Working Paper Number 134
December 2007

Macro Aid Effectiveness Research: A Guide for the Perplexed

By David Roodman

Abstract

Like many public policy debates, that over whether foreign aid works carries on in two worlds. Within the research world, it plays out in the form of papers full of technical language, formulas, and numbers. Outside, the arguments are plainer and the audience broader, but those academic studies remain a touchstone. While avoiding jargon, this paper reviews recent, contending studies of how much foreign aid affects country-level outcomes such as economic growth and school attendance rates. This particular kind of study is ambitious: it is far easier to evaluate a school-building project, say, on whether the school was built and children filled its seats than to determine whether all aid, or large subcomponents of it, made the economy grow faster. Because of its ambition, this literature has attracted attention from those hoping for clear answers on whether aid "works." On balance, the quantitative approach to exploring grand questions about aid effectiveness, which began 40 years ago, was worth trying and is probably worth pursuing somewhat further. But the literature will probably continue to disappoint as often as it offers hope. Perhaps the biggest challenge is going beyond documenting correlations to demonstrating causation—not just that aid went hand-in-hand with economic growth, but caused it. Aid has eradicated diseases, prevented famines, and done many other good things. But given the limited and noisy data available, its effects on growth in particular probably cannot be detected.

Macro Aid Effectiveness Research:

A Guide for the Perplexed

David Roodman[1]

Center for Global Development

[1] David Roodman is a research fellow at the Center for Global Development. He thanks Ethan Kapstein, Ruth Levine, Steven Radelet, Dani Rodrik, and Arvind Subramanian for comments.

Introduction

The old argument about whether foreign aid "works" continues to percolate in the popular press. Between the stalwarts of Jeffrey Sachs and William Easterly, Paul Collier has prominently sought a middle ground while Arvind Subramanian has taken a position that is at least as pessimistic as Easterly's. What should people who care about making aid work, but who cannot understand the technicalities of the arguments, conclude from all the contention?

This note offers some answers on the question of what we know about aid effectiveness—but they are not all pat answers. The paper works on two tracks: 1) describing consensus where it exists (it is often overshadowed by controversy) and 2) stating as fact what the sources of controversy are, where they exist.

Like many public policy debates, that over whether foreign aid brings overall progress in receiving countries carries on in two worlds. Within the research world, it plays out in the form of papers full of technical language, formulas, and numbers. Outside, the arguments are plainer and the audience broader, but those academic studies remain a touchstone. Policymakers and concerned citizens have looked to the researchers for insight on two questions. First, does aid work on average? If the researchers think so, that buttresses the case for broad aid increases, such as urged by the former U.K. Prime Minister Tony Blair's Commission for Africa in 2005. If not, that reinforces skepticism. Second, regardless of the overall averages, can aid be expected to work better in certain kinds of countries—ones that have democratic governments, perhaps, or that are particularly vulnerable to global economic shocks? Evidence of this kind offers officials not only a rationale for giving aid, but a recipe for giving it.

The Center for Global Development has indeed been a center of the analytical work. In 2002, the then-resident Easterly, along with Ross Levine and CGD's David Roodman (ELR), coauthored a CGD working paper that challenged earlier work by World Bank economists Craig Burnside and David Dollar.[1] Burnside and Dollar had famously found that aid works in countries with good economic policies. ELR showed that these results were fragile, being sensitive to small changes in the data set. Roodman went on to widen the critique to other studies. Then, in a 2004 working paper, CGD's Michael Clemens, Steven Radelet, and Rikhil Bhavnani made the case that the aid that could plausibly raise growth within a few years in fact does.[2] Flatly contradicting them in a 2005 were two IMF economists, Raghuram Rajan and Arvind Subramanian; Subramanian is now a CGD-Peterson Institute joint fellow.[3] He recently penned a *Wall Street Journal* op-ed broadcasting his conclusion that aid has no significant effect on growth and may in fact corrode recipient countries' governance and competitiveness.[4]

Bluntly, a question that is probably on the minds of many who follow these matters is, *How can smart people draw such contradictory conclusions from the same data?* Where lies the truth?

---

[1] William Easterly, Ross Levine, and David Roodman, "Aid, Policies, and Growth: Comment," *American Economic Review* 94(3), pp. 774–80, June 2004, previously published as CGD Working Paper 26, February 2003.
[2] Michael Clemens, Steven Radelet, and Rikhil Bhavnani, "Counting Chickens When They Hatch: The Short-Term Effect of Aid on Growth," Working Paper 44, Center for Global Development, Washington, DC, July 2004.
[3] Raghuram G. Rajan and Arvind Subramanian, "Aid and Growth: What Does the Cross-Country Evidence Really Show?" *Review of Economics and Statistics*, forthcoming.
[4] Arvind Subramanian, "A Farewell to Alms," *Wall Street Journal*, August 22, 2007.

The effects of any given aid project ripple through a society in complex ways. Aid *effectiveness* can therefore defined at many levels. Was a school built? Did children come? Did they learn? When they grew up, did they have fewer children of their own? Did they find more rewarding and productive work? Did economic output go up? Did poverty or inequality fall? As they should, aid agency evaluators have assessed thousands of projects on the more proximate outcomes. They have documented many failures, but also some monumental successes, such as the Green Revolution, which ended a long history of famines in India, and the eradication of smallpox. (Though, in general, Ruth Levine and other analysts have argued that evaluators have been far less rigorous than they should be.[5])

This paper focuses on the literature about effectiveness with respect to the more diffuse and distant goals, especially economic growth. This "macro" literature has attracted particular attention because it promises, seemingly, to answer grand questions about aid policy, and because it has been contentious. This is not meant to imply that studies of effectiveness with respect to more immediate goals are less useful. If anything, this review argues the contrary.

1 Perspective

Aid has often worked. In many of the successes, improvements in outcomes such as vaccination rates were so dramatic—and their importance to people's welfare so obvious—that statistical analysis has easily confirmed their value. The World Health Organization, for example, led the successful campaign to rid the world of smallpox and the Pan American Health Organization did the same for eliminating polio from the Western Hemisphere. Donors also played an important supporting role in the Green Revolution.[6]

Twentieth-century advances in the manipulation of nature made both of these successes possible. Where science is less central, as in education, it is harder to find such spectacular achievements. The late sociologist Peter Rossi noted that "[i]n the social program field, nothing has yet been invented which is as effective in its way as the small pox vaccine was for the field of public health."[7] Nevertheless, rigorous evaluations have demonstrated the effectiveness of specific social interventions. A growing body of evidence from random assignment evaluations of conditional cash transfers (in Brazil, Mexico, Nicaragua, and elsewhere) shows improvements in

[5] The Evaluation Gap Working Group, *When Will We Ever Learn? Improving Lives through Impact Evaluation* (Washington, DC: Center for Global Development, 2006).
[6] Ruth Levine and the What Works Working Group, *Millions Saved: Proven Successes in Global Health* (Washington, DC: Center for Global Development, 2004).
[7] Peter H. Rossi, "The Iron Law of Evaluation and Other Metallic Rules," *Research in Social Problems and Public Policy* 4, pp. 3–20 (1987).

school attendance and some health indicators.[8] Several well-evaluated secondary school scholarship programs for girls find major improvements in retention and completion.[9]

On the other hand, there is also consensus that aid has often *not* worked, and even done harm. Billions of dollars in grants and loans to the rapacious Mobutu Sese Seko left ordinary Zairians no better off, to say the least. U.S. aid to the Marcos regime in the 1980s may have delayed democracy in the Philippines.

Perhaps more relevant to understanding the challenge of measuring aid effectiveness are the prosaic, small-scale failures that pepper the aid landscape and mix in statistically with the successes. After all, helping to improve a society from the outside is no easy task. In its most recent review of World Bank–financed projects, the Bank's Independent Evaluation Group (IEG) found that 22.2% of projects in Africa completed in 2001–05 had unsatisfactory outcomes, and 20.1% were probably unsustainable, meaning that they did not bring lasting benefits.[10] The success rates reported by the IEG should be taken with grains of salt since they are based in large measure on Bank staff self-reports and have recently improved across the board. But the point stands that when success and failure mix together so closely, aid effectiveness becomes harder to demonstrate in the large.

This mix of successes and failures fits the venture capitalist's perspective: every aid effort is a calculated risk, or at least an experiment. Success for the aid enterprise should manifest as a few spectacular successes, a lot of smaller successes, and lot of disappointments. Unfortunately, where the balance lies is unclear because the aid donor's bottom line is far harder to measure than the capitalist's. And those spectacular life-saving successes such as the Green Revolution may not show up in macro studies of aid effectiveness since more "capitas" does not necessarily mean more gross domestic product (GDP) per capita.

Another important point of consensus—not to mention common sense—is that as the consequences of an aid project ripple out and diffuse, they become harder to detect. It is not hard to tell if a road was paved. But it may be impossible to discern whether *all aid* raises *total output per capita* from a national economy in the *long run*. Human societies are complex and subject to many influences. Ethical considerations may stop governments from carrying out macro-level experiments that would help us tease apart cause and effect by randomly granting aid to some countries and withholding it from others. For that would institutionalize a policy of treating

---

[8] Eliana Cardoso and Andre Portela Souza, "The Impact of Cash Transfers on Chile Labor and School Attendance in Brazil," Paper No. 04-W07, Department of Economics, Vanderbilt University, Nashville, TN, 2004; Saul Morris, Pedro Olinto, et al. "Conditional Cash Transfers are Associated with a Small Reduction in the Rate of Weight Gain of Preschool Children in Northeast Brazil," American Society for Nutritional Science, Bethesda, MD, 2004.; Orazio Attanasio, Costas Meghir, and Ana Santiago, "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to evaluate Progresa," manuscript, Institute for Fiscal Studies, London, 2005; John Maluccio and Rafael Flores, "Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan Red de Proteccion Social," Discussion Paper 184, IFPRI, Washington, DC, 2004.

[9] Deon Filmer and Norbert Schady, "Getting Girls into School: Evidence from a Scholarship Program in Cambodia," Policy Research Working Paper 3910, World Bank, Washington, DC, 2006; Shahidur Khandker, Mark Pitt, and Nobuhiko Fuwa, "Subsidy to Promote Girls' Secondary Education: The Female Stipend Program in Bangladesh," World Bank, Brown University, and Chiba University, 2003.

[10] World Bank, Independent Evaluation Group, *Annual Review of Development Effectiveness* (Washington, DC: 2006), Table A.2.

equally needy people differently. So somewhere between the extremes of short-term project effectiveness and long-term impacts on economic growth lies a fuzzy boundary, the limit to what we can know about aid effectiveness.

The high-profile controversy in the research world is over where this limit lies. The optimists believe that whether aid in general, or at least some major subcategory of aid, has improved society-level outcomes such as economic growth is both knowable and known. The skeptics are unconvinced on both counts. Some of them doubt the effectiveness of aid in general. Others are open to the possibility that it does more good than harm, but doubt the ability of econometrics to detect what is probably a weak signal within the noisy and limited data available from developing countries. They doubt aid regressions (regressions being economists' favorite statistical tool) more than aid itself, a skepticism that may be rooted in the intellectual revolution triggered by the twentieth century's encounter with hard limits to human knowledge. Werner Heisenberg discovered that an observer cannot simultaneously measure the position and velocity of a particle with perfect accuracy. Kurt Gödel showed that there are true mathematical statements that are unprovable and false ones that are irrefutable.

The rest of this paper is not an academic literature review, but a guide to recent controversies for the perplexed non-expert. It briefly reviews the history of "macro" aid research—that on country-level outcomes such as economic growth and school enrollment rates—emphasizing recent developments. It explores some of the deep causes of controversy in the literature and confusion in the public mind. And it explains the essence of current debates as dispassionately as it can manage. On balance, it tends to side with the aid *regression* skeptics.

2 History

Scientific inquiry into the macro effects of aid began in the 1960s, and now includes more than 100 papers.[11] Amidst all the debate and findings, the literature has reviewed itself at least four times.[12] Since the research has progressed organically and unplanned it is not surprising that each review demarcates "phases" or "generations" in the literature differently. And the reviewers, partisans to the debates, have interpreted the history differently.

One thing that is clear is that the scholarship has steadily improved, and for two reasons. First, the quality and quantity of the underlying data have risen. The OECD's Development Assistance Committee, based in Paris, began collecting aid statistics in the 1960s. Its database has grown as more countries and agencies have entered the aid business, as the years of data have accumulated, and as detail has improved. Though the data are still far short of the ideal for many purposes, the growth in its quantity has given researchers more power to detect statistical relationships. Much the same can be said for data on GDP, infant mortality, and many other indicators.

[11] Hristos Doucouliagos and Martin Paldam, "The Aid Effectiveness Literature: The Sad Results of 40 Years of Research," Working Paper 2005–15, Department of Economics, University of Aarhus, Aarhus, Denmark, 2005.
[12] Paul Mosley, "Aid, Savings and Growth Revisited," *Oxford Bulletin of Economics and Statistics* 42(2), pp. 79–96 (1980); Henrik Hansen and Finn Tarp, "Aid Effectiveness Disputed," *Journal of International Development* 12(3), pp. 375–98 (April 2000); Clemens, Radelet, an Bhavnani, op. cit. note 2; Mark McGillivray, Simon Feeney, Nils Hermes, and Robert Lensink, "It Works; It Doesn't; It Can, But That Depends…: 50 Years of Controversy over the Macroeconomic Impact of Development Aid," Working Paper 2005/24, World Institute for Development Economics Research, Helsinki (August 2005).

Meanwhile, the arrival of the microcomputer has led to an explosion of innovation in econometrics, which is the application of statistics to social science questions. Calculations that would once have been theoretical curiosities became practical and useful. Computer power thus fueled repeated cycles of innovation within econometrics, mostly having to do with the creation of new "estimators," which are formulas or procedures for estimating such numbers as the effect of smoking on weight or that of aid on growth. A cycle of innovation begins when one econometrician criticizes an estimator for embodying dubious assumptions about the real world, such as that there is no reverse causation from economic growth to aid levels. Another responds with a more complicated estimator designed to circumvent the problem. The cycle repeats. As a result, the econometric techniques used to study aid effectiveness today are far more sophisticated than those deployed in the 1960s and 1970s. (However, this sophistication also creates a risk that researchers, not fully understanding the behavior of these estimators, will unwittingly misuse them, as discussed below in section 3.2.)

Has improved scholarship brought improved understanding? A peculiar fact about the reviews of this literature is that there is controversy over how much controversy there is. Henrik Hansen and Finn Tarp (HT) argue that the majority of studies since 1970 have found that aid does good and almost none have found that it does harm.[13] The telling of Clemens, Radelet, and Bhavnani (CRB) has more the air of continuing controversy.[14] In the interpretation of Mark McGillivray and coauthors, only since 1996, with the arrival of a new generation of sophisticated analyses, did the light of clarity shine on the field.[15]

All do agree that something important happened in the mid-1990s. In 1994, Peter Boone at the London School of Economics circulated a working paper that found no significant effect of aid on savings or growth.[16] CRB describe what happened next:

> Boone's is the paper that launched a thousand regressions. Many researchers have taken Boone's findings as confirmation of a "macro-micro paradox": that so many aid-funded projects report positive micro-level economic returns somehow undetectable at the macro-level. Here…the literature splits into two strands—one trying to explain the paradox and the other denying its existence.[17]

CRB label the researchers associated with the two strands the "apologists" and the "gainsayers." (Some inhabit both strands through different papers.) The apologists accept Boone's contention that aid does not work on *average* but seek conditions under which it is nevertheless effective—say, in countries with good governance, however defined, or high vulnerability to natural disasters. Foremost among the "apologists" have been Craig Burnside and David Dollar, who found in a 1997 World Bank working paper that aid works in a good policy environment.[18] For them, such a propitious environment was indicated by openness to trade, low inflation, and low government budget deficits. Meanwhile, the gainsayers argued that aid does in fact work on

[13] Hansen and Tarp, op. cit. note 12.
[14] Clemens, Radelet, and Bhavnani, op. cit. note 2.
[15] McGillivray *et al.*, op. cit. note 12.
[16] Boone, Peter, "The Impact of Foreign Aid on Savings and Growth," Centre for Economic Performance Working Paper No. 677, London School of Economics, 1994.
[17] Clemens, Radelet, and Bhavnani, op. cit. note 2, p. 7.
[18] Craig Burnside and David Dollar, "Aid, Policies, and Growth," *American Economic Review* 90(4), pp. 847–68 (September 2000).

average. Their strongest voice, at least until the arrival of CRB in 2004, was the Danish duo of Hansen and Tarp.[19]

The apologist school thrived at first. Alongside Burnside and Dollar, Paul Collier and Dollar found that aid works best in countries with stronger government institutions and economic policies.[20] Jakob Svensson wrote that it raises growth in democracies.[21] Collier and Jan Dehn said aid *surges* are helpful in countries experiencing sharp price drops for key commodity exports.[22] Patrick Guillaumont and Lisa Chauvet told a similar story, of aid cushioning the countries most exposed to global economic shocks and natural disasters.[23] Collier and Anke Hoeffler argued that aid works best in countries that are just emerging from civil war *and* that have good policies.[24] Carl-Johan Dalgaard, along with Hansen and Tarp, said that aid works outside the tropics but not in them.[25] And there were more.

But the apologists' credibility appears to have declined in academia, if not in policy circles. In 2001, Carl-Johan Dalgaard and Hansen demonstrated the fragility of the famous Burnside and Dollar paper to small changes in the data sample.[26] Easterly, Levine, and Roodman hammered home the point, adding a few additional countries and years of data. In "The Anarchy of Numbers," Roodman demonstrated fragility in, among others, the works of Collier and Dollar, Collier and Dehn, Collier and Hoeffler.[27] (Despite this doubt, the publication of Collier's thoughtful *Bottom Billion*, built on his own scholarship, may be reinforcing some of these ideas in policy circles.[28])

The gainsayers also suffered under Roodman's scrutiny: the results of HT disappeared when Roodman added more data. Thus at the end of 2003, the landscape of macroeconomic aid research featured a lot of wreckage. In the policy world, that a left a vacuum of guidance about whether and where aid is likely to work.

Within a year, Clemens, Radelet, and Bhavnani bid fair to fill the vacuum. Future reviews of the literature may declare that "Counting Chickens When They Hatch" was the first of a new generation of studies. Its chief contribution is to narrow the aid variable to parallel the outcome

---

[19] Henrik Hansen and Finn Tarp, "Aid and Growth Regressions," *Journal of Development Economics* 64(2), pp. 547–70 (2001).
[20] Paul Collier and David Dollar, "Development Effectiveness: What Have We Learnt?" *The Economic Journal* 114(496), pp. F244–71 (2004).
[21] Jakob Svensson, "Aid, Growth and Democracy," *Economics and Politics* 11(3), pp. 275–97 (1999).
[22] Paul Collier and Jan Dehn, "Aid, Shocks, and Growth," Working Paper 2688, World Bank, Washington, DC, October 2001.
[23] Patrick Guillaumont and Lisa Chauvet, "Aid and Performance: A Reassessment," *Journal of Development Studies* 37(6), pp. 66–92 (August 2001).
[24] Paul Collier and Anke Hoeffler, "Aid, Policy and Growth in Post-Conflict Societies," *European Economic Review* 48(5), pp. 1125–45 (2004).
[25] Carl-Johan Dalgaard, Henrik Hansen, and Finn Tarp, "On the Empirics of Foreign Aid and Growth," *The Economic Journal* 114(496), pp. F191–216, June 2004.
[26] Carl-Johan Dalgaard and Henrik Hansen, "On Aid, Growth and Good Policies," *Journal of Development Studies* 37(6), pp. 17–41 (2001).
[27] David Roodman, "The Anarchy of Numbers: Aid, Development, and Cross-country Empirics," *World Bank Economic Review* 21(2), pp. 255–77 (May 2007).
[28] Paul Collier, *The Bottom Billion: Why the Poorest Countries are Failing and What Can Be Done About It* (New York: Oxford University Press, 2007).

studied, namely economic growth over the half-decade following aid disbursal. Their "short-impact" aid variable includes budget and balance of payments support and aid for infrastructure and industry, all of which can reasonably be expected to affect economic growth within a few years. It leaves out humanitarian assistance and "long-impact" assistance such as for education and health, which may only show up in GDP after a generation, if at all. This adjustment, CRB argued, eliminates a serious incoherence in previous studies.

Meanwhile, however, the IMF's chief economist, Raghuram Rajan, and its head of macroeconomic studies, Arvind Subramanian, embarked on their own project to correct what they saw as a serious deficiency in published work. Their concern was not about the mismatch between the aid and growth variables but about the techniques deployed to remove reverse causation and other forms of "endogeneity"—more technically, about the choice of "instruments." As discussed in section 3.1, what distinguishes econometrics from statistics generally is its obsession with causality. For the social scientist, it is not enough to know that, say, microcredit borrowing goes hand-in-hand with business success. She wants to know whether microcredit indeed causes some of that success—or whether some third variable such as entrepreneurial ability is at the root of both the borrowing and the happy outcome (omitted variable bias), or whether entrepreneurial success leads lenders to provide more microcredit (reverse causation bias). Rajan and Subramanian (RS) believed that past attempts to purge the data of omitted variable causation and reverse causation were inadequate to buttress confident claims of causality. They argued that the choice of instruments should be guided by a well-developed theoretical story. Using instruments they submit are superior, RS found no reliable effect of aid on growth.

In the last few years, a new crop of papers has appeared that copy CRB in narrowing the aid variable, or in tracking an outcome more specific than economic growth, or both. The German economists Katharina Michaelowa and Anke Weber find evidence that aid for education increases primary school enrollment.[29] So do the trio of Axel Dreher, Peter Nunnenkamp, and Rainer Thiele.[30] Prachi Mishra and David Newhouse at the IMF reach the same conclusion for health aid and infant mortality.[31] The reining in of ambition—measuring aid effectiveness against narrower goals—is refreshing. But it remains to be seen if these studies will transcend the problems of their predecessors.

3 Reasons Smart People Draw Opposite Conclusions from the Same Data

This section reviews some deep causes of controversy within the research community about the macro effects of aid, as well as ways that the media filter and distort the debate. With that background established, the section after returns to critique some of the key studies just mentioned.

[29] Katharina Michaelowa and Anke Weber, "Aid Effectiveness Reconsidered: Panel Data Evidence for the Education Sector," Discussion Paper 264, Hamburg Institute of International Economics, revised 2006.
[30] Axel Dreher, Peter Nunnenkamp, and Rainer Thiele, "Does Aid for Education Educate Children? Evidence from Panel Data," draft, January 2007.
[31] Prachi Mishra and David Newhouse, "Health Aid and Infant Mortality," Working Paper 07/100, International Monetary Fund, Washington, DC, April 2007.

3.1 Correlation vs. Causation

The greatest source of controversy in the macro literature arises from the distinction between *correlation* and *causality*. The aid and aid regression skeptics may accept correlations—statistical associations between aid and development outcomes—but they doubt the optimists' claims to have shown that aid *affects* those outcomes.

It may seem strange that social scientists argue over how well they can perceive something like the influence of aid on national development. Astronomers debate how a star got the way it is, but not whether it is there. What makes social scientists different is that they are interested in causality in human behavior and human societies. Were the question of aid effectiveness simply a matter of measuring correlation—is economic growth higher or poverty rarer where aid is greater?—there would be far less controversy. But much of econometrics was developed in order to go beyond correlation to causality. It is supposed to help researchers decide which information in a data set can confirm or refute a causal story of interest, and which information should be thrown away as "endogenous." When the unit of analysis is the country, an extremely complex system in which, roughly speaking, everything affects everything else, the task for econometrics is daunting indeed.

The ongoing renaissance in econometrics has produced many clever techniques for reducing endogeneity and for assessing how well a particular technique is working. However, none of these approaches is perfect; indeed, it is a mathematical inevitability that none ever will be. Not only can endogeneity never be fully removed but we cannot ever be sure of how much we have removed in any particular instance.

Endogenous causation is widely understood as a serious possibility in the study of aid and growth. For example, economic growth can affect aid levels (reverse causation), thereby obscuring the effect we are interested in, of aid on growth: donors may flock to fast-growing countries as promising, or avoid them as less needy. Or, since aid levels are measured in most aid-growth studies as a share of recipient's GDP, aid receipts may fail to keep up with GDP, so that as GDP goes up, aid/GDP goes down. Then there is the possibility of an omitted-variable: good governance, say, might affect both variables of interest, attracting aid and raising growth. All these dynamics could create correlations between aid and growth quite apart from any effect of the first on the second. The same goes for other outcomes, such as infant mortality.

The principal technique economists have used to purge data of endogenous information, common since the mid-1990s, has been to "instrument" aid with variables that are thought to be correlated with the outcome of interest *only* through their relationship with aid. If the instruments does correlate with the outcome, then aid must have worked. Whether a country was once a French colony or is of geopolitical importance to the United States, for example, arguably affects how much aid it gets, but not how fast it grows, so indicators of geopolitical status may make good instruments.

Instrumentation can be seen as an imperfect substitute for performing controlled experiments. If a donor randomly aided some countries and not others, and then observed that the aided countries did better, that would be strong evidence of aid effectiveness, assuming a reasonably large sample of countries. The random decision about whether to aid a country would be just like an instrument, something that could only have affected development through aid. Absent

experiments, econometricians must derive their instruments from "quasi-experiments." France's heavy aid to its former colonies in western Africa owes to a historical accident that once gave France dominion over that part of the world. This is the basis for using former French colonial status as an instrument for aid.

But quasi-experiments are rarely as clean as real ones. Is it obvious that having been a French colony *only* affects economic growth through aid receipts? Au contraire, a widely cited paper concludes that the origin of country's legal system (England, France, Germany, or Scandinavia) also affects the development of its financial system, which affects growth.[32] In technical terms, the French colonial history instrument may not be "valid." Another potential problem is instrument weakness. Suppose it turns out that former French colonies hardly get more aid on average than other countries. Then there is hardly any quasi-experiment with which to study aid effectiveness.

A final pitfall of instrumentation is that there is such a thing as too many of them. Researchers are free to incorporate many instruments into their regressions—one for having been a French colony, another for having been a British colony, and so on. But for technical reasons, as the number of instruments creeps toward the number of data points in a study, a collection of instruments that are individually valid can be collectively invalid.

How then does one judge specific instruments and the papers that use them? One approach researchers take is technical. When theoretical econometricians are not developing new estimators, they spend most of their time working on "specification tests" which are mathematical procedures meant to detect phenomena such as invalid or weak instruments. As a rule, the tests are imperfect but helpful. The discussion of specific papers in section 4 invokes such tests.

Another approach to judging instruments is conceptual. Does a choice of instrument appear grounded in a plausible story about how the world works, a theory? In their aid-growth paper, Rajan and Subramanian exhort aid researchers to take more care in theorizing around their instruments. The iconic example is the 2001 paper of economists Daron Acemoglu, Simon Johnson, and James A. Robinson, who studied the effects of institutions on long-run economic performance. Their preferred measure of institutions is an index of the strength of private property rights. Through a discourse on global economic history since the rise of the European colonial powers, they motivate a creative instrument for institutions: the mortality rate among European settlers in various colonies between the seventeenth an nineteenth centuries. Colonies where European settlers lived better, they suggest, developed into European offshoots that place more checks on the power of government. Countries where European settlers died off in droves from malaria and other diseases—notably, tropical countries—tended to become little more than extractive outposts for the Europeans, which they exploited but did not truly settle; and this left a far different governmental legacy to the present.

A good theory can discipline econometric analysis and makes it more convincing than an *ad hoc* statistical game it might otherwise appear to be. Nevertheless, one should not make a fetish of

[32] Ross Levine, Norman Loayza, and Thorsten Beck, "Financial Intermediation and Growth: Causality and Causes," *Journal of Monetary Economics* 46, pp. 31–77 (2000).

theory. A good theory can lead to bad instruments, and good instruments can sometimes be found without much theory. In the end, unfortunately, there are no simple rules of thumb for judging instruments, thus for determining how well a study has buttressed its causal claims.

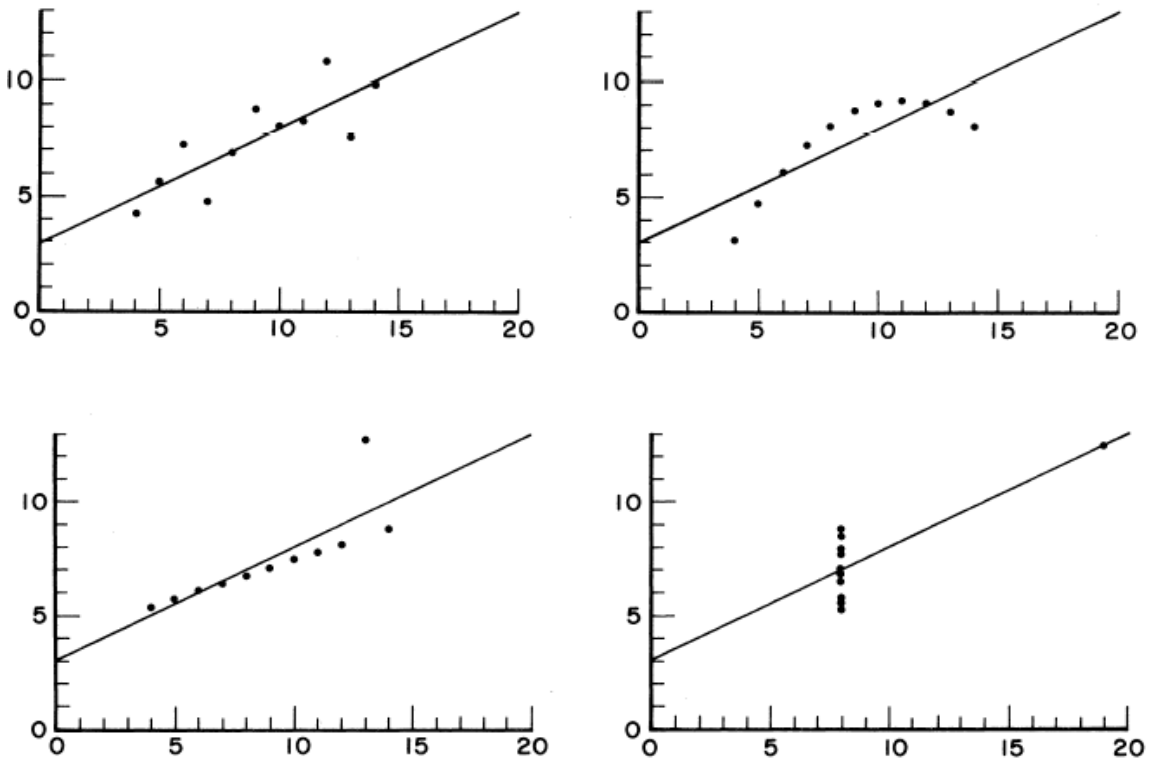
An alternative approach to studying causality, developed by Nobel laureate Clive Granger, should be noted.[33] It is rare in the aid literature.[34] Rather than trying to filter the information in the data set to remove endogeneity, Granger causality testing relies on time's arrow. If B tends to happen after A, there is a good chance B is caused by A. Causality usually flows forward through time. This suggests a simple way to check whether aid affects development: test whether information about past aid levels helps predict future growth rates. Granger testing has its own limitations because people *anticipate* events. If the stock market spikes just before the Federal Reserve cuts interest rates, that could be because traders see the cut coming, rather than the market move causing the cut. Nevertheless, like all imperfect techniques, Granger testing can offer qualified insight.

3.2 The Black Box Problem

The rapid progression of econometrics toward more sophisticated techniques has already been mentioned. Much of that progress in theory has indeed led to progress in practice. But not always. Sophisticated estimators, typically encapsulated in bits of software, amplify an old danger in econometrics, the black box problem.

Among the simplest estimators is Ordinary Least Squares (OLS), which can be visualized as finding the line that best fits a set of data points on a two-dimensional scatter plot of, for example, aid versus growth in 100 countries during the 1990s. However, even this staple of undergraduate textbooks is complicated enough that it is hard to perform by hand, as it involves the inversion of matrices. Computers make OLS easy. But in the process of hiding the complexity of the analysis they also hide the complexity of that which is analyzed. The following graphs, lifted from a classic paper by Francis John Anscombe, illustrate.[35] Each of the four scatter plots shows a hypothetical set of 11 observations on two variables, which could be aid and growth. Clearly each implies a different kind of relationship between the two variables. Yet all four yield *exactly* the same numerical results if plugged into a computer program that does OLS. They all have the same best-fit line.

[33] Clive W.J. Granger, "Investigating Causal Relation by Econometric and Cross-sectional Methods," *Econometrica* 37 (3), pp. 424–438, August 1969.
[34] One example is B. Mak Avin and Francisco Barillas, "Foreign Aid, Poverty Reduction, and Democracy," *Applied Economics* 34, pp. 2151–56, 2002.
[35] F.J. Anscombe, "Graphs in Statistical Analysis," *The American Statistician* 27(1), pp. 17–21, February 1973.



Anscombe built these examples to dramatize the value of graphing as part of a statistical analysis: a picture is worth a thousand numbers. The point here is broader: it is easy to deploy an econometric estimator without fully appreciating what is going on behind the columns of numbers spat out as results.

Another potential source of trouble in even this simplest of estimators is the use of "nearly collinear variables." If one ran an OLS regression of math test scores among elementary school children on their age measured in months and their age measured in days, the OLS procedure would not work. It would treat the two age variables as separate drivers of test scores yet fail to distinguish between them because they move in lockstep. Because OLS fails obviously in such cases, perfect collinearity is almost unheard of in published regressions. But *near*-collinearity can go undetected and produce results that look ordinary yet are sensitive to small changes in the underlying data, or that exaggerate the significance of small correlations.[36] These phenomena are akin to the fact that 1 / (1 – 0.999999) (which equals a million), differs radically from 1 / (1 – 1.000001) (negative one million) even though the difference between 0.999999 and 1.000001 is tiny. A small change in a number that goes into this formula causes a huge change in the number that comes out.

A common source of near-collinearity in the aid-growth literature has been the use of "quadratic terms" and "interaction terms." In their famous paper finding that aid raises growth in a good

[36] Christopher J. Green and Eric Kiernan, "Multicollinearity and Measurement Error in Econometric Financial Modeling," *The Manchester School* 42(4), pp. 257–69, December 1989; Aris Spanos and Anya McGirk, "The Problem of Near-multicollinearity Revisited: Erratic vs Systemic Volatility," *Journal of Econometrics* 180, pp. 365–93 (2002).

policy environment, Burnside and Dollar reach this conclusion by constructing a variable equal to the product of aid receipts and an index of economic policy quality.[37] Aid×policy takes its highest values in countries with good policies *and* much aid, and it was the apparent correlation of this term with growth that was the crux of the Burnside and Dollar thesis. In response to this influential work, other economists introduced their own interaction terms—aid×economic vulnerability, aid×negative export price shocks, aid×democracy—all meant to study how local conditions influence aid effectiveness.[38] Others, including HT and CRB, interacted aid with itself, producing aid^2. The squared term explodes in value at high levels, and can be used to test for diminishing returns to aid in the most aid-dependent countries.

Some of these interaction terms are nearly collinear with the variables from which they are made. For example, the influential Collier and Hoeffler study on aid to countries emerging from conflict finds that aid is particularly effective when given to countries a few years after a civil war has ended and if the country has good economic policies.[39] They base this conclusion on the statistical significance of a *triple* interaction term, aid×policy×post-conflict status. It turns out that this term is correlated 0.985 with the simpler aid×post-conflict status interaction variable, also included in some of their regressions.[40] Since 1.00 would indicate perfect correlation, it is unlikely that the estimator can reliably distinguish between these two interaction terms. Any apparent conclusions about the importance of economic policy to aid effectiveness in post-conflict environments may simply be about the importance of aid in post-conflict environments. No amount of staring at their results tables, however, will reveal this ambiguity.

And the black box problem grows worse with more complicated estimators. Most of the assertions theoreticians make about the superiority of newer estimators, such as the Generalized Method of Moments (GMM), are only certain for infinite data sets. Whether they are actually superior to simpler methods on finite, real-world data sets is generally less clear. Between the complexity and the uncertainty, the new estimators are more easily misused, however unintentionally.

The earliest aid-growth studies, starting in 1970, used OLS. In the mid-1990s, a somewhat fancier technique called Two-Stage Least Squares came into common use. It is the most straightforward technique for incorporating instruments in order to reduce endogeneity. More recently the advanced "difference GMM" and "system GMM" estimators have become popular in aid analysis and well beyond.[41] The arrival of the Roodman's "xtabond2" command for Stata in late 2004 raised their popularity. Ironically, a new CGD working paper by Roodman

[37] Burnside and Dollar, op. cit. note 18.
[38] Guillaumont and Chauvet, op. cit. note 24; Collier and Dehn, op. cit. note 24; Svensson, op. cit. note 24.
[39] Collier and Hoeffler, op. cit. note 24.
[40] David Roodman, "The Anarchy of Numbers: Aid, Development, and Cross-country Empirics," CGD Working Paper 32, July 2004.
[41] Manuel Arellano and Stephen Bond, "Some Tests of Specification for Panel Data: Monte Carlo Evidence and An Application to Employment Equations," *The Review of Economic Studies* 58(2), pp. 277–97 (1991); Richard Blundell and Stephen Bond., "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics* 87, pp. 11–143 (1998); David Roodman, "How to Do xtabond2: An Introduction to 'Difference' and 'System' GMM in Stata," Working Paper 103, Center for Global Development, Washington, DC, December 2006. Macro aid effectiveness papers that use the GMM estimators include Hansen and Tarp, op. cit. note 19; Dalgaard, Hansen, and Tarp, op. cit. note 25; Michaelowa and Weber, op. cit. note 29; Dreher, Nunnenkamp, and Thiele, op. cit. note 30; and Mishra and Newhouse, op. cit. note 31.

demonstrates that the complexity of these estimators has backfired even in some widely cited papers on the determinants of economic growth, producing results that appear to have been purged of endogeneity but which in fact have not.[42]

3.3 Publication Biases

Another source of the apparently puzzling state of affairs in the macro study of aid effectiveness is the set of incentives that researchers face to generate certain kinds of findings, which can effectively filter which results see the light of day, and which reach the public. These filters can loosely be called publication biases, stipulating a broad definition of "publication." In academia, "publication" has a particular meaning—acceptance in a journal. Here, the term is interpreted to include everything from researchers' decisions about which results to write up, to circulation of working papers, to publication in popular outlets as well as journals. One reason that some observers of the macro aid effectiveness literature feel whipsawed is that academic and popular outlets generate sometimes-contradictory incentives. Academia generally favors the publication of statistically significant results. Finding that things are not correlated is less interesting. But the popular media reward those ready to take public positions on the extremes, whether they see powerful correlations or none: witness the high profiles accorded to Sachs and Easterly.

Consider the academic biases first. A given regression can be run in a practically infinite number of ways—varying the estimator, or which data points are included, or which variables are controlled for, or which interaction terms are used. The laws of chance say that even if there is no relationship between the variables of interest, some of the variations will produce evidence of an improbable degree of correlation. Even a fair coin will sometimes come up five heads in a row. Every step in the process of research tends to favor the selection of those regressions that show significant results. A researcher who has just labored to assemble a data set on civil wars in developing countries since 1970, or to build a complicated mathematical model of how aid raises growth in good policy environment, will feel a strong temptation to zero in on the preliminary regressions that show her variable to be important. Sometimes it is called "specification search" or "letting the data decide." Researchers may challenge their own results obtained this way with less fervor than they ought over econometric issues such as endogeneity and collinearity.[43] Research assistants may do all these things unbeknownst to their supervisors.

The filtering continues during the publication process. Tight for time, a researcher may be more likely to write up the projects with significant results, deferring others with the best of intentions. And if two researchers with the highest standards study the same topic in somewhat different ways, the one finding the significant result is more likely to win publication in a prestigious journal.[44] Meanwhile, there is less career reward for researchers to spend time replicating and

---

[42] David Roodman, "A Short Note on the Theme of Too Many Instruments," Working Paper 125, Center for Global Development, Washington, DC, August 2007. The studies questioned are Kristin J. Forbes, "A Reassessment of the Relationship between Inequality and Growth," *American Economic Review* 90(4), pp. 869–87 (2000), on inequality and growth, and Levine, Loayza, and Beck, op cit. note 32, on financial sector development and growth.
[43] Michael C. Lovell, "Data Mining," *Review of Economics and Statistics* 65(1), pp. 1–12, February 1983.
[44] Theodore D. Sterling, "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa," *Journal of the American Statistical Association* 54(285), pp. 30–34, March 1959; Edgar L. Feige, "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy* 83(6), pp. 1291–96, December 1975; Frank Denton, "Data Mining as an Industry," *Review of Economics and Statistics* 67(1), pp. 124–27, February 1985.

checking the work of others.[45] Who the researchers work for sometimes matters too. Certainly it is hard to imagine that World Bank researchers like Burnside and Dollar would have published a paper finding no evidence of aid effectiveness in the prestigious *American Economic Review*. Then-World Bank economist William Easterly publicly doubted their optimism in his first book, and ended up at CGD. It must be conceded that Rajan and Subramanian did publish and publicize a skeptical paper about aid while at the IMF. The IMF, however, is not as centrally involved as the World Bank is in foreign aid as that term is usually understood.

The respected UCLA economist Edward Leamer characterized the state of affairs this way in his call to "take the con out of econometrics":

> The econometric art as it is practiced at the computer terminal involves fitting many, perhaps thousands, of statistical models. One or several that the researcher finds pleasing are selected for reporting purposes. This search for a model is often well intentioned, but there can be no doubt that such a specification search invalidates the traditional theories of inference….[A]ll the concepts of traditional theory…utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose.
> …
> This is a sad and decidedly unscientific state of affairs we find ourselves in. Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously. Like elaborately plumed birds who have long since lost the ability to procreate but not the desire, we preen and strut and display our *t*-values.[46]

It would be unfair to leave the impression, however, that these foibles only afflict seekers of significant results, for those trying to demonstrate *lack* of significance can mine, however unwittingly, for findings that confirm their biases too. And academia does make some space for the skeptics. The top-tier *American Economic Review* published the Easterly, Levine, and Roodman comment on Burnside and Dollar. The respected *Review of Economic Statistics* has accepted Rajan and Subramanian's paper.[47] But the profession as a whole is still biased toward significant results. As economists J. Bradford DeLong and Kevin Lang pointed out in their paper, "Are All Economic Hypotheses False?", economics journals are filled with an improbable number of significant findings—improbable, that is, unless one accepts the possibility of publication bias.[48]

Another consequence of "invalidat[ing] the traditional theories of inference" is that seemingly strong results, on the effectiveness or ineffectiveness of aid, say, can turn out to be fragile—liable to disappear in response to small changes in the data. Most estimators do not just estimate numbers like the effect of aid on growth; they also calculate how likely it is that the observed

---

[45] Gordon Tullock, "Publication Decisions and Tests of Significance—A Comment," *Journal of the American Statistical Association* 54(287), pp. 593, September 1959; Daniel S. Hamermesh, "Replication in Economics," Working Paper 13026, National Bureau of Economic Research, Cambridge, MA, April 2007.
[46] Edward E. Leamer, "Let's Take the Con out of Econometrics," *American Economic Review* 73(1), pp. 31–43, March 1983. *t*-values measure statistical significance.
[47] Easterly, Levine, and Roodman, op. cit. note 1; Rajan and Subramanian, op. cit. note 3.
[48] J. Bradford De Long and Kevin Lang, "Are all Economic Hypotheses False?" *Journal of Political Economy* 100(6), pp. 1257–72 (December 1992).

correlation would surface if in fact there is no relationship.[49] If that probability is low, one can infer that the relationship is almost certainly real. But if many regressions are run in a search for the most pleasing answer, that invalidates the logic. It is the difference between flipping a coin five times and getting heads every time and flipping it thousands of times until one gets five heads in a row—and only reporting that event. Only the first scenario suggests a biased coin. The second is a recipe for fragility: when someone else repeats the *reported* experiment of flipping . the coin five times, he will probably observe a different outcome. This helps explain the fragility in many aid-growth studies, as documented by Dalgaard and Hansen; Easterly, Levine, and Roodman; and Roodman (see section 2).

Operating according to different principles, the press rewards the extremists on both sides of any issue, and that creates incentives that seep somewhat into academia. The reasons are well-known: the human mind is drawn to conflict; the underdog is inherently attractive; in the United States at least, there is the ideal of "objectivity" in the media, which created a tradition of according both sides in a debate equal weight. This can frustrate the attempts of economists who avoid extremes to broadcast more nuanced messages. Emphasizing the poles can also generate cognitive dissonance for the lay reader. Thus in different ways academia and the press conspire against the reasonable middle.

4 The Current Debate

In light of the above, this section attempts to objectively summarize for the lay reader the latest debates over the macroeconomic effects of aid, by reviewing the contentions of major papers and the contentiousness they have engendered. The word "attempt" is chosen advisedly, for the task is impossible twice over. First, almost anyone who understands the arguments well has become invested in a published position. Second, the deep disagreements are inherently technical, about whether certain constructed variables are correlated with certain other variables that cannot even be observed. At some point, the concerned lay reader may need to draw her own conclusions from the fact that experts disagree. The account picks up after Easterly, Levine, and Roodman's strong challenge to Burnside and Dollar destroyed a conventional wisdom that aid had been shown to work in "good policy environments," and after Roodman's "Anarchy of Numbers" had broadened the critique of fragility to more studies.

*4.1 Dalgaard, Hansen, and Tarp (***The Economic Journal***, 2004)—The Aid-Tropics Link*

Before joining with Dalgaard to publish in the top British economics journal, Danish economists Hansen and Tarp had led the "gainsayers" in arguing that aid does work on average, if with diminishing returns at high levels. In the new paper, the trio (DHT) joined the "apologist" camp, emphasizing that aid may work in some places but not others.[50] In particular, DHT found that aid works outside the tropics, but not in them, a thesis that resonates with the work of Acemoglu, Johnson, and Robinson about the deep sources of economic development (described at the end of section 3.1). DHT build their model on a foundation of some intricate theory and apply it using sophisticated, instrumenting estimators such as difference and system GMM. The pattern they find is strikingly strong: aid raises economic growth outside the tropics but not in them.

[49] More precisely, estimators calculate the probability of observing the aid-growth relationship that is actually observed if the true effect is zero.
[50] Dalgaard, Hansen, and Tarp, op. cit. note 25.

It turns out, however, that the pattern is really the story of a few countries: most of all Jordan, but also Syria, Egypt, and Botswana. The first three countries, and 30 percent of Botswana, are outside the tropics. All had episodes in which high aid and high growth coincided. Removing these four countries eliminates the DHT results.[51] Since Jordan, Syria, and Egypt are linked by common historical experiences, the DHT finding about aid outside the tropics essentially distills down to two cases, one of Middle Eastern countries that are not major oil producers, the other of Botswana, an African nation historically tied to the economy of South Africa.

This raises two questions. Looking behind the numbers for these few countries, does history suggest that aid raised growth in them? And if so, how much can we generalize from their experiences? As for the first question, aid may well have raised growth in these countries. But other stories are plausible too. In the oil boom years of the 1970s and 1980s, Jordan received 5–15% of GDP in aid, mostly from other Middle Eastern nations. Simultaneously, the boom directly stimulated Jordan's economy. Its Port of Aqaba is an important transshipment point for oil, for example. Meanwhile, Botswana's economic success may be due in part to its membership in the South African Customs Union, which deprived the government of an independent trade policy and the associated, dangerous opportunities for government officials to meddle with import and export rules in order to extract bribes from traders. Why did such an economically helpful union arise in the southern cone of Africa? Perhaps because the non-tropical climate was more hospitable to European settlers, leading to the creation of the South African state. Botswana's "non-tropical" institutional advantage and the resulting growth may have made aid more effective; or they may have attracted more aid.

A more thorough examination of the histories of these four countries might support the DHT story of aid raising growth. Or they might not. It seems difficult, however, to generalize from them to Guatemala and India and Laos and the universe of poor countries.

4.2 Clemens, Radelet, and Bhavnani (CGD Working Paper, 2004)—The Short-term Impact of Aid

Recall that the major innovation introduced by Clemens, Radelet, and Bhavnani is to recognize that aid comes in many forms and serves many purposes. Much of it cannot be expected to raise economic growth, especially within the timeframes usually studied. They zeroed in on "short-impact" aid—budget and balance of payments support and aid for infrastructure and industry. CRB found a strong, immediate effect of short-impact aid on growth and demonstrated the robustness of their results to many changes in the data and estimator.[52]

Facing contradictions from other authors, CRB have made several points in their own defense. First, their study incorporates a number of design choices that together distinguish it from all others. Rather than combining aid given to countries and debt payments received from them into one "net aid" variable, they enter separate variables for gross aid disbursements and debt service. Copying Hansen and Tarp, they include both gross aid and its square to allow for diminishing returns to aid.[53] And of course they include only "short" aid in an analysis that, by design, can only pick up effects that appear within a few years. Theory, moreover, backs these choices. It

[51] Roodman, op. cit. note 27.
[52] Clemens, Radelet, and Bhavnani, op. cit. note 2.
[53] Hansen and Tarp, op. cit. note 19.

would be improbable for short aid to have exactly the same dollar-for-dollar effect on growth as debt servicing (with opposite sign). Receiving aid funds for a road is not the same as writing a check from the general treasury to pay back an aid loan. Entering aid disbursements separately from debt servicing in the statistical work acknowledges this reality. Similarly, standard economic growth theory says that there are diminishing returns to capital investment, so one would expect the benefits of aid to taper off as it rises.

Finally, CRB's specification tests suggest that their instruments are valid, so that the correlation they uncover can be interpreted as causation from aid to growth. In particular, they report that their regressions are free of a phenomenon known as autocorrelation, whose significance is explained just below.

These are serious arguments, though not airtight. Almost any choice a researcher makes can be justified on the grounds of some theory, since theories are just thoughtful stories that simplify reality. Just as when choosing instruments, theory is useful discipline but not a magic talisman. None of the contemporary aid-growth studies is fully grounded in any comprehensive theory that justifies, say, the use of four-year periods in a country's history as the unit of observation, or the inclusion of a variable indicating whether the country was embroiled in civil war four years ago. And none could be. So to some extent all those who invoke theory in their favor live in glass houses.

CRB's invocations of theory should therefore be taken in a constructive spirit as a list of their study's virtues. Against these strengths, there are also concerns. For one, by entering both aid and its square, CRB (and HT) create a collinearity problem, since the countries with high aid also have high aid^2. For technical reasons, the problem is particularly severe in instrumented regressions.

A more serious worry about CRB, HT, and other studies with similar instruments has to do with these instruments. For example, CRB's instrument set includes, for each four-period in a country's history, the level of short-impact aid it received in the *previous* four-year period. For technical reasons, such "lagged" instruments can only be relied upon to expunge reverse and third-variable causation if the regressions are free of "autocorrelation," meaning that there is no significant tendency for China, say, to repeatedly grow faster than the regression model predicts, or for Argentina to repeatedly grow more slowly. It turns out that the CRB aid-growth regressions, like many others, are autocorrelated.[54] Autocorrelation has often gone undetected because extreme growth volatility in a few countries, such as Gabon and Nicaragua, masks the overall pattern of growth persistence unless removed.[55]

The presence of autocorrelation does not automatically make the CRB instruments invalid. And even if they are invalid, meaning that reverse and third-variable causation cannot be ruled out in the aid-growth relationship they find, that does not mean that short-impact aid does not raise growth. Nevertheless, the shakiness of the instruments makes it is hard to be as confident in the causal claims about short-impact aid as one would like.

---

[54] Rajan and Subramanian, op. cit. note 3, raised this general issue, and Roodman, op. cit. note 51, shows autocorrelation in several studies.

[55] David Roodman, "How Consequences become Causes: Through the Looking-Glass, and What OLS Found There," Working Paper, Center for Global Development, forthcoming.

4.3 Rajan and Subramanian (IMF Working Paper, 2005; The Review of Economics and Statistics*, forthcoming)—Encompassing Other Studies, Striving for Better Instruments*

As mentioned earlier, a starting point for Rajan and Subramanian (RS) was the worry that the instrumentation strategies in most aid-growth studies are doubtful and not grounded in theory.[56] They observed that influential papers such as those of Burnside and Dollar use instruments for aid based on policy quality or other variables that are probably caused by growth. But such instruments are invalid if they are *directly* related to growth, outside any link via aid. Burnside and Dollar's instrument set, for instance, includes policy×population and similar terms. But growth shocks are well-known to directly affect such "policy" indicators as budget deficits, which arguably invalidates them as instruments.

RS proceeded to run two kinds of regressions, both with instruments they argue are superior. In response to the cacophony of competing stories of aid effectiveness that had built up in the literature (described in section 2), RS use this framework to systematically test many of the stories, in turn. The first kind of regression, unusual since the mid-1990s, have just a single data point for each country, using aid and growth averaged over a long time frame such as 1960–2000. For these "cross-section" regressions, RS construct a novel instrument based primarily on past colonial history and language affinity (whether donor and recipient share a common language). In the second group are "panel" regressions, with one observation for each country and five-year period, as is now common. Here, RS use the popular "GMM" estimators mentioned in section 3.2, which in a complicated way use past values of aid and other variables to instrument current values.

Despite trying many variations, RS find no consistent effect of aid on growth in either group of regressions. They conclude that the effect, if any, is too small to be detected.

Interestingly, however, these two skilled economists who dedicated themselves to solving the instrumentation problem ultimately struggled to do so. As set forth in the original IMF working paper, the novel instrument for the cross-section regression turns out to have serious problems. It effectively predicts, for example, that the United Kingdom gives the same absolute amount of aid to Kiribati and India since both are English-speaking former colonies. After dividing by recipient GDP, tiny Kiribati is then predicted to get vastly more aid for its size than India. So extreme is the effect that the original RS cross-section results were driven by the numbers for a few small-island states. Importantly, the final version of the study largely eliminated this mathematical problem, and this does not change the results.

The panel regressions also exhibit a basic instrumentation problem, though it too may not affect the results. There appear to be about two instruments for every three data points. As discussed in section 3.1, numerous instruments are collectively invalid no matter how valid each one is alone.[57] It is somewhat surprising that in a paper dedicated to instrumenting well, all the panel regressions have this problem. Again, RS reported that if the instrument problem is fixed, by truncating the instrument collection, their results do not change.

[56] Rajan and Subramanian, op. cit. note 3.
[57] Roodman, op. cit. note 42.

On balance, the final RS instrumentation strategies are not perfect, as none will ever be. But because RS reached a negative conclusion, the remaining imperfections may not greatly undermine their thesis. The spirit of their endeavor was to do an instrumented aid-growth study as well as can be done; it leads them to the conclusion that aid's impact is too small to detect by the means available to today's economists. Aid may be poorly instrumented in their study—but arguably instrumented about as well as it can be. Or it may be well instrumented and yet does not correlate with growth. Both possibilities fit the conclusion.

4.4 Roodman (CGD Working Paper, forthcoming)—OLS in Wonderland

A forthcoming working paper, in addition to casting some of the above critiques in technical language, tells a mathematical story of how statistical findings that aid works can arise from something quite opposite.[58] Under the right circumstances, the tendency for GDP growth to *reduce* aid as a share of the recipient's GDP (which is usually how aid is measured) can look like aid *raising* growth—an effect with the opposite sign and opposite direction.

For intuition, imagine two countries that are identical at time t except that one's economy is growing faster than the other. Since the countries have the same GDP at t, the faster-growing one has a *smaller* GDP at $t - 1$. So, all else equal, the faster-growing country had *higher* aid/GDP at time $t - 1$. Thus can higher aid/GDP precede—and appear to cause—higher growth. And if one plays the tape forward from time t to time $t + 1$, the faster-growing country ends up with lower aid/GDP in the future. Turning to the data, Roodman finds that growth today is indeed positively correlated with past aid and negatively correlated with future aid. This pattern *could* appear if aid causes growth, which then "repels" further aid, donors shifting to less successful countries. But this is complicated story. Roodman invokes Occam's Razor, the principle that the simpler theory is more likely correct, in favor of his own more mathematical explanation.

Roodman also applies Granger causality testing (mentioned at the end of section 3.1) to the aid-growth relationship. It turns out that knowing how much aid a country received does not help predict its economic growth in the following years. But knowing how much it grew does help predict how much aid it will subsequently get, as a share of GDP. Aid tends to follow growth rather than vice versa.

Roodman's analysis confirms that aid researchers have been right to worry about reverse causation. The dominant link between aid and growth appears to be that the second influences the first. Only good instruments can remove that first-order connection so that the subtler influences of aid on growth can shine through. And history shows that a good instrument is hard to find.

4.5 New Sector-level Studies

As mentioned in section 2, the last few years have seen the circulation of several studies that carry on the spirit of CRB, narrowing the aid variable and, in these papers, the outcome of interest too. Given the difficulties encountered in previous work, it seems smart to scale back the ambition somewhat by narrowing the variables. Michaelowa and Weber and, separately, Dreher, Nunnenkamp, and Thiele look at whether aid for education raises primary enrollment rates.[59]

[58] Idem, op. cit. note 55.
[59] Michaelowa and Weber, op. cit. note 29; Dreher, Nunnenkamp, and Thiele, op. cit. note 30.

Mishra and Newhouse of the IMF do the same for health aid and infant mortality.[60] The studies are similar in many respects. All use "panel" data sets with one observation every four or five years for each country, and, like the RS panel regressions, rely heavily on GMM estimators as implemented in Roodman's software. All find aid to be effective. And all face similar questions about the validity of their instruments

Michaelowa and Weber run their initial regressions on a panel with one observation for each country and five-year periods, taking commendable care in choosing their instruments and limiting their number. They find no consistent effect of education aid on the number of children who go to school. Then they switch to a panel with separate observations for each *year*. In some of these, education aid pops out as positively related to education. The authors then hone in on one of these regressions for further analysis, despite poor results on specification tests that suggest that causal claims should be doubted. They are refreshingly frank in explaining why they mine the data: "we attempt to ensure that our final results indicate the most optimistic, rather than the most pessimistic results for the effect of aid on educational outcomes."

The Dreher, Nunnenkamp, and Thiele study, also on education aid, is less rigorous in containing the instrument count. As explained in Roodman's "Short Note on the Theme of Too Many Instruments," this instrument proliferation may produce estimation results that are invalid and results on specification test, which are meant to check instrument validity, that are falsely reassuring.[61] Once again the upshot is doubt about whether researchers can convincingly go beyond demonstrating correlations to proving causation.

Following common practice, the Mishra and Newhouse regressions, which relate health aid to infant mortality, include the "lagged dependent variable" as a control, the estimated coefficient on which is almost exactly 1.0. In other words, the lower a country's infant mortality was in one five-year period, the lower it was in the next, all else equal, and the relationship is exactly one-to-one. Unfortunately, one technical assumption necessary for the validity of the instruments they use is that this relationship be *less* than one-for-one. This would have been the case if gains in infant mortality appeared to fade over time unless continually reinforced by aid, economic growth, or other factors.

It does not appear that narrowing the focus has yet helped researchers find better ways to instrument aid. As a result, proof that aid works at the macro level remains elusive.

5 Which Way Forward?

What should we learn from this history? The quantitative approach to answering grand questions about aid effectiveness has repeatedly offered hope and repeatedly disappointed. The recent move toward narrowing aid variables, pioneered by CRB, is worth pursuing farther. It reduces the ambition of the literature. But grounds for optimism seem narrow.

[60] Mishra and Newhouse, op. cit. note 31.

[61] Roodman, op. cit. note 42. Torben G. Andersen and Bent E. Sørensen, "GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study," *Journal of Business and Economic Statistics* 14(3), pp. 328–52 (July 1996); Clive G. Bowsher, "On Testing Overidentifying Restrictions in Dynamic Panel Data Models," *Economics Letters* 77, pp. 211–20 (2002).

Amid all the public interest in the literature, a question that too often goes unasked is how practically relevant it is. The "apologist" studies that ask where or when aid works do have clear policy relevance. The iconic example is that of Burnside and Dollar, whose paper clearly influenced the design of the U.S. Millennium Challenge Account. But these are also the studies that are now least trusted within academia, however reasonable their conclusions about the real world. The relevance of the "gainsayers'" contention that aid works on average, if with diminishing returns, is also murky. Does a study finding that aid generally does or does not raise economic growth affect how much is given? Deep national characteristics probably have more to do with why the Danes and Dutch give far more foreign aid per person than Americans and Japanese. And the global aid increase in the last ten years probably has more to do with U.S.-led wars in Iraq and Afghanistan and a shift in aid politics tracing to the Jubilee 2000 debt cancellation movement.

Finally, if a new sectoral study makes a convincing case that a certain kind of aid—health aid, education aid, infrastructure aid—is effective, that is reassuring, but leaves unanswered important practical questions. Should aid be moved from other sectors to the one shown to be effective? Which of the myriad forms of aid within the sector are responsible for the success? How can the aid be improved?

Other forms of analysis, with their own virtues and limitations, are therefore needed to round out our understanding of foreign aid. Rigorous evaluations of precisely defined interventions can provide more solid information about when aid-backed programs work. Less narrow in scope and less formally rigorous are case studies that bring out the full richness of a national context as it influences aid effectiveness—detail that gets lost in the numbers of quantitative analysis.

On balance, it seems that the macro research has attracted attention out of proportion to its value. Even the optimists caution against taking their results too literally and recognize how limited is the insight they can give policymakers. Meanwhile, the skeptics tend to believe that limits to human knowledge were long ago reached and cannot be breached. Attacking smaller, practical questions, such as about the effects in various contexts of microfinance or roads, is more likely to achieve what ought to be the primary purpose of studying aid effectiveness, which is to improve it.