of instruments in Boone (1996) is valid (or the model is correctly specified). We find similar evidence when comparing the Hansen (1982) test statistics with and without the size instruments.¹²

Moreover, by relaxing the excludability of population size in column 2, the p -value on the smaller set of (weak) political instruments triples. Treating the size instrument as strong and the political instruments as weak a priori, we fail to reject the validity of population size on the basis of a Hausman-type test (see notes to the table) for the validity of a strong instrument in the presence of weak instruments (Hahn, Ham, and Moon 2011), which delivers a p -value of 0.05.¹³ Turning to columns 3–5 for Burnside and Dollar (2000), the message is less clear. Yet, we do find relatively lower p -values for the difference-in-Hansen tests pertaining to the validity of the population size instruments. Taken together,

¹¹ Here, we extend and elaborate upon related points raised in Clemens et al. (2012), which provides more detail on these two seminal aid and growth studies.

¹² See Hayashi (2000, 220, 232–34) for a discussion of these tests of overidentifying restrictions based on the difference-in-Hansen or C statistic.

¹³ While informative as a heuristic test, the asymptotic properties of this test have been criticized by Guggenberger (2009) and the authors themselves.

TABLE 3—UNPACKING THE SOURCES OF IDENTIFICATION IN SEMINAL AID-GROWTH REGRESSIONS

Study	Boone (1996)		Burnside and Dollar (2000)		
Population IVs in 2nd stage ^a	—	Yes	—	Yes	Yes
Policy IVs in 2nd stage ^b	—	—	—	—	Yes
	(1)	(2)	(3)	(4)	(5)
Point estimate: Aid/GDP	0.235 (0.198)	−0.782 (0.818)	−0.119 (0.180)	−0.206 (0.441)	0.363 (1.190)
CLR confidence set: Aid/GDP	[−0.125, 0.562] (−∞, 0.261] ∪ [8.674, ∞) [−0.523, 0.231] [−1.314, 0.609] [−1.923, 3.582]				
Observations	132	132	275	275	275
Hansen test all instruments (<i>p</i> -value)	0.123	0.368	0.194	0.290	0.122
Hansen test excl. size instruments (<i>p</i> -value)	0.197	—	0.313	—	—
Difference-in-Hansen test (<i>p</i> -value) ^c	0.112	—	0.154	—	—
Hansen test excl. policy instruments (<i>p</i> -value)	—	—	0.078	0.169	—
Difference-in-Hansen test (<i>p</i> -value) ^c	—	—	0.799	0.456	—
Hansen test excl. size & policy instruments (<i>p</i> -value)	—	—	0.237	—	—
Difference-in-Hansen test (<i>p</i> -value) ^c	—	—	0.230	—	—
Kleibergen-Paap LM test (<i>p</i> -value)	0.004	0.201	< 0.0001	0.057	0.124
Cragg-Donald Wald stat	15.70	1.77	19.74	7.04	4.65
H_0 : relative OLS bias > 10 percent (<i>p</i> -value)	0.010	0.724	< 0.001	0.510	0.409
H_0 : relative OLS bias > 30 percent (<i>p</i> -value)	< 0.001	0.442	< 0.001	0.014	0.162
Kleibergen-Paap Wald stat	7.57	1.69	15.76	5.90	2.77
H_0 : relative OLS bias > 10 percent (<i>p</i> -value)	0.283	0.734	0.001	0.687	0.603
H_0 : relative OLS bias > 30 percent (<i>p</i> -value)	0.028	0.455	< 0.001	0.043	0.313

Notes: The dependent variable in all specifications is average annual growth in GDP per capita over the period. The regressions replicated and modified are Boone (1996, table 4, column 5, row 3) and Burnside and Dollar (2000, table 4, column 3 2SLS). The other coefficients are suppressed, but details on the replication of the original studies can be found in Clemens et al. (2012), which reported an abbreviated version of figures in this table.

^aThe log population instrument is included in the second stage of the Boone (1996) regression, and the following instruments are included in the second stage of the Burnside and Dollar (2000) regression: log population, log population × policy, and (log population)² × policy.

^bThe following instruments are included in the second stage of the Burnside and Dollar (2000) regression: log initial income × policy, (log initial income)² × policy, and lagged arms imports/total imports × policy.

^cThe null hypothesis of the difference-in-Hansen test (or *C* statistic, see Hayashi 2000) is that the given suspect instruments are valid. This test is not robust to weak instruments. Applying an alternative Hausman test that is robust to weak instruments albeit problematic for other reasons (see Hahn, Ham and Moon 2011), we fail to reject the hypothesis that log population is a valid instrument in column 1 (*p*-value of 0.053) for Boone (1996). Applying the same test to column 3 for Burnside and Dollar (2000), we fail to reject that the size, policy, or size and policy instruments combined are valid with *p*-values of 0.411, 0.622, and 0.202 respectively. Following the original papers, we retain the degrees-of-freedom adjustment to the Kleibergen-Paap *F* and LM statistics based on country-level clustered standard errors in Boone and robust standard errors in Burnside and Dollar (2000). See the notes to Table 1 for more details on the CLR confidence set as well as the Kleibergen-Paap and Cragg-Donald tests.

these tests, and the associated point and set estimates for aid/GDP, provide additional evidence of the difficulties that arise when weak instruments are valid and strong instruments are invalid.¹⁴

¹⁴Of course, one must also recognize that these tests of overidentifying restrictions ultimately hinge on the untestable assumption that at least one of the instruments is valid.

This problem also extends beyond pooled cross-section models to dynamic panel regressions with numerous nonsize-based instruments. As an example, we consider the ten year panel regressions in Hausmann, Hwang, and Rodrik (2007). The authors utilize two estimators: a pooled 2SLS estimator with log population and log area as instruments, and the Blundell and Bond (1998) dynamic panel system GMM estimator with instrumental variables that include log population and log area, as well as a standard set of lagged covariates employed in this popular estimation strategy (see Section III below for a detailed discussion of this estimator).

Table 4 demonstrates that the key dynamic panel result in Hausmann, Hwang, and Rodrik (2007) hinges on the excludability of country size from the levels equation—despite plausibly valid moment conditions comprising lagged levels and differences of the endogenous variables. The statistically and economically significant effect of time-varying export product diversity (initial EXPY) on economic growth is driven primarily by its covariation with slowly changing log population and time-invariant log area. Columns 1 and 2 replicate the results from columns 6 and 8 of table 9 in Hausmann, Hwang, and Rodrik (2007).¹⁵ In column 3, we do not exploit moment conditions for log population and log area in the difference equation. Doing so leaves the (difference-in-)Hansen test statistics and inference largely unchanged. In column 4, we do not exploit moment conditions for the size instruments in the levels equation, and in column 5, we do not exploit these moment conditions in either equation. As we treat country size as nonexcludable in increasingly more equations in the system, we are less likely to reject the null of valid identifying restrictions. Of course, there may be other circumstances in which population size and/or area could be excludable from one equation, neither, or both.

Not unlike the previous examples from the aid and growth literature, identification of the panel regressions in Hausmann, Hwang, and Rodrik (2007) depends crucially on size-based instruments that, if they are valid in this setting, require causal pathways identified in other studies to be incorrect.¹⁶ While Hausmann, Hwang, and Rodrik (2007) mention this being a potential problem in their pooled 2SLS specification,¹⁷ they do not consider how or why the system GMM estimator fails to solve the problem.

We can go beyond mere suspicion that residuals in some of these studies are correlated with the endogenous variables in the other studies. Table 5 shows this within the Rajan and Subramanian (2008) framework. Here we perform ten cross-section OLS regressions, each with a candidate growth determinant x_j ($j = 1, \dots, 10$) on the left-hand side that has been omitted from the Rajan and Subramanian (2008) regressions, $x_j = \beta_j \ln population + \mathbf{Z}'\Theta_j + u$, where the \mathbf{Z} are the second-stage regressors

¹⁵ Despite utilizing their original Stata code and dataset, the system GMM replication in column 2 differs slightly, albeit immaterially, from the published results. See online Appendix E.

¹⁶ As we show in online Appendix D, this same set of results does not hold in the longer, five-year panel periodization in Hausmann, Hwang, and Rodrik (2007). Given the higher frequency and additional periods in this specification, the variation in the system GMM instruments comprised of lagged levels and differences of endogenous covariates swamps the potentially nonexcludable variation in country size. We cannot reject the null of valid overidentifying restrictions implied by the full instrument matrix or the size instruments alone.

¹⁷ “The variables used as instruments [log population and log area] fail the overidentification test in columns (2) and (6) [pooled 2SLS], most likely because they are persistent series akin to country fixed effects in a panel. Reassuringly, columns (4) and (8) show that the GMM setup where lagged levels and differences are used as instruments passes both the overidentification test and exhibits no second order correlation” (Hausmann, Hwang, and Rodrik 2007, footnote 9).

TABLE 4—UNPACKING THE SOURCES OF IDENTIFICATION IN 10-YEAR PANELS OF HAUSMANN, HWANG, AND RODRIK (2007)

Estimator	IV	GMM-SYS	GMM-SYS	GMM-SYS	GMM-SYS	GMM-SYS
GMM-SYS moment conditions?	No	Yes	Yes	Yes	Yes	Yes
Size instruments?	Yes	Yes	Yes, lev. eq.	Yes, diff. eq.	No	Yes
Size excluded?	Yes	Yes	Yes	Yes	Yes	No
	(1)	(2)	(3)	(4)	(5)	(6)
log initial GDP/capita	−0.038 (4.425)	−0.013 (1.567)	−0.015 (1.687)	0.003 (0.233)	0.011 (0.984)	0.011 (1.157)
log initial EXPY	0.092 (4.598)	0.043 (2.315)	0.047 (2.600)	0.008 (0.213)	−0.017 (0.777)	−0.017 (0.796)
log human capital	0.004 (1.766)	0.005 (0.652)	0.004 (0.515)	0.000 (0.024)	0.007 (1.242)	0.005 (0.959)
log area						−0.004 (3.315)
log population						0.007 (3.267)
Observations	299	299	299	299	299	299
Number of periods	3	3	3	3	3	3
Number of countries	79	79	79	79	79	79
Number of instruments	2	18	18	18	16	18
Hansen test (<i>p</i> -value)	0.001	0.093	0.090	0.103	0.192	0.186
Hansen test excl. size instruments (<i>p</i> -value)	—	0.146	0.154	0.165	—	—
Difference-in-Hansen test (<i>p</i> -value) ^a	—	0.125	0.108	0.120	—	—
Kleibergen-Paap LM test (<i>p</i> -value)	< 0.001	—	—	—	—	—
Cragg-Donald Wald stat	17.47	—	—	—	—	—
<i>H</i> ₀ : <i>t</i> -test size > 25 percent (<i>p</i> -value)	< 0.001	—	—	—	—	—
Kleibergen-Paap Wald stat	15.20	—	—	—	—	—
<i>H</i> ₀ : <i>t</i> -test size > 25 percent (<i>p</i> -value)	< 0.001	—	—	—	—	—

Notes: The dependent variable in all specifications is average annual growth in GDP per capita over the period. The size instruments include log population and log area. The internal instruments refer to the lagged levels and lagged differences of endogenous right-hand side variables in the respective difference and levels equations of the dynamic panel GMM system of equations. Columns 1 and 2 are based on table 9, columns 6 and 8 of Hausmann, Hwang, and Rodrik (2007). We use Stata code and data provided by one of the authors, Jason Hwang, but the estimates in column 2 slightly differ from those reported in the published version of their paper. Following the original paper, we report heteroskedasticity-robust standard errors in parentheses and retain associated degrees of freedom adjustments for the first-stage test statistics.

^aThe null hypothesis of the Difference-in-Hansen test is that the size instruments are valid. See the notes to Table 1 for more details on the Kleibergen-Paap and Cragg-Donald tests, which apply in column 1 to the endogenous log initial EXPY.

(including a constant) treated as exogenous by Rajan and Subramanian (2008). The table reports the point estimate and standard error for β_j in each case, beginning with the estimate for Aid/GDP from the Rajan and Subramanian (2008) study. Log population has a statistically significant partial relationship with several variables that are plausible growth determinants, in addition to foreign aid. These include trade (Frankel and Romer 1999), foreign direct investment (Borensztein, De Gregorio, and Lee 1998), education expenditure (Bosworth and Collins 2003), inequality (Forbes 2000), government consumption (found to correlate with country size by

TABLE 5—LOG POPULATION IS ASSOCIATED WITH OMITTED GROWTH DETERMINANTS IN THE RAJAN AND SUBRAMANIAN SPECIFICATION, 1970–2000 CROSS-SECTION

Dependent variable	log population regressor		Observations
	Coefficient	SE	
Aid/GDP	−1.925	(0.340)	78
Trade/GDP	−13.680	(2.497)	77
FDI/GDP	−0.537	(0.183)	77
Education Expenditure/GDP	−0.423	(0.179)	75
Gini coefficient	−2.452	(0.991)	62
Government Consumption/GDP	−1.399	(0.352)	78
Manufacturing Value Added/GDP	1.529	(0.398)	76
Military Personnel/Total Labor Force	−0.263	(0.123)	78
Private Capital Flows/GDP	−2.548	(1.057)	77
Public Debt Service/GNI	−0.396	(0.229)	73
Savings/GDP	3.245	(1.502)	78

Notes: Each of the rows in the table correspond to a regression of the given dependent variable X listed in column 1 on log population and the additional covariates \mathbf{Z} other than aid/GDP in the baseline 1970–2000 cross-section specification of Rajan and Subramanian (2008), Table 4, column 2: $x_i = \beta \ln population_i + \mathbf{Z}_i' \Theta + u_i$. Only the point estimates and standard errors for log population are reported. The standard errors are robust to heteroskedasticity. The sample sizes change depending on the number of available observations for the given dependent variable, all of which come from the World Bank's *World Development Indicators 2007* (Aid/GDP, Trade/GDP, FDI/GDP, Education Expenditure/GDP, Gini Coefficient, Government Consumption/GDP, Manufacturing Value Added/GDP, Military Personnel/Total Labor Force, Private Capital Flows/GDP, Public Debt Service/GNI, and Savings/GDP).

Alesina and Wacziarg 1998, and acknowledged as a robust growth determinant by Sala-i-Martin, Doppelhofer, and Miller 2004), alongside multiple others.¹⁸

III. When Valid Instruments Are Weak

So far we have discussed cases of (mostly) strong instruments whose invalidity is difficult to detect. We turn now to cases of plausibly valid instruments whose weakness is difficult to detect.

The advent of dynamic panel GMM has been a boon to growth empiricists. These estimators take advantage of moment conditions not exploited in earlier dynamic panel two-stage least squares (2SLS) estimators. Whereas the Anderson and Hsiao (1982) estimator, for example, only exploits a single lag of endogenous right-hand side variables as instruments, the GMM estimator of Arellano and Bond (1991) (hereinafter Arellano-Bond) exploits deeper lags beyond the first or second, zeroing out lagged values that would be treated as missing in Anderson and Hsiao's 2SLS framework. Arellano and Bond's (1991) estimator, sometimes referred to as "difference" GMM, thus provides additional overidentifying restrictions without sacrificing sample size. The related system estimator of Arellano and Bover (1995) and Blundell and Bond (1998) (hereinafter Blundell-Bond)

¹⁸ A further complication arises when relaxing the assumption of linearity in the endogenous variables of interest. Although one could construct nonlinear functions of valid instruments to meet the necessary rank conditions in specifications with endogenous quadratic or interaction terms, the larger instrument set often proves weak in practice. We explore this issue further in online Appendix C using the Rajan and Subramanian (2008) framework.

imposes additional moment conditions allowing one to exclude once- or twice-lagged differences from an additional estimating equation in levels.¹⁹ Deeper lags are redundant given the Arellano-Bond moment conditions. Both estimators can accommodate additional instruments as well.

The general dynamic panel estimating equation is of the form

$$(3) \quad g_{i,t} = \beta \ln y_{i,t-1} + \mathbf{x}'_{i,t} \gamma + \psi_i + \nu_{i,t},$$

where $y_{i,t-1}$ is GDP per capita in country i at time $t - 1$ from the World Development Indicators or Penn World Table; $g_{i,t}$ is percentage growth ($\Delta \ln y_{i,t}$);²⁰ $\mathbf{x}_{i,t}$ is a vector of growth determinants; ψ_i is a country fixed effect; $\nu_{i,t}$ is an idiosyncratic shock, $t = 1, \dots, T$; and $i = 1, \dots, N$.

Arellano-Bond estimation transforms equation (3) into first-differences and exploits the moment conditions $E(\ln y_{i,t-j} \Delta \nu_{i,t}) = 0$ and $E(\mathbf{x}_{i,t-k} \Delta \nu_{i,t}) = 0$ for $t = 3, \dots, T$, $j = 2, \dots, t - 1$, and $k = k', \dots, t - 1$. While researchers commonly instrument for the lagged dependent variable to address dynamic panel bias, most have a particular interest in some possibly singleton subset of growth determinants in \mathbf{x} . Here, the literature goes one of two ways. Some authors treat that specific subset as endogenous or predetermined, where $k' = 2$ and $k' = 1$, respectively. Others treat all elements of \mathbf{x} as endogenous, in which case $k' = 2$. Another key choice concerns the number of moment conditions. Asymptotically, one would want to use the full set of lags, but as Roodman (2009b) and others show—and as we reaffirm below—such choices can have important finite-sample consequences.

Developed in response to the well-known weak instruments problem in difference GMM, the Blundell-Bond estimator augments the Arellano-Bond difference (DIF) equation with a levels (LEV) equation. Specifically, this popular estimator exploits an additional set of moment conditions, $E(\omega_{i,t} \Delta \ln y_{i,t-1}) = 0$ and $E(\omega_{i,t} \Delta \mathbf{x}_{i,t-1}) = 0$ for $t = 3, \dots, T$, where $\omega_{i,t} = \psi_i + \nu_{i,t}$ and $\mathbf{x}_{i,t}$ is assumed to be endogenously determined. These moment conditions are valid under joint mean stationarity of the $\ln y_{i,t}$ and $\mathbf{x}_{i,t}$ processes, but also under weaker albeit less plausible conditions (see Blundell, Bond, and Windmeijer 2000). This provides exclusion restrictions (based on lagged differences) for the growth determinants in equation (3) in levels. In theory, these moment conditions offer a credible identification strategy for researchers aiming to test the canonical Solow growth model or to highlight a salient source of heterogeneity in growth rates across countries.²¹

Often, however, a crucial question goes unexplored in applications of this new econometric technology. How much of the variance in the endogenous variables is

¹⁹ Caselli, Esquivel, and Lefort (1996) and Levine, Loayza, and Beck (2000) were, respectively, the first to employ the Arellano and Bond (1991) and Blundell and Bond (1998) estimators in the empirical growth literature.

²⁰ Strictly speaking, among the papers revisited here, Voitchovsky (2005) uses the dependent variable $\Delta \ln y_{i,t}$, while Levine, Loayza, and Beck (2000), Hausmann, Hwang, and Rodrik (2007), and Rajan and Subramanian (2008) instead use the (very closely related) period-average annual per capita growth rate. The two exceptions are Hauk and Wacziarg (2009) and DeJong and Ripoll (2006), whose regressand is the level $\ln y_{i,t}$, which is also amenable to a growth interpretation, given the inclusion of lagged log income on the right-hand side.

²¹ Bond, Hoeffler, and Temple (2001) characterize the appropriateness of the moment conditions in the context of estimating the Solow model, while Hauk and Wacziarg (2009) point out that, at least in theory, *exogenous* growth models do not necessarily prescribe the use of an instrumental variables framework.

explained by the instruments? A standard test for weak instruments in dynamic panel GMM regressions does not currently exist, so measuring instrument strength empirically is nontrivial.²² Until now, skeptical researchers have been mostly concerned with finite-sample biases stemming from weak instruments in the Arellano-Bond estimator and violations of the initial conditions assumption in the Blundell-Bond estimator.²³ What most have failed to address, however, is a potentially equally important problem, weak instruments in Blundell-Bond. Although generally thought to be more robust to weak instruments than difference GMM, recent work shows that this system estimator can also suffer from serious weak instrument biases (Hayakawa 2009; Bun and Windmeijer 2010). In practice, most applications of system GMM simply assume that instruments are strong. We argue that instrument strength is an empirical question that can and should be directly tested.

Below we investigate instrument strength in a variety of applications of system GMM, first in simulated data, and then in several influential growth regressions recently published in top field and general-interest journals. We follow a simple approach to assessing instrument strength in dynamic panel GMM regressions advanced *analytically* by Bun and Windmeijer (2010) and Hayakawa (2009) and *heuristically* in various settings (Blundell and Bond 2000; Dollar and Kraay 2003; Roodman 2009a; Newey and Windmeijer 2009).

Specifically, we construct the GMM instrument matrix for the difference and levels equation of the system estimator, and carry out the corresponding regressions using 2SLS.²⁴ This permits simple and transparent tests of instrument strength in a closely related setting. Blundell, Bond, and Windmeijer (2000) demonstrate that the system estimator is a weighted average of the difference and levels equations with the weights on the levels equation moments increasing in the weakness of the difference equation instruments.²⁵ So, if instrumentation of contemporaneous

²² See Stock and Wright (2000) on why the weak-instrument diagnostics for linear instrumental variables regression do not carry over to the more general setting of GMM.

²³ Bobba and Coviello (2007), for example, demonstrate that the null result in Acemoglu et al. (2005) is reversed upon augmenting the weakly instrumented difference estimator with the levels equation in the system estimator. By necessity, we discuss weak instruments in the DIF equation of the system estimator, but we explicitly leave the validity issue aside as it has been thoroughly addressed elsewhere (Roodman 2009b; Hauk and Wacziarg 2009).

²⁴ For the DIF and LEV equations, this instrument matrix, originally due to Holtz-Eakin, Newey, and Rosen (1988), takes the form (see Roodman 2009a):

$$\underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \ln y_{i1} & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & \ln y_{i2} & \ln y_{i1} & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \ln y_{i3} & \ln y_{i2} & \ln y_{i1} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}}_{\text{DIF}}; \quad \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & \dots \\ \Delta \ln y_{i2} & 0 & 0 & 0 & \dots \\ 0 & \Delta \ln y_{i3} & 0 & 0 & \dots \\ 0 & 0 & \Delta \ln y_{i4} & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}}_{\text{LEV}},$$

where, for presentational purposes, we restrict attention to the respective moment conditions $E(\ln y_{it}, t-j\Delta \nu_{it}) = 0$ and $E(\omega_{it}, \Delta \ln y_{it-1}) = 0$ in a five-period panel.

²⁵ The authors make this simple yet powerful point in the panel AR(1) model without covariates. The system estimator delivers the autoregressive point estimate of $\hat{\alpha}_s = \tilde{\delta}\hat{\alpha}_d + (1 - \tilde{\delta})\hat{\alpha}_l$, where $\hat{\alpha}_l$ is the point estimate from the LEV equation; $\hat{\alpha}_d$ is the point estimate from the DIF equation; and $\tilde{\delta} = \frac{\pi_d' \mathbf{Z}_d \mathbf{Z}_d \pi_d}{\hat{\pi}_d' \mathbf{Z}_d \mathbf{Z}_d \hat{\pi}_d + \hat{\pi}_l' \mathbf{Z}_l \mathbf{Z}_l \hat{\pi}_l}$, where $\hat{\pi}_j$ are the equivalent first-stage estimates using the instruments \mathbf{Z}_j for $j = l, d$ in the LEV and DIF equation, respectively. Their familiar setup motivates our heuristic use of the 2SLS analogues.

differences by once, twice, or multiple lagged levels are weak, and instrumentation of contemporaneous levels by lagged differences is weak, this casts great doubt on the ability of GMM estimators to yield strong identification as used in these settings. Formalizing this intuition, Bun and Windmeijer (2010) demonstrate the explicit connection between cross-sectional concentration parameters (from the familiar Stock and Yogo 2005 setup) and instrument strength in the panel 2SLS equations of the type that we estimate.

Extending the setup in Bun and Windmeijer (2010) to the common case of multiple endogenous variables, we examine whether the additional moment conditions used in system GMM are actually strong enough to compensate for the well-established weak instruments in difference GMM estimation of growth models. We appeal to the results of Blundell, Bond, and Windmeijer (2000) and Stock and Yogo (2005) in justifying our extension of the AR(1) analytics to the case of multiple endogenous regressors. In particular, we do not examine the strength of identification in the individual first-stage regressions in isolation, but rather, we test whether the instruments jointly explain enough variation in the multiple endogenous regressors to identify unbiased causal effects in the structural equation (3). These multivariate versions of the tests described in Section IIC allow us to characterize under- and weak-identification in the Blundell-Bond estimator. The rank-based LM test for underidentification due to Kleibergen and Paap (2006) readily applies to the panel 2SLS context here. Bun and Windmeijer (2010) provide evidence that the weak-instruments testing methods derived in the cross-section “are also informative about absolute and relative 2SLS bias when exploiting the whole panel.” Although these tests should be considered heuristic in the panel setting considered here, their use is certainly preferable to ignoring the problem.²⁶

Despite the large number of instruments in several specifications considered below, the Stock and Yogo (2005) weak instruments tests offer a powerful diagnostic tool.²⁷ Their critical values are available for up to 100 instruments. Those critical values exhibit a slow rate of decay as instruments increase beyond 30 or 40, and numerical results suggest their procedure is consistent for any number of instruments (Stock and Yogo 2005, 90). According to Stock and Yogo, “Viewed as a test, the procedure has good power, especially when the number of instruments is large.”

A. Monte Carlo Results

Our first step is to show that the system GMM estimator can often have poor size and power properties, depending crucially on the extent of endogeneity and on the strength of instrumentation. This is a different focus than the simulations in

²⁶One reason for caution is that in panels of the type studied here, both cross-section heteroskedasticity and time-series heteroskedasticity are likely, which means that the conventional F -statistic is problematic (see Bun and de Haan 2010).

²⁷Hall, Inoue, and Shin (2008) develop a method for using the Stock and Yogo (2005) diagnostics to select optimal instruments in GMM regressions via examination of corresponding 2SLS regressions. Their results show that a Stock and Yogo (2005) pretest can actually be more powerful than using weak-instrument robust inference procedures (see Section IIIE) with the full set of possibly suboptimal instruments.

Blundell, Bond, and Windmeijer (2000) and Bun and Windmeijer (2010), though we follow their specification, which differs slightly from equation (3). We simulate

$$(4) \quad y_{i,t} = \beta y_{i,t-1} + \gamma d_{i,t} + \psi_i + \nu_{i,t}$$

$$d_{i,t} = \zeta d_{i,t-1} + \theta_i + \phi_{i,t} \quad t = 1, \dots, 6; \quad i = 1, \dots, 100,$$

where the initial conditions $y_{i,0} = (\psi_i + (\gamma\theta_i/(1-\zeta)))/(1-\beta) + \nu_{i,0}$ and $d_{i,0} = (\theta_i)/(1-\zeta) + \phi_{i,0}$ are sufficient to impose mean stationarity, an assumption on which the consistency of the Blundell-Bond estimator is predicated. The errors are distributed as

$$(5) \quad \nu_{i,t}, \phi_{i,t} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \omega \\ \omega & \sigma^2 \end{pmatrix}\right) \quad \text{and} \quad \psi_i, \theta_i \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

The correlation coefficient for the shocks is $\rho = \frac{\omega}{\sigma^2}$. All simulation results employ the Windmeijer (2005) two-step correction, cluster standard errors by groups i , include time dummies in all equations, treat $y_{i,t-1}$ and $d_{i,t}$ as endogenous, and include the full set of available lags in the difference equation instrument matrix.²⁸

Figure 1 shows results from this simulation with $\gamma = 0.3$ and $\beta = 0.2$ based on 500 repetitions. The horizontal axis shows different assumed values of $\zeta \in \{0.1, 0.2, \dots, 0.9\}$, indicating the persistence of d over time, and the vertical axis compares the estimated $\hat{\gamma}$ (solid black line) to the true γ (dotted red line). The dashed lines show the average 95 percent confidence interval on $\hat{\gamma}$ across all repetitions. The top part of the figure shows the results for the difference GMM estimator, the bottom part for the system GMM estimator. Each small panel of the figure shows a different combination of the extent of endogeneity $\omega \in \{-0.1, -0.5, -0.9\}$, and the shock variance $\sigma^2 \in \{0.1, 0.5, 1, 5, 10\}$.

The magnitude of σ^2 , which implicitly captures the ratio of the variance in idiosyncratic shocks to the variance in country fixed effects, has fundamental effects on instrument strength. While the theoretical apparatus in Blundell and Bond (1998) presumes that $\sigma^2 = 1$, it is more likely that $\sigma^2 < 1$ for typical applications in the empirical growth literature including those we consider in the next section. That is, the time-invariant heterogeneity in income levels across countries is likely to swamp the within-country variation in idiosyncratic shocks. The theoretical channel from low σ^2 to weak instruments is borne out in Bun and Windmeijer (2010), among others.

The performance of the difference GMM estimators is poor. In no case does the estimate of $\hat{\gamma}$ both reject the hypothesis that $\gamma = 0$ and fail to reject the hypothesis that $\gamma = 0.3$. For the more negative values of ω , bias is so extensive that the true value of γ is often rejected. The downwardly biased difference GMM estimates are consistent with the Monte Carlo findings in Blundell and Bond (1998), although

²⁸Note that this setup is consistent with the growth formulation in equation (3) after simply “relabeling” equation (4): relabeling $y_{i,t}$ with $\ln y_{i,t}$, subtracting $\ln y_{i,t-1}$ from both sides, and relabeling β with $\tilde{\beta} = \beta - 1$. Our variance formulation is analogous to the factor loadings representation of endogeneity in Blundell, Bond, and Windmeijer (2000).

our results seem to imply biases even at quite low levels of persistence in d and y . The system GMM estimator performs better: The estimate of $\hat{\gamma}$ only rejects the true value when ζ is low, that is when d is not sufficiently persistent over time. It is able to reject the hypothesis $\gamma = 0$, but only for high levels of $\zeta > 0.6$, which is consistent with the original motivation for the system estimator in Blundell and Bond (1998). Additionally, whereas the difference GMM estimator is unaffected by the magnitude of σ^2 , the system GMM estimator performs poorly for low σ^2 when the degree of endogeneity is not extreme ($\omega > -0.9$).

Figure 2 suggests that these biases are related to problems of weak- or under-identification. Here, the vertical axis shows the p -value from the Kleibergen-Paap LM test of the null hypothesis that the 2SLS regressions for the difference and levels equations corresponding to Figure 1 are underidentified or rank-deficient. In the upper part of Figure 2, current differences are instrumented by the same GMM instrument matrix of lagged levels used in the difference GMM estimates of Figure 1. In the lower part of Figure 2, current levels are instrumented by the same matrix of lagged differences used in the levels equation of the system GMM estimates in Figure 1. LM test p -values greater than 0.1 (or more conservatively 0.05) point to potentially severe underidentification.

A clear pattern emerges. When instrumentation is weak in the 2SLS equations of Figure 2, the performance of the corresponding difference and system GMM estimates of γ is poor in Figure 1. When instrumentation in the 2SLS levels equation is strong (e.g., when $\zeta > 0.6$ and $\sigma^2 > 1$), the estimates $\hat{\gamma}$ show excellent size and power properties. In online Appendix D, we repeat the same exercise with $\beta = 0.8$, so that y is more persistent over time, with essentially the same result.

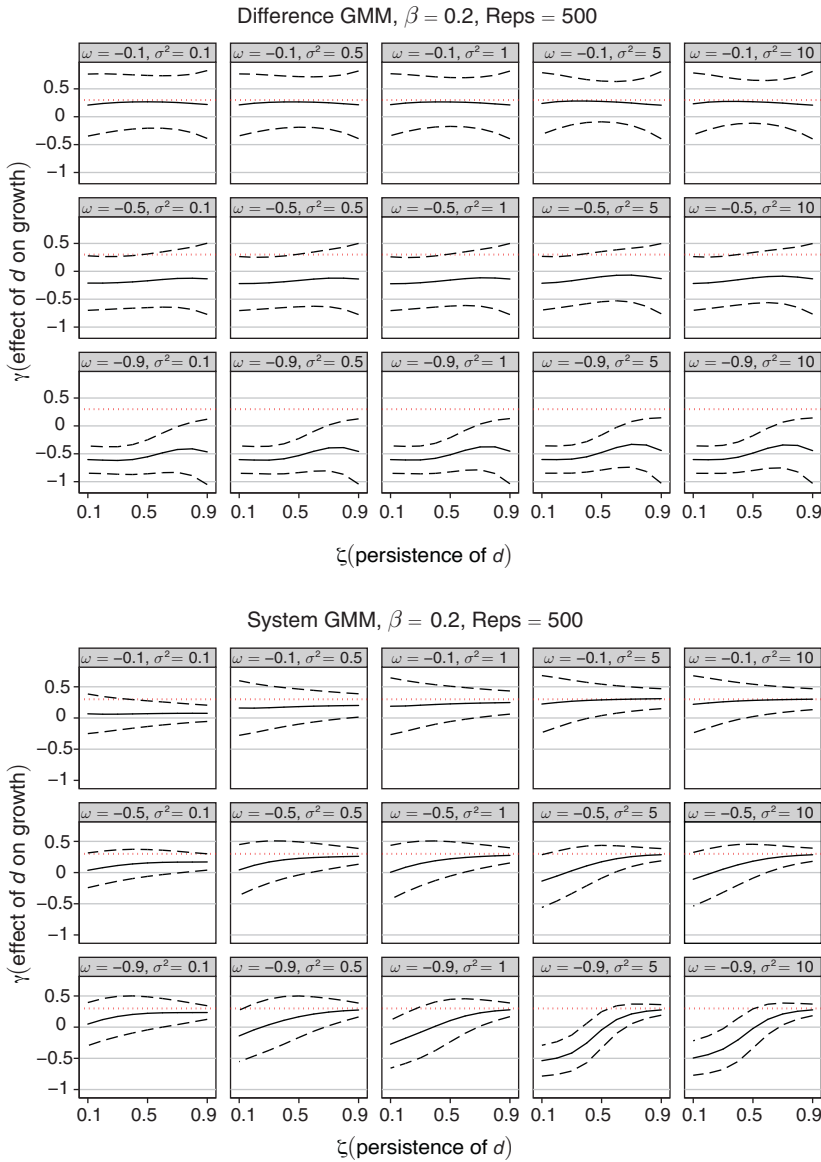
Figures 1 and 2 are sobering. Under reasonable parameter assumptions, the system GMM estimator is capable of leading a researcher to spurious conclusions—that d does not cause growth when it does, or that d has a negative effect on growth when the true effect is positive. A major part of the problem appears to be that in many cases there is no good reason to believe that lagged levels of the regressors explain a large portion of the variance in current differences, or vice versa. In these simulation results, this is transparent by construction. We now proceed to illustrate that this concern may be far from hypothetical and may apply to recently published growth regressions.

B. Financial Intermediation: Abundant Instruments versus Strong Instruments

Table 6 revisits the dynamic panel GMM results of Levine, Loayza, and Beck (2000) using the original data.²⁹ Column 1 reproduces a representative regression of growth on “liquid liabilities” (their Table 5, column 1). Column 2 gives the results of the closest reproduction of this regression we could achieve using the authors’ original dataset, and the results match relatively well.³⁰ Again, we test for

²⁹ Levine, Loayza, and Beck (2000) conduct similar regressions with three different endogenous measures of financial intermediation. See online Appendix D for a discussion of similar results using the other two measures.

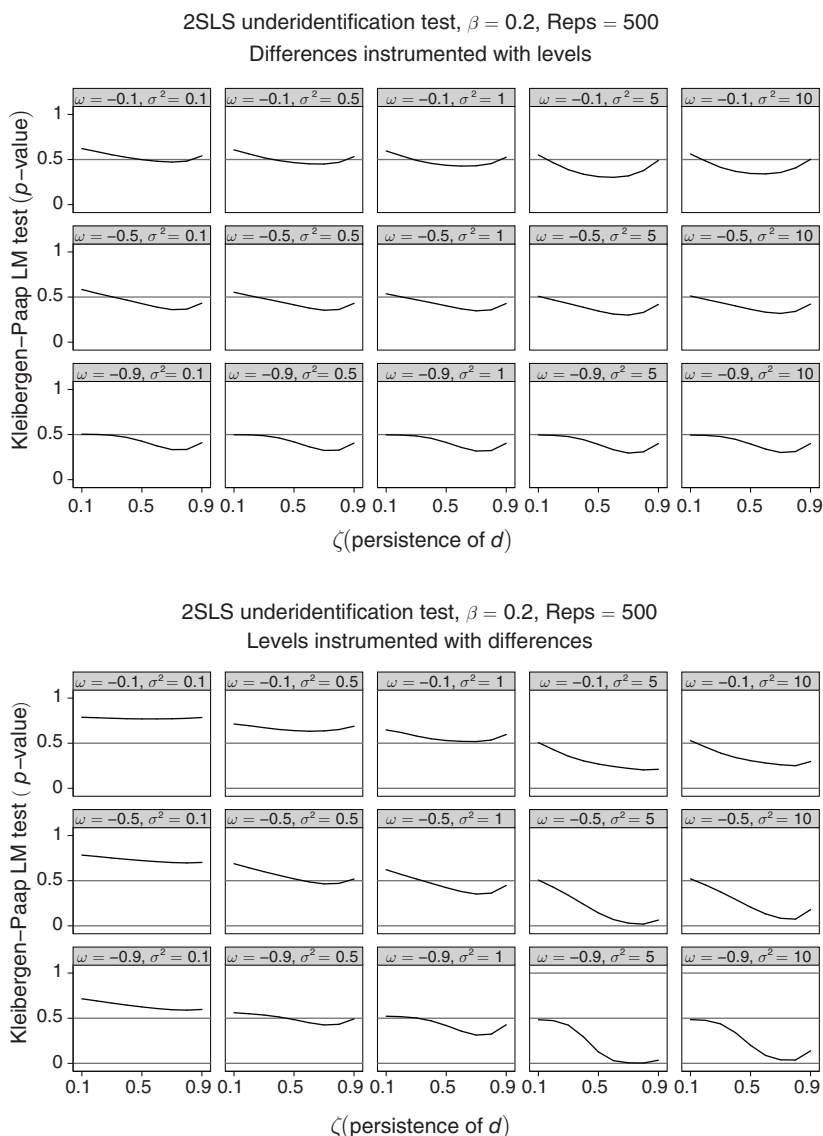
³⁰ Our replication uses the original DPD96 Gauss program employed by the authors. The remaining specifications in the table use Stata software. The number of observations reported by Gauss differs from that reported by Stata (compare columns 2 and 3) for reasons discussed in online Appendix E, where we also provide details on our attempted replication of their results including a full elaboration of the point estimates suppressed in column 2 of Table 6.

FIGURE 1. POWER AND SIZE PROPERTIES OF GMM ESTIMATORS IN SIMULATION RESULTS, $\beta = 0.2$

Notes: The graphs show parameter estimates and 95 percent confidence intervals from simulations of the model in equation (4) based on 500 draws of a sample size of 600 with 100 cross-sectional units and six time periods, fixed $\beta = 0.2$, varying $\zeta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, varying degrees of endogeneity $\omega \in \{-0.1, -0.5, -0.9\}$, and alternative variances of the idiosyncratic shock, $\sigma^2 \in \{0.1, 0.5, 1, 5, 10\}$, where the variance of cross-sectional heterogeneity is fixed at 1. The dashed red line shows the true value of $\gamma = 0.3$ in the simulations.

underidentification (Kleibergen-Paap LM test) and weak instruments (Cragg-Donald and Kleibergen-Paap Wald tests).

Column 3 carries out the same regression using simple pooled OLS. In columns 4 and 5, we purge the country fixed effects from the regression by first-differencing (FD) and within-transformation (FE). While weak instruments typically bias difference GMM estimates downward, Bun and Windmeijer (2010) demonstrate how

FIGURE 2. WEAK IDENTIFICATION IN SIMULATION RESULTS, $\beta = 0.2$

Notes: The graphs show p -values from a Kleibergen-Paap LM test for (the null of) underidentification in the levels and differences equations from simulations of the model in equation (4) as detailed in the notes to Figure 1. See the notes to Table 1 for details on the Kleibergen-Paap test.

system GMM estimates can be biased upward. This bias increases in the ratio of the variance of the fixed effects to the variance of the idiosyncratic shocks. Recall that our simulation results in the preceding section similarly present the least biases for low values of this ratio, or high values of σ^2 . In column 5, the estimated ratio of variances is approximately 5 (i.e., $\hat{\sigma}_\nu^2/\hat{\sigma}_\psi^2 \approx 0.2$ in equation (3)).³¹ Column 6 regresses

³¹ We estimate these variance terms using the Baltagi and Chang (1994) method, which typically exhibits superior finite sample performance in unbalanced panels such as those commonly used in the growth literature.

difference growth on differenced regressors, instrumented by lagged regressor levels analogous to the difference GMM estimator. Both the Kleibergen-Paap LM test of underidentification and the Cragg-Donald and Kleibergen-Paap Wald-type statistics show that instrumentation is very weak, far too weak for instrumentation to remove a substantial portion of OLS bias.³²

An additional problem lurks below the surface: The sample contains 74 countries, and 75 different instrumental variables are used in the system estimator.³³ The large number of instruments relative to the number of groups may actually result in a failure to expunge the endogenous components of the right-hand side variables, thereby biasing the coefficient estimates towards those from the OLS estimator (see Beck and Levine 2004; Calderón, Chong, and Loayza 2002; Roodman 2009b). In the limiting case, a 2SLS regression that has one instrument for each observation would show strong instrumentation but would produce coefficients exactly equal to those produced by OLS, and would not address endogeneity bias at all. The problem is perhaps even more serious in panels in which the cross-sectional variation dominates the within variation, as is common in growth regressions. Until recently, the literature has offered little guidance on the appropriate number of instruments relative to the number of groups and time periods.

Roodman (2009b) discusses a practical method for addressing this problem of “too many instruments” in dynamic panel GMM estimation. He suggests first restricting the number of lagged levels used in the instrument matrix for the difference equation, but since Levine, Loayza, and Beck (2000) restrict their original matrix to a single lag, we must try an alternative approach. By “collapsing” the instrument matrix, we can effectively combine the instruments into smaller sets while retaining the same information from the original 75 column instrument matrix. The “collapsed” matrix contains one instrument for each lag depth instead of one instrument for each period *and* lag depth as in the conventional dynamic panel GMM instrument matrix.³⁴ Roodman suggests that a liberal rule of thumb is to become concerned when the number of instruments is close to the number of groups, as in the present case. Column 7 shows the results with the instrument matrix collapsed. Again, we cannot reject the null

³² It is worth noting that the extremely high *p*-values we obtain in a number of specifications (i.e., failure to reject the null of weak instruments) are not uncommon (see Yogo 2004). Nor are they indicative of underpowered or biased tests as, for example, a *p*-value of one in a Hansen test of overidentifying restrictions may be in the presence of “too many instruments” (see Bowsher 2002; Roodman 2009b).

³³ In both the levels and difference equations, 35 lagged regressors are used as instrumental variables—1 for each of the 7 endogenous right-hand side variables in each of the 5 periods—along with the 5-period dummies included in the main equation.

³⁴ Collapsing leads to the following changes in the general, full DIF and LEV instrument matrices in footnote 25,

$$\underbrace{\begin{bmatrix} 0 & 0 & 0 & \dots \\ \ln y_{i1} & 0 & 0 & \dots \\ \ln y_{i2} & \ln y_{i1} & 0 & \dots \\ \ln y_{i3} & \ln y_{i2} & \ln y_{i1} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}}_{\text{DIF-Collapsed}}; \quad \underbrace{\begin{bmatrix} 0 & \dots \\ \Delta \ln y_{i2} & \dots \\ \Delta \ln y_{i3} & \dots \\ \Delta \ln y_{i4} & \dots \\ \vdots & \ddots \end{bmatrix}}_{\text{LEV-Collapsed}},$$

where the first column in DIF-Collapsed corresponds to the first lag collapsed across periods 3–5, the second column to the second lag collapsed across periods 4–5, etc.

TABLE 6—WEAK INSTRUMENTS IN DYNAMIC PANEL REGRESSIONS USING LIQUID LIABILITIES
IN LEVINE, LOAYZA, AND BECK (2000)

Estimator	GMM-SYS	GMM-SYS	OLS	OLS-FD	OLS-FE
Collapsed IV matrix	No (1)	No (2)	— (3)	— (4)	— (5)
Liquid liabilities	2.952 (0.001)	2.834 (0.001)	1.692 (0.000)	1.095 (0.122)	0.851 (0.296)
log initial GDP/capita	−0.742 (0.001)	−0.792 (0.001)	−0.400 (0.025)	−13.609 (0.000)	−7.478 (0.000)
<i>Other parameter estimates omitted</i>					
Observations	359	359	345	323	345
Number of countries	74	74	74	74	74
Number of instruments	75	75	—	—	—
IV: Lagged levels	Yes	Yes	—	—	—
IV: Lagged differences	Yes	Yes	—	—	—
	(6)	(7)		(8)	(9)
		Difference equation		Levels equation	
Estimator		2SLS	2SLS	2SLS	2SLS
Collapsed IV matrix		No	Yes	No	Yes
Liquid liabilities		−0.747 (0.705)	−15.403 (0.702)	2.830 (0.002)	2.285 (0.321)
log initial GDP/capita		−12.435 (0.000)	−12.335 (0.355)	0.339 (0.619)	1.839 (0.423)
<i>Other parameter estimates omitted</i>					
Observations		323	323	345	345
Number of countries		74	74	74	74
Number of instruments		40	12	40	12
IV: Lagged levels		Yes	Yes	No	No
IV: Lagged differences		No	No	Yes	Yes
Kleibergen-Paap LM test (<i>p</i> -value)		0.780	0.580	0.559	0.200
Cragg-Donald Wald stat		0.59	0.04	0.72	0.25
H_0 : relative OLS bias > 30 percent (<i>p</i> -value)		1.000	1.000	1.000	0.997
Kleibergen-Paap Wald stat		0.67	0.06	1.12	0.25
H_0 : relative OLS bias > 30 percent (<i>p</i> -value)		1.000	1.000	1.000	0.998

Notes: The dependent variable in all specifications is average annual growth in GDP per capita each period. Column 1 reproduces the published version of Levine, Loayza, and Beck (2000, table 5, columns 1 and 2) and reports our best attempted replication using the `DPD96` program for Gauss, the publicly available dataset, and a Gauss program used to generate their results provided by Thorsten Beck. Further details on the difference in sample sizes across columns, our replication efforts, and the associated differences in the Gauss and Stata programs for dynamic panel GMM regressions can be found in online Appendix E. The following variables are included in the regressions but suppressed in the table here for presentational purposes: government size, openness to trade, inflation, average years of secondary schooling, black market premium, time period dummies and a constant. The first five of these variables are treated as endogenous. Following the original paper, we report *p*-values in parentheses. See the notes to Table 1 for more details on the Kleibergen-Paap and Cragg-Donald tests, which apply in columns 6–9 to the full set of endogenous right-hand-side variables.

of underidentification, and weak instruments imply substantial bias of the 2SLS estimates relative to pure OLS. Weak identification is not an artifact of too many instruments. Instrumentation this weak—no matter how valid—is incapable of testing hypotheses about coefficients in the main regression.

To test for weak instruments in the system estimator, we must also examine the levels equation independently of, but in the same manner as, the difference

equation. Columns 8 and 9 conduct this parallel exercise for the levels equation. Since the difference equation is so weakly instrumented, the burden of strong identification in the system estimator relies on the levels equation moments. In column 8, the level of growth is regressed on the level of the regressors in a two-stage least squares framework, instrumented by the same lagged differences, as in the levels equation of the system GMM estimator. Once again, instrumentation is too weak to address any substantial portion of OLS bias, thereby casting doubt on the system GMM point estimate for liquid liabilities in column 2, which is remarkably close to that for the levels equation in column 8. Using a collapsed instrument matrix in column 9 leaves these primary conclusions unchanged.

C. Weak Aid or Weak Instruments?

Table 7 repeats this analysis for an entirely different set of regressions. It revisits the dynamic panel results of Rajan and Subramanian (2008) using the original data. Columns 1 and 2 exactly replicate their main Arellano-Bond (table 9, column 1) and Blundell-Bond (table 10, column 1) results. Column 3 shows the simple pooled OLS result, which appears quite similar to the system estimate in the preceding column. Columns 4 and 5 purge country fixed effects from the regression in column 3 via first-differencing (FD) and within-transformation (FE) with results similar to those for the Arellano-Bond estimator in column 1. Given that the estimated ratio of the variance of the time-invariant individual effects to the variance of idiosyncratic shocks is around three in column 5, this evidence suggests that instrumentation in these dynamic panel GMM regressions may be too weak to improve upon OLS.

Following the approach above, in column 6, we estimate the difference component of the system estimator in a 2SLS regression with exactly the same sequential moment conditions. Using the Kleibergen-Paap LM test, we cannot reject the null of underidentification, suggesting that identification is too weak to conduct meaningful hypothesis tests based on the difference equation alone. Although the Kleibergen-Paap Wald statistic appears high, and we can reject large relative OLS bias on the basis of Stock and Yogo diagnostics, the perceived strength turns out to be a statistical artifact of the large, unrestricted GMM instrument matrix. After collapsing the 120 column instrument matrix, the Kleibergen-Paap Wald statistic falls dramatically in column 7, and we cannot reject the null that the difference equation exhibits more than 30 percent of the OLS bias. Columns 8 and 9 repeat the same exercise for the levels equation of the system estimator. Column 8 demonstrates underidentification and weak instruments in the standard wide instrument matrix, and collapsing does little to help. These results suggest that the similarity between the biased OLS estimates in columns 3–5 and the dynamic panel GMM estimates in columns 1 and 2 is not a coincidence. Weak instruments in both the difference and levels equations render hypothesis tests on the system GMM point estimate for aid/GDP unreliable.

D. Beyond Aid and Credit

The findings above are not peculiar to the specifications used in these two studies. In this subsection, we further examine the weak instruments problem in other

TABLE 7—WEAK INSTRUMENTS IN DYNAMIC PANEL REGRESSIONS OF RAJAN AND SUBRAMANIAN (2008)

Estimator	GMM-DIF	GMM-SYS	OLS	OLS-FD	OLS-FE
Collapsed IV matrix	No (1)	No (2)	— (3)	— (4)	— (5)
Aid/GDP	−0.151 (0.077)	−0.054 (0.114)	−0.037 (0.053)	−0.236 (0.066)	−0.224 (0.067)
Initial log GDP/capita	−8.347 (1.543)	−2.456 (1.057)	−1.514 (0.517)	−13.245 (1.839)	−7.960 (1.307)
<i>Other parameter estimates omitted</i>					
Observations	359	359	345	323	345
Number of countries	74	74	74	74	74
Number of instruments	75	75	—	—	—
IV: Lagged levels	Yes	Yes	—	—	—
IV: Lagged differences	Yes	Yes	—	—	—
Lags used	2nd–7th	2nd–7th	—	—	—
	(6)	(7)		(8)	(9)
Difference equation			Levels equation		
Estimator	2SLS	2SLS	2SLS	2SLS	
Collapsed IV matrix	No	Yes	No	Yes	
Aid/GDPs	−0.220 (0.086)	−0.355 (0.157)	0.116 (0.079)	0.470 (0.710)	
Initial log GDP/capita	−11.060 (1.980)	−10.535 (3.355)	0.117 (1.454)	10.193 (15.689)	
<i>Other parameter estimates omitted</i>					
Observations	167	167	239	239	
Number of countries	68	68	72	72	
Number of instruments	120	52	41	17	
IV: Lagged levels	Yes	Yes	No	No	
IV: Lagged differences	No	No	Yes	Yes	
Lags used	2nd–7th	2nd–7th	2nd	2nd	
Kleibergen-Paap LM test (p -value)	0.522	0.698	0.765	0.413	
Cragg-Donald Wald stat	0.66	0.43	0.41	0.06	
H_0 : relative OLS bias > 30 percent (p -value)	1.000	1.000	1.000	1.000	
Kleibergen-Paap Wald stat	9.89	1.36	0.69	0.07	
H_0 : relative OLS bias > 30 percent (p -value)	< 0.001	1.000	1.000	1.000	

Notes: The dependent variable in all specifications is average annual growth in GDP per capita each period. Column 1 exactly replicates Rajan and Subramanian (2008, table 9, column 1), and column 2 exactly replicates Table 10, column 1 in Rajan and Subramanian. The following variables are included in the regressions but suppressed in the table here for presentational purposes: life expectancy, institutional quality, log inflation, M2/GDP, budget balance/GDP, revolutions, ethnic fractionalization, geography, time period dummies, dummies for countries in sub-Saharan Africa, East Asia, and a constant. The first six of these variables are treated as endogenous. Following the original paper, we report heteroskedasticity-robust standard errors in parentheses and retain associated degrees of freedom adjustments for the first-stage test statistics. See the notes to Table 1 for more details on the Kleibergen-Paap and Cragg-Donald tests, which apply in columns 6–9 to the full set of endogenous right-hand-side variables.

recently published empirical applications. The goal is to highlight features of the data and panel setup that give rise to different outcomes in terms of the strength of identification. Table 8 reports weak instruments diagnostics for the baseline system GMM specifications in four studies published within the last five years. As before, we emphasize not the particular findings of each study but rather the sources and

quality of identification. For each study, we report results based on the Kleibergen and Paap LM test for underidentification and the weak instrument tests based on the Kleibergen-Paap and Cragg-Donald Wald statistics for the 2SLS estimates of the difference (DIF) and levels (LEV) equations separately with the full and collapsed instrument matrices.³⁵

The first panel of the table unpacks the Blundell-Bond estimates of the Solow growth equations in Table 13 of Hauk and Wacziarg (2009). They treat all augmented Solow regressors—physical capital, human capital, population, and lagged income—as endogenous, and instrument with the full set of available lags in the difference equation. In only one specification—the levels equation with a collapsed instrument matrix—do we fail to reject the null of underidentification. Yet, weak instruments still afflict the system GMM estimates, as we cannot reject the null hypothesis that the 2SLS estimates maintain a nontrivial portion of the OLS bias. Compared with results in the previous subsections, however, it seems that the instruments explain some of the variation in the four endogenous variables in the canonical Solow model estimated over a sufficiently long panel.

Next, we return to Hausmann, Hwang, and Rodrik's (2007) results, examining their longer panel employing a five-year periodization. The longer panel affords a richer degree of within-country variation.³⁶ We can reject the null of underidentification in both the difference and levels equations based on the full, unrestricted instrument matrix. However, we cannot rule out underidentification when collapsing the instrument matrix for the difference equation. Nor can we rule out that weak instruments leave much of the OLS bias in the four 2SLS specifications. This is concerning since plausibly invalid country size instruments account for a nontrivial amount of the instrument strength captured by the underidentification and weak instrument test statistics and, especially under failures of validity, the 2SLS bias can be worse than the OLS bias. Using the Hahn, Ham, and Moon (2011) test for instrument validity, we strongly reject the validity of the lagged difference in log population as an instrument in the 2SLS levels equation specifications.

The third paper we consider is due to Voitchovsky (2005) who analyzes the impacts of inequality on economic growth using a short unbalanced panel of OECD countries from 1970–1995. We report our diagnostics for the baseline specification in column 4, Table 2 (Voitchovsky 2005, 287), in which five growth determinants are treated as endogenous—lagged income per capita, contemporaneous investment, lagged schooling, lagged Gini coefficient, and lagged ratio of the 90/75th percentile of the income distribution.³⁷ The Kleibergen-Paap LM test suggests that the DIF equation is underidentified. Yet, the large Kleibergen-Paap Wald statistic allows us to reject the null that weak instruments bias the 2SLS point estimates. This seeming anomaly disappears when collapsing the instrument matrix and hence

³⁵Our replications of Hauk and Wacziarg (2009); Hausmann, Hwang, and Rodrik (2007); and DeJong and Ripoll (2006) are exact, relying on the original data and code provided by the authors. Our replication of Voitchovsky (2005), for which original code is unavailable, yields slightly different results than the published version. See online Appendix E.

³⁶The ratio of the variance in country fixed effects to the variance in idiosyncratic heterogeneity is approximately one compared to the shorter panel with ten-year periodization (see Table 4), where that ratio exceeds two.

³⁷See online Appendix E for a discussion of the movement conditions used.

TABLE 8—CHARACTERIZING WEAK- AND UNDER-IDENTIFICATION
IN OTHER DYNAMIC PANEL GROWTH REGRESSIONS

Equation	Sample size	No. of endog. vars.	No. of instruments	KP LM test ^a (<i>p</i> -test)	CD <i>F</i> -Stat	rel. OLS bias > 30 percent (<i>p</i> -value)	KP <i>F</i> -Stat	rel. OLS bias > 30 percent (<i>p</i> -value)
Hauk and Wacziarg (2009): panel with 69 countries, 8 periods								
DIF	414	4	102	0.075	1.84	1.000	4.34	0.002
DIF-Collapsed	414	4	27	0.012	1.67	0.995	2.43	0.861
LEV	483	4	27	0.042	1.74	0.992	1.83	0.987
LEV-Collapsed	483	4	4	0.161	0.94	0.783	0.49	0.913
Hausmann, Hwang, and Rodrik (2007) ^a : panel with 79 countries, 8 periods								
DIF	525	3	56	0.042	2.23	0.992	2.78	0.848
DIF-Collapsed	525	3	14	0.244	1.70	0.930	0.86	0.999
LEV	604	3	20	0.001	2.81	0.596	3.42	0.288
LEV-Collapsed	604	3	5	0.015	1.82	0.508	2.10	0.431
Voitchovsky (2005): panel with 21 countries, 5 periods								
DIF	61	5	24	0.641	0.56	1.000	103.8	< 0.001
DIF-Collapsed	61	5	6	0.932	0.01	1.000	0.02	1.000
LEV	82	2	11	0.318	0.65	0.999	1.18	0.978
LEV-Collapsed	82	2	2	0.013	2.73	0.312	3.70	0.184
DeJong and Ripoll (2006): panel with 60 countries, 5 periods								
DIF	20	8	64	0.257	0.81	1.000	73.4	< 0.001
DIF-Collapsed	200	8	28	0.397	0.63	0.995	0.87	1.000
LEV	260	8	28	0.830	0.39	1.000	0.69	1.000
LEV-Collapsed	260	8	8	0.107	0.22	0.999	0.32	1.000

Notes: We follow the original papers in utilizing heteroskedasticity-robust standard errors and associated degrees of freedom adjustments for the first-stage test statistics. All estimates are obtained using 2SLS. Details on the DIF- and LEV-(Collapsed) instrument matrices are provided in the text. In the LEV-Collapsed row for Voitchovsky (2005), we use the critical values for the bias test based on three instruments since two instrument critical values cannot be calculated for the case of two endogenous variables. See the notes to Table 1 for more details on the Kleibergen-Paap and Cragg-Donald tests, which apply in each regression to the full set of endogenous right-hand-side variables described in the text.

^aThis study includes additional “external” instruments, log population and log area, which affect the diagnostic tests. Treating the “internal” GMM instruments as potentially weak and the external size instruments as likely strong (see Table 4), we apply the Hahn, Ham, and Moon (2011) test of the null hypothesis that the size instruments are valid and find *p*-values of 0.188 for the DIF equation, 0.228 for the DIF-Collapsed equation, < 0.001 for the LEV equation, and < 0.001 for the LEV-Collapsed equation. These results suggest that log population provides a valid instrument in the difference equations, but the difference in log population fails the exclusion restriction in the levels equations. Difference-in-Hansen tests, although not robust to weak instruments, yield similar insights.

can be explained by the large number of instruments (24) relative to the number of countries (21). While we cannot reject that the original LEV equation is underidentified, collapsing the instrument matrix delivers an exactly identified IV equation in which the two instrumental variables pass the LM rank test.

Lastly, we examine the system GMM estimates from DeJong and Ripoll (2006), a study arguing that the relationship between trade openness and economic growth depends on initial income. We consider the authors’ baseline estimates from the fourth column in table 2 of DeJong and Ripoll (2006, 631). The regressions examine *eight* growth determinants: life expectancy, female schooling, male schooling,

lagged income per capita, ad valorem tariffs (import duties as share of imports), tariffs \times initial income/capita, investment/GDP, and government spending/GDP. The first four growth determinants are treated as predetermined and the latter four as endogenous. Regardless of the specification, we fail to reject that the structural equation is underidentified, casting doubt on the ability of the system GMM estimator to solve the weak instruments problem evident in these 2SLS regressions.³⁸

Collectively, the simulation results and analysis of the six papers considered above sound a warning note about the credibility of unexamined growth empirics using difference and system GMM estimation. We have shown that with a weakly instrumented levels equation, system GMM estimates can exhibit biases similar in magnitude to uncorrected OLS variants. However, unlike the initial conditions restrictions on which the system GMM estimator is predicated, weak instruments can be diagnosed and (partially) addressed in many settings.

E. Weak-Instrument Robust Inference

In this final subsection, we go beyond documenting the pervasiveness of weak instruments in system GMM to characterize the implications for inference about parameters of interest. Using the Hausmann, Hwang, and Rodrik (2007) and Hauk and Wacziarg (2009) studies as examples, we conduct weak-instrument robust inference on the 2SLS difference and levels equations.³⁹ With a single endogenous variable, the CLR approach of Moreira (2003), which we used in Section IIC, permits inference that is immune to the damaging effects of weak instruments. However, other methods are required for regressions with multiple endogenous variables.

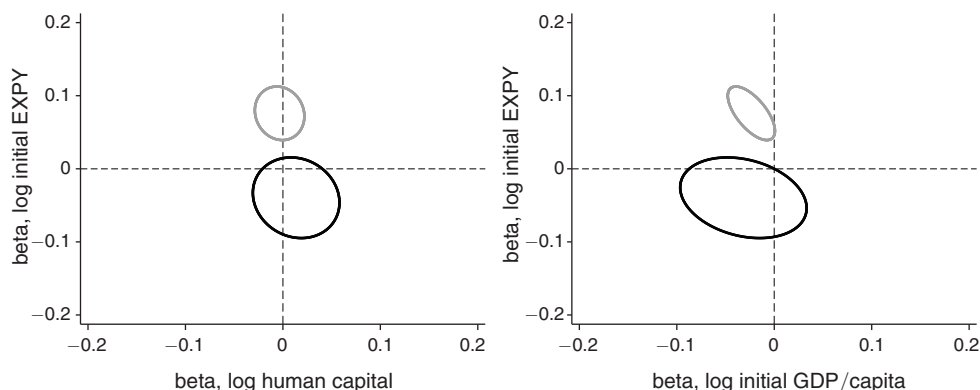
Here, we utilize the Kleibergen (2002) testing procedure, which has better power properties than the conventional Anderson and Rubin (1949) test in the presence of many instruments—the norm in the dynamic panel context. We describe the testing procedure in detail in online Appendix B. Using the resulting *K*-statistic, we can derive joint confidence sets for multiple endogenous variables. Although computationally intensive, the Kleibergen procedure is robust not only to (many) weak instruments but also to invalid instruments (Doko and Dufour 2008). This is important given concerns about the validity of the moment conditions in the levels equation of system GMM.

Figure 3, panel A plots two-dimensional weak-instrument robust confidence sets for subsets of the three endogenous variables in the Hausmann, Hwang, and Rodrik (2007) five-yearly panel: log initial export diversity (EXPY), log human capital, and log initial GDP per capita. The 95 percent confidence ellipses in the graphs represent (approximately) the boundary of the maximal area level set over the third endogenous variable in the full three-dimensional confidence ellipsoids. In the top graph, we cannot reject that both log initial export diversity and human capital have zero effect on economic growth in the 2SLS difference equation.

³⁸ Not unlike the Voitchovsky (2005) result, the large Kleibergen-Paap Wald statistic in the DIF equation disappears when using the collapsed instrument matrix and is incongruent with the more reliable underidentification test.

³⁹ Given our illustrative purposes here, we use these two studies for reasons of computational practicality. Their relatively small number of endogenous variables are more amenable to the test procedure we deploy here than the large number of endogenous variables in some of the other studies considered above.

Panel A. Hausmann, Hwangand, and Rodrik (2007)



Panel B. Hauk and Wacziarg (2009)

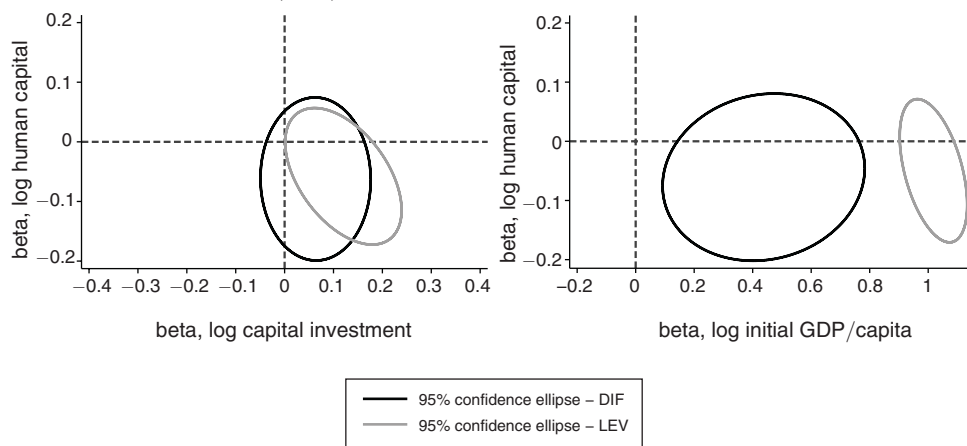


FIGURE 3. WEAK-INSTRUMENT ROBUST CONFIDENCE SETS

Notes: The graphs in panel A (quadrants I and II) are the 95 percent weak-instrument robust confidence ellipses for two of the three endogenous variables in the 2SLS analogues of the difference (DIF) and levels (LEV) equations in the system GMM estimates of the dynamic panel regressions in Hausmann, Hwang, and Rodrik (2007). The confidence regions are obtained through a three-dimensional grid-search procedure over the domain -0.2 to 0.2 at increments of 0.01 for each of the three variables. The graphs in panel B (quadrants III and IV) are the 95 percent weak-instrument robust confidence ellipses for two of the three endogenous variables in the 2SLS analogues of the difference (DIF) and levels (LEV) equations in the system GMM estimates of the dynamic panel regressions in Hauk and Wacziarg (2009). The confidence regions are obtained through a four-dimensional grid-search procedure over a domain comprising the original point estimates and exceeding zero from above or below. The domain spans the x - and y -axis for the variables shown in these graphs. The procedure is based on the approach developed in Kleibergen (2002) (see online Appendix B), which is shown to have higher power than the more familiar Anderson-Rubin test in the presence of many instruments. The ellipses are means-centered with a boundary constant of 4.

Turning to the levels equation, however, we cannot reject the null hypothesis that log initial export diversity has a positive effect. The same general pattern for export diversity holds when examining the two-dimensional confidence ellipse with log initial GDP per capita in the bottom figure. On the basis of these figures, we conclude that the key system GMM point estimate in Hausmann, Hwang, and Rodrik (2007) is robust to the weak instruments problem identified in Table 8. This is reassuring

given that we could strongly reject underidentification of the levels equation using the Kleibergen-Paap LM test.

In Figure 3, panel B, we plot two-dimensional weak-instrument robust confidence sets for subsets of the four endogenous variables in Hauk and Wacziarg (2009): physical capital, human capital, population, and lagged income. We cannot reject that log human capital and log physical capital have null effects on economic growth in both the 2SLS difference and levels equations. The bottom right graph reaffirms the null result for human capital when taking a different two-dimensional representation of the four-dimensional confidence ellipsoid. While the levels and difference equations are not underidentified (see Table 8), weak instruments may have implications for inference in the augmented Solow model.

IV. Lessons

We demonstrate that invalid and weak instruments continue to be commonly used in the growth literature. This suggests that the warnings of Durlauf, Johnson and Temple (2005) and others on this subject have gone unheard. Weak and/or invalid instruments do not assist researchers in conducting meaningful hypothesis tests about the causes of growth. Continued use of problematic instruments in the growth literature risks pushing all of its findings further towards irrelevance.

Many of the papers discussed here contain explicit policy implications based on their results. Without strong and valid identification of causal relationships, such exercises may or may not carry policy implications, and require further investigation. Nevertheless, these studies remain valuable contributions to the literature for other reasons—especially their innovations in method.

We certainly do not recommend that economists refrain from pursuing pressing research questions until perfect methods arrive. But we suggest a handful of guidelines for the next generation of growth empirics:

- *Generalize the theoretical underpinnings of an instrument to account for other published results with the same instrument.* When an instrument has been used elsewhere in the literature, new users of that instrument bear the burden of showing that other important findings using that instrument do not invalidate its use in the new case. This can be done using a somewhat more generalized model that comprises causal pathways explored elsewhere with that instrument. Accounting for all plausible pathways through a “unified growth theory” is too high a standard, but accounting for prominent published pathways should be a minimum standard.
- *Deploy the latest tools for probing validity.* Perfect instruments for growth determinants will remain elusive, but many underutilized tools exist to shine brighter light on the instruments we have. The Hahn, Ham, and Moon (2011) test used in this paper probes the validity of strong instruments in the presence of other weak ones. Imbens (2003) lays out a transparent method of assessing the sensitivity of a growth effect estimate to a given degree of correlation between instrument and error. Kraay (2008) and Conley, Hansen, and Rossi (2012) explore how

to conduct second-stage inference accounting for prior uncertainty about the excludability of the instrument. Ashley (2009) shows how the discrepancy between OLS and IV estimates can be used to estimate the degree of bias under any given assumption about the degree to which the exclusion restrictions are violated.

- *Open the black box of GMM.* It is no longer sufficient to assert that the mere use of system GMM adequately addresses the risk of weak instrumentation in dynamic panel models. As applied econometricians wait for an analog of the Stock and Yogo (2005) weak instrument diagnostics suitable for dynamic panel GMM estimation, its use must be complemented by supportive evidence that the instruments explain a sufficient degree of the variance of the endogenous regressors (and not simply because so many instruments are used). Papers exploring growth determinants should explore the strength of candidate instruments in analogous two-stage least squares regressions, should explore robustness to collapsing of the instrument matrix, should utilize optimal instrument selection procedures tailored to dynamic panel GMM (Okui 2009), and should explore methods robust to weak instruments (Kleibergen 2002; Kleibergen and Mavroeidis 2009). Robust inference procedures now provide growth researchers with the means to go beyond merely identifying weak instruments to characterizing their implications for inference about key structural parameters.

APPENDIX A: BACKGROUND ON THE INSTRUMENT

Rajan and Subramanian (2008) construct an instrumental variable for the aid receipts in a “zero-stage” specification by regressing bilateral aid flows as a fraction of recipient GDP on recipient and donor characteristics. They use the resulting coefficients to calculate predicted bilateral aid flows. They sum these predicted bilateral flows across donors to arrive at predicted total aid receipts for each recipient country as a fraction of recipient GDP. This predicted total, a constructed instrument for true aid receipts, becomes the excluded instrument in a series of two-stage least squares regressions of economic growth on aid receipts and a set of control variables. The instrument is $a_{dr} \equiv \frac{A_{dr}}{Y_r} = \sum_{j=0}^7 \beta_j I_{j,dr} + \sum_{j=0}^5 \beta_{j+8} (\ln P_d - \ln P_r) I_{j,dr} + v_{dr}$, where A_{dr} is dollars of aid given by donor d to recipient r ; Y_r is the GDP of r ; β_0 through β_{13} are regression coefficients; P_d is donor-country population; and P_r is recipient-country population. The I s are a set of time-invariant country dummy variables describing the country dyad: a current or past colonial relationship (I_1); a current or past colonial relationship with the United Kingdom (I_2), France (I_3), Spain (I_4), or Portugal (I_5); common language (I_6); and a current colonial relationship (I_7). Finally, $I_{0,dr} = 1 \ \forall \ d, r$ and v_{dr} is an error term. The estimated coefficient vector $\hat{\beta}$ is then used to generate predicted bilateral flows \bar{a}_{dr} , which are summed across donors to create the constructed instrument $\bar{a}_r = \sum_d \bar{a}_{dr}$, which then instruments for aid receipts $a_r \equiv A_r/Y_r$ in the cross-section growth regression $g_r = \gamma_1 a_r + \mathbf{X}_r' \Theta + u_r$, where g_r is real GDP per capita growth, \mathbf{X}_r is a vector of country characteristics, γ_1 is a regression coefficient, Θ is a vector of regression coefficients, and u_r is an error term.

REFERENCES

- Acemoglu, Daron.** 2010. "Theory, General Equilibrium, and Political Economy in Development Economics." *Journal of Economic Perspectives* 24 (3): 17–32.
- Acemoglu, Daron, Simon Johnson, and James A. Robinson.** 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91 (5): 1369–1401.
- Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared.** 2005. "From Education to Democracy?" *American Economic Review* 95 (2): 44–49.
- Albouy, David Y.** Forthcoming. "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment." *American Economic Review*.
- Alesina, Alberto, and Romain Wacziarg.** 1998. "Openness, Country Size and Government." *Journal of Public Economics* 69 (3): 305–21.
- Alfaro, Laura, Areendam Chanda, Sebnem Kalemli-Ozcan, and Selin Sayek.** 2004. "FDI and Economic Growth: The Role of Local Financial Markets." *Journal of International Economics* 64 (1): 89–112.
- Anderson, Theodore W., and Cheng Hsiao.** 1982. "Formulation and Estimation of Dynamic Models Using Panel Data." *Journal of Econometrics* 18 (1): 47–82.
- Anderson, Theodore W., and Herman Rubin.** 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63.
- Angeles, Luis, and Kyriakos C. Neanidis.** 2009. "Aid Effectiveness: The Role of the Local Elite." *Journal of Development Economics* 90 (1): 120–34.
- 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): 277–97.
- Arellano, Manuel, and Olympia Bover.** 1995. "Another Look at the Instrumental Variable Estimation of Error-Components Models." *Journal of Econometrics* 68 (1): 29–51.
- Ashley, Richard.** 2009. "Assessing the Credibility of Instrumental Variables Inference with Imperfect Instruments via Sensitivity Analysis." *Journal of Applied Econometrics* 24 (2): 325–37.
- Baltagi, Badi H., and Young-Jae Chang.** 1994. "Incomplete Panels: A Comparative Study of Alternative Estimators for the Unbalanced One-Way Error Component Regression Model." *Journal of Econometrics* 62 (2): 67–89.
- Barro, Robert J.** 1991. "Economic Growth in a Cross Section of Countries." *Quarterly Journal of Economics* 106 (2): 407–43.
- Barro, Robert J.** 1996. "Democracy and Growth." *Journal of Economic Growth* 1 (1): 1–27.
- Baum, Christopher F., Mark E. Schaffer, and Stephen Stillman.** 2007. "Enhanced Routines for Instrumental Variables/GMM Estimation and Testing." *Stata Journal* 7 (4): 465–506.
- Baumol, William J.** 1986. "Productivity Growth, Convergence, and Welfare: What the Long-Run Data Show." *American Economic Review* 76 (5): 1072–85.
- Bazzi, Samuel, and Michael A. Clemens.** 2013. "Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth: Dataset." *American Economic Journal: Macroeconomics*. <http://dx.doi.org/10.1257/mic.5.2.152>.
- Beck, Thorsten, Asli Demirgüç-Kunt, and Ross Levine.** 2005. "SMEs, Growth, and Poverty: Cross-Country Evidence." *Journal of Economic Growth* 10 (3): 199–229.
- Beck, Thorsten, and Ross Levine.** 2004. "Stock Markets, Banks, and Growth: Panel Evidence." *Journal of Banking and Finance* 28 (3): 423–42.
- Blundell, Richard W., and Stephen R. Bond.** 1998. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87 (1): 115–43.
- Blundell, Richard W., and Stephen R. Bond.** 2000. "GMM Estimation with Persistent Panel Data: An Application to Production Functions." *Econometric Reviews* 19 (3): 321–40.
- Blundell, Richard W., Stephen R. Bond, and Frank Windmeijer.** 2000. "Estimation in Dynamic Panel Data Models: Improving on the Performance of the Standard GMM Estimator." *Advances in Econometrics* 15 (1): 53–92.
- Bobba, Matteo, and Decio Coviello.** 2007. "Weak Instruments and Weak Identification, in Estimating the Effects of Education, on Democracy." *Economics Letters* 96 (3): 301–06.
- Bond, Stephen R., Anke Hoeffler, and Jonathan Temple.** 2001. "GMM Estimation of Empirical Growth Models." Center for Economic Policy Research (CEPR) Discussion Paper 3048.
- Boone, Peter.** 1996. "Politics and the Effectiveness of Foreign Aid." *European Economic Review* 40 (2): 289–329.
- Borensztein, Eduardo, Jose De Gregorio, and Jong-Wha Lee.** 1998. "How Does Foreign Direct Investment Affect Economic Growth?" *Journal of International Economics* 45 (1): 115–35.
- Bosworth, Barry P., and Susan M. Collins.** 2003. "The Empirics of Growth: An Update." *Brookings Papers on Economic Activity* 33 (2): 113–79.

- Bowsher, Clive G.** 2002. "On Testing Overidentifying Restrictions in Dynamic Panel Data Models." *Economics Letters* 77 (2): 211–20.
- Brock, William A., and Steven N. Durlauf.** 2001. "Growth Empirics and Reality." *World Bank Economic Review* 15 (2): 229–72.
- Bun, Maurice J. G., and Monique de Haan.** 2010. "Weak Instruments and the First Stage F Statistic in IV Models with a Nonscalar Error Covariance Structure." University of Amsterdam Working Paper 2.
- Bun, Maurice J. G., and Frank Windmeijer.** 2010. "The Weak Instrument Problem of the System GMM Estimator in Dynamic Panel Data Models." *Econometrics Journal* 13 (1): 95–126.
- Burnside, Craig, and David Dollar.** 2000. "Aid, Policies, and Growth." *American Economic Review* 90 (4): 847–68.
- Burnside, Craig, and David Dollar.** 2000. "Aid, Policies, and Growth: Dataset." *American Economic Review*. <http://www.cgdev.org/content/publications/detail/11941> (accessed August, 2008).
- Calderón, César, Alberto Chong, and Norman Loayza.** 2002. "Determinants of Current Account Deficits in Developing Countries." *B. E. Journal of Economics: Contributions to Macroeconomics* 2 (1): Article 2.
- Caselli, Francesco, Gerardo Esquivel, and Fernando Lefort.** 1996. "Reopening the Convergence Debate: A New Look at Cross-Country Growth Empirics." *Journal of Economic Growth* 1 (3): 363–89.
- Clemens, Michael A., Steven Radelet, Rikhil R. Bhavnani, and Samuel Bazzi.** 2012. "Counting Chickens When They Hatch: Timing and the Effects of Aid on Growth." *Economic Journal* 122 (561): 590–617.
- Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi.** 2012. "Plausibly Exogenous." *Review of Economics and Statistics* 94 (1): 260–72.
- Cragg, John G., and Stephen G. Donald.** 1993. "Testing Identifiability and Specification in Instrumental Variable Models." *Econometric Theory* 9 (2): 222–40.
- Deaton, Angus S.** 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.
- DeJong, David N., and Marla Ripoll.** 2006. "Tariffs and Growth: An Empirical Exploration of Contingent Relationships." *Review of Economics and Statistics* 88 (4): 625–40.
- DeJong, David N., and Marla Ripoll.** 2006. "Tariffs and Growth: An Empirical Exploration of Contingent Relationships: Dataset." *Review of Economics and Statistics*. Data provided by the authors (accessed August 2008).
- Djankov, Simeon, Rafael la Porta, Florencio López-de Silanes, and Andrei Shleifer.** 2003. "Courts." *Quarterly Journal of Economics* 118 (2): 453–517.
- Djankov, Simeon, Jose G. Montalvo, and Marta Reynal-Querol.** 2008. "The Curse of Aid." *Journal of Economic Growth* 13 (3): 169–94.
- Doko, Firmin, and Jean-Marie Dufour.** 2008. "Instrument Endogeneity and Identification-Robust Tests: Some Analytical Results." *Journal of Statistical Planning and Inference* 138 (9): 2649–61.
- Dollar, David, and Aart Kraay.** 2003. "Institutions, Trade, and Growth." *Journal of Monetary Economics* 50 (1): 133–62.
- Durlauf, Steven N., Paul A. Johnson, and Jonathan R. W. Temple.** 2005. "Growth Econometrics." In *Handbook of Economic Growth*, Vol. 1A, edited by Philippe Aghion and Steven N. Durlauf, 555–677. Amsterdam: North-Holland.
- Easterly, William.** 2009. "Can the West Save Africa?" *Journal of Economic Literature* 47 (2): 373–447.
- Easterly, William, Ross Levine, and David Roodman.** 2004. "Aid, Policies, and Growth: Comment." *American Economic Review* 94 (3): 774–80.
- Fernández, Carmen, Eduardo Ley, and Mark F. J. Steel.** 2001. "Model Uncertainty in Cross-Country Growth Regressions." *Journal of Applied Econometrics* 16 (5): 563–76.
- Feyrer, James, and Bruce Sacerdote.** 2009. "Colonialism and Modern Income: Islands as Natural Experiments." *Review of Economics and Statistics* 91 (2): 245–62.
- Forbes, Kristin J.** 2000. "A Reassessment of the Relationship between Inequality and Growth." *American Economic Review* 90 (4): 869–87.
- Frankel, Jeffrey A., and David Romer.** 1999. "Does Trade Cause Growth?" *American Economic Review* 89 (3): 379–99.
- Frankel, Jeffery A., and Andrew Rose.** 2002. "An Estimate of the Effect of Common Currencies on Trade and Income." *Quarterly Journal of Economics* 117 (2): 437–66.
- Friedman, Eric, Simon Johnson, Daniel Kaufmann, and Pablo Zoido-Lobaton.** 2000. "Dodging the Grabbing Hand: The Determinants of Unofficial Activity in 69 Countries." *Journal of Public Economics* 76 (3): 459–93.
- Giavazzi, Francesco, and Guido Tabellini.** 2005. "Economic and Political Liberalizations." *Journal of Monetary Economics* 52 (2): 1297–1330.

- Glaeser, Edward L., Rafael La Porta, Florencio López-de Silanes, and Andrei Shleifer. 2004. "Do Institutions Cause Growth?" *Journal of Economic Growth* 9 (3): 271–303.
- Guggenberger, Patrik. 2009. "The Impact of a Hausman Pretest on the Asymptotic Size of a Hypothesis Test." *Econometric Theory* 26 (2): 369–82.
- Hahn, Jinyong, John C. Ham, and Hyungsik Roger Moon. 2011. "The Hausman Test and Weak Instruments." *Journal of Econometrics* 160 (2): 289–99.
- Hahn, Jinyong, and Jerry Hausman. 2005. "Estimation with Valid and Invalid Instruments." *Annales d'Économie et de Statistique* 79/80 (1): 25–57.
- Hall, Alastair, Atsushi Inoue, and Changmook Shin. 2008. "Entropy-Based Moment Selection in the Presence of Weak Identification." *Econometric Reviews* 27 (5): 398–427.
- Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" *Quarterly Journal of Economics* 114 (1): 83–116.
- Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (4): 1029–54.
- Hauk, William R., Jr., and Romain Wacziarg. 2009. "A Monte Carlo Study of Growth Regressions." *Journal of Economic Growth* 14 (2): 103–47.
- Hauk, William R., Jr., and Romain Wacziarg. 2009. "A Monte Carlo Study of Growth Regressions: Dataset." *Journal of Economic Growth*. Data provided by the authors (accessed August 2008).
- Hausmann, Ricardo, Jason Hwang, and Dani Rodrik. 2007. "What You Export Matters." *Journal of Economic Growth* 12 (1): 1–25.
- Hausmann, Ricardo, Jason Hwang, and Dani Rodrik. 2007. "What You Export Matters: Dataset." *Journal of Economic Growth*. Data provided by the authors (accessed August 2008).
- Hayakawa, Kazuhiko. 2009. "A Simple Efficient Instrumental Variable Estimator for Panel AR(p) Models When Both N and T Are Large." *Econometric Theory* 25 (3): 873–90.
- Hayashi, Fumio. 2000. *Econometrics*. Princeton: Princeton University Press.
- Holtz-Eakin, Douglas, Whitney Newey, and Harvey S. Rosen. 1988. "Estimating Vector Autoregressions with Panel Data." *Econometrica* 56 (6): 1371–95.
- Imbens, Guido W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." *American Economic Review* 93 (2): 126–32.
- Jones, Benjamin F., and Benjamin A. Olken. 2005. "Do Leaders Matter? National Leadership and Growth since World War II." *Quarterly Journal of Economics* 120 (3): 835–64.
- Kleibergen, Frank. 2002. "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression." *Econometrica* 70 (5): 1781–1803.
- Kleibergen, Frank, and Sophocles Mavroeidis. 2009. "Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve." *Journal of Business and Economic Statistics* 27 (2): 293–311.
- Kleibergen, Frank, and Richard Paap. 2006. "Generalized Reduced Rank Tests Using the Singular Value Decomposition." *Journal of Econometrics* 133 (1): 97–126.
- Kraay, Aart. 2008. "Instrumental Variables Regression with Honestly Uncertain Exclusion Restrictions." World Bank Policy Research Working Paper 4632.
- Levine, Ross, Norman Loayza, and Thorsten Beck. 2000. "Financial Intermediation and Growth: Causality and Causes." *Journal of Monetary Economics* 46 (1): 31–77.
- Levine, Ross, Norman Loayza, and Thorsten Beck. 2000. "Financial Intermediation and Growth: Causality and Causes: Dataset." *Journal of Monetary Economics*. <http://go.worldbank.org/40TPPEYOC0> (accessed August 2008).
- Lundberg, Mattias, and Lyn Squire. 2003. "The Simultaneous Evolution of Growth and Inequality." *Economic Journal* 113 (487): 326–44.
- Mankiw, N. Gregory. 1995. "The Growth of Nations." *Brookings Papers on Economic Activity* 25 (1): 275–310.
- Mauro, Paolo. 1995. "Corruption and Growth." *Quarterly Journal of Economics* 110 (3): 681–712.
- Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71 (4): 1027–48.
- Murray, Michael P. 2006. "Avoiding Invalid Instruments and Coping with Weak Instruments." *Journal of Economic Perspectives* 20 (4): 111–32.
- Newey, Whitney K., and Frank Windmeijer. 2009. "Generalized Method of Moments with Many Weak Moment Conditions." *Econometrica* 77 (3): 687–719.
- Noguer, Marta, and Marc Siscart. 2005. "Trade Raises Income: A Precise and More Robust Result." *Journal of International Economics* 65 (2): 447–60.
- Okui, Ryo. 2009. "The Optimal Choice of Moments in Dynamic Panel Data Models." *Journal of Econometrics* 151 (1): 1–16.

- Papaioannou, Elias, and Gregorios Siourounis.** 2008. "Democratisation and Growth." *Economic Journal* 118 (532): 1520–51.
- Persson, Torsten, and Guido Tabellini.** 2006. "Democracy and Economic Development: The Devil Is in the Details." *American Economic Review* 96 (2): 319–24.
- Persson, Torsten, and Guido Tabellini.** 2007. "The Growth Effect of Democracy: Is It Heterogenous and How Can It Be Estimated?" National Bureau of Economic Research (NBER) Working Paper 13150.
- Rajan, Raghuram G., and Arvind Subramanian.** 2008. "Aid and Growth: What Does the Cross-Country Evidence Really Show?" *Review of Economics and Statistics* 90 (4): 643–65.
- Rajan, Raghuram G., and Arvind Subramanian.** 2008. "Aid and Growth: What Does the Cross-Country Evidence Really Show?: Dataset." *Review of Economics and Statistics*. Data provided by the authors (accessed January 2008).
- Rodríguez, Francisco, and Dani Rodrik.** 2001. "Trade Policy and Economic Growth: A Skeptic's Guide to the Cross-National Evidence." In *NBER Macroeconomics Annual 2000*, Vol. 15, edited by Ben S. Bernanke and Kenneth Rogoff, 261–325. Cambridge, MA: MIT Press.
- Rodrik, Dani, and Romain Wacziarg.** 2005. "Do Democratic Transitions Produce Bad Economic Outcomes?" *American Economic Review* 95 (2): 50–55.
- Roodman, David.** 2009a. "How to Do xtabond2: An Introduction To Difference and System GMM in Stata." *Stata Journal* 9 (1): 86–136.
- Roodman, David.** 2009b. "A Note on the Theme of Too Many Instruments." *Oxford Bulletin of Economics and Statistics* 71 (1): 135–58.
- Rose, Andrew K.** 2006. "Size Really Doesn't Matter: In Search of a National Scale Effect." *Journal of the Japanese and International Economies* 20 (4): 482–507.
- Sala-I-Martin, Xavier.** 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87 (2): 178–83.
- Sala-I-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller.** 2004. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." *American Economic Review* 94 (4): 813–35.
- Spolaore, Enrico, and Romain Wacziarg.** 2005. "Borders and Growth." *Journal of Economic Growth* 10 (4): 331–86.
- Stock, James H., and Jonathan H. Wright.** 2000. "GMM with Weak Identification." *Econometrica* 68 (5): 1055–96.
- Stock, James H., and Motohiro Yogo.** 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by Donald W. K. Andrews and James H. Stock, 80–108. New York: Cambridge University Press.
- Tavares, Jose, and Romain Wacziarg.** 2001. "How Democracy Affects Growth." *European Economic Review* 45 (8): 1341–78.
- Temple, Jonathan.** 1999. "The New Growth Evidence." *Journal of Economic Literature* 37 (1): 112–56.
- Voitchovsky, Sarah.** 2005. "Does the Profile of Income Inequality Matter for Economic Growth?: Distinguishing between the Effects of Inequality in Different Parts of the Income Distribution." *Journal of Economic Growth* 10 (3): 273–96.
- Voitchovsky, Sarah.** 2005. "Does the Profile of Income Inequality Matter for Economic Growth?: Dataset." *Journal of Economic Growth*. Data provided by the authors (accessed August 2008).
- Windmeijer, Frank.** 2005. "A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators." *Journal of Econometrics* 126 (1): 25–51.
- Yogo, Motohiro.** 2004. "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak." *Review of Economics and Statistics* 86 (3): 797–810.