WORKING PAPER NUMBER 32
JULY 2004

Center for Global Development

The Anarchy of Numbers: Aid, Development, and Cross-country Empirics

by David Roodman

Abstract

Contemporary econometric literature has generated a large number of stories of the relationship between how much foreign aid a countries receives and how it grows. All the stories hinge on the statistical significance in cross-country regressions of a quadratic term involving aid. Among the stories are that aid raises growth (on average) 1) in countries where economic policies are good; 2) in countries where policies are good and a civil war recently ended; 3) in all countries, but with diminishing returns; 4) in countries outside the tropics; 5) in countries with difficult economic environments, characterized by declining or volatile terms of trade, natural disasters, or low population; or 6) when aid increases in countries experiencing negative export price shocks. The diversity of results prima facie suggests that many are fragile. Easterly *et al.* (2004) find the aid-policy story (Burnside and Dollar, 2000) to be fragile in the face of an expansion of the data set in years and countries. The present study expands that analysis by applying more tests, and to more studies. Each test involves altering just one aspect of the regressions. All 19 tests are derived from sources of variation that are minimally arbitrary. Twelve derive from specification differences between studies, what Leamer (1983) calls "whimsy." Three derive from doubts about the appropriateness of the definition of one variable in one study. The remaining four derive from the passage of time, which allows sample expansion. This design allows an examination of the role of "whimsy" in the results that are tested while minimizing "whimsy" in the testing itself. Among the stories examined, the aid-policy link proves weakest, while the aid-tropics link is most robust.

The Anarchy of Numbers: Aid, Development, and Cross-country Empirics

David Roodman[1]

Center for Global Development

This draft: July 2004

Abstract: Contemporary econometric literature has generated a large number of stories of the relationship between how much foreign aid a countries receives and how it grows. All the stories hinge on the statistical significance in cross-country regressions of a quadratic term involving aid. Among the stories are that aid raises growth (on average) 1) in countries where economic policies are good; 2) in countries where policies are good and a civil war recently ended; 3) in all countries, but with diminishing returns; 4) in countries outside the tropics; 5) in countries with difficult economic environments, characterized by declining or volatile terms of trade, natural disasters, or low population; or 6) when aid increases in countries experiencing negative export price shocks. The diversity of results prima facie suggests that many are fragile. Easterly *et al.* (2004) find the aid-policy story (Burnside and Dollar, 2000) to be fragile in the face of an expansion of the data set in years and countries. The present study expands that analysis by applying more tests, and to more studies. Each test involves altering just one aspect of the regressions. All 19 tests are derived from sources of variation that are minimally arbitrary. Twelve derive from specification differences between studies, what Leamer (1983) calls "whimsy." Three derive from doubts about the appropriateness of the definition of one variable in one study. The remaining four derive from the passage of time, which allows sample expansion. This design allows an examination of the role of "whimsy" in the results that are tested while minimizing "whimsy" in the testing itself. Among the stories examined, the aid-policy link proves weakest, while the aid-tropics link is most robust.

[1] The author thanks William Easterly and Michael Clemens for advice, Craig Burnside, Lisa Chauvet, Jan Dehn, and Henrik Hansen for data and assistance, and Stephen Knack, Patrick Guillaumont, and participants in an August 2003 seminar at the Center for Global Development for valuable comments. Responsibility for the final product rests with the author. All judgments, opinions, and errors are those of the authors alone and do not represent the views of the Center for Global Development, its staff, or board of directors.

Introduction

In the early weeks of 1981 economist Edward Leamer gave a speech at the University of Toronto, in which he bemoaned the state of econometrics. Econometrics sought the status of a science, with regressions its analog for the reproducible experiments of chemistry or physics. Yet an essential part of econometric "experiments" was too often arbitrary, opaque, and unrepeatable. Adapting the speech for the *American Economic Review*, he wrote:

> 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. The concepts of unbiasedness, consistency, efficiency, maximum-likelihood estimation, in fact, all 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. (Leamer, 1983)

The way out of the quagmire, Leamer argued, was for econometricians to explore large regions of "specification space," systematically analyzing the relationship between assumptions and conclusions:

> [A]n inference is not believable if it is fragile, if it can be reversed by minor changes in assumptions. As consumers of research, we correctly reserve judgment on an inference until it stands up to a study of fragility, usually by other researchers advocating opposite opinions. It is, however, much more efficient for individual researchers to perform their own sensitivity analyses, and we ought to be demanding much more complete and more honest reporting of the fragility of claimed inferences….The job of a researcher is then to report economically and informatively the mapping from assumptions into inferences. (Leamer, 1983)

One econometric debate that has worrying hallmarks of Leamer's syndrome is that on the

effectiveness of foreign aid to developing countries. Since Griffin and Enos (1970), econometricians have parried over the question of how aid affects economic growth in recipient nations. Prominent in the contemporary work, Burnside and Dollar (2000) concluded, "aid has a positive effect on growth in a good policy environment," presumably because good policies increase the productivity of aid-financed investment. Their evidence: the statistical significance in cross-country panel growth regressions of an interaction term of total aid received and an indicator of the quality of recipient economic policies.

The work of Burnside and Dollar has brought corroborations (Collier and Dehn, 2001; Collier and Hoeffler 2002; Collier and Dollar, 2002, 2004) and challenges. From the ongoing debate have emerged several stories of the relationship between aid and growth, each of which turns on a particular quadratic or cubic term involving aid. The stories are not incompatible, but most papers support only one, a few two. Hansen and Tarp (2001) find that entering the square of aid drives out the significance of Burnside and Dollar's aid×policy, and makes the simple aid term significant too: aid works on average, but with diminishing returns. Guillaumont and Chauvet (2001) also fail to find significance for aid×policy, and instead offer evidence that aid works best in countries with difficult economic environments, characterized by volatile and declining terms of trade, low population, and natural disasters, perhaps because aid finances institutions that help cushion the shocks. In the same vein, Collier and Dehn (2001) find that *increasing* aid to countries experiencing negative export price shocks raises growth. Meanwhile, Collier and Hoeffler (2002) offer a triple-interaction term: aid works particularly well in countries that are recovering from civil war *and* that have good policies. Last, Hansen and Tarp, along with Dalgaard, tell a new story: aid raises growth outside the tropics but not in them, low latitudes apparently being associated with some combination of inhospitable natural environments and poor

policies and institutions (Dalgaard *et al.*, 2004).

These papers naturally differ not only in their conclusions and policy implications but also their specifications. Within the group, there are two different choices of period length in the panel data sets, three definitions of "policy," three of aid, and four choices of control variable set. Though probably none of the choices are truly made on a whim, these differences appear to be examples of what Leamer called "whimsy." From Leamer's point of view, the studies taken together represent a small and unsystematic sampling of specification space. Without further analysis, it is impossible to know whether the results reveal solid underlying regularities in the data or are fragile artifacts of "whimsy."

In this paper, I therefore examine the possibility of fragility in this literature systematically. Since by the laws of chance any regression can be broken with enough experimentation, it is essential for credibility that the testing suite itself be un-whimsical. I derive my tests therefore from three canonical sources of variation: the "whimsy" already present in the original specifications; doubts about the appropriateness of the definition of one variable, the Collier and Dehn negative shock variable; and the passage of time, which allows expansion of data sets. Each test ideally involves varying just one aspect of the regressions, although sometimes additional changes are made so that the modified regressions pass specification tests. In all, I subject 8 regressions from 7 of the most prominent studies to this systematic test suite.

The rest of this paper runs as follows. Section I places the recent empirical work on aid in historical context and reviews the approaches and conclusions of the studies that are tested for robustness. Section II describes the testing regime. Section III reports results. Section IV concludes.

I. History

It can be said that foreign aid for poor countries was born on the steps of the U.S. Capitol just

after noon on January 20, 1949. At that moment, the re-elected President Harry Truman gave his inaugural address, in which he enunciated a historically novel vision of the relationship between the democratic west and what is now called the "developing world." The poorer nations were no longer just colonies or sources of raw materials for the West to exploit, but potential allies in the Cold War, which must be won over to capitalism and democracy. By sharing its capital and technical know-how, the United States could make poor nations richer and expand the sphere of freedom.

As Truman spoke, American assistance for the rapid, aid-financed *redevelopment* of Western Europe was already underway, and would soon succeed unambiguously. But the effectiveness of aid for *development* in poorer countries has for decades been a point of intense controversy. Bauer (1976), Hancock (1989), Easterly (2001), and Reusse (2002), among others, have pointed out how foreign assistance has been undermined by geopolitics, pressures from domestic commercial interests for contracts, arrogance and perverse incentives within aid-giving institutions, and the sheer difficulty of fostering development from the outside. Against these criticisms, defenders have pointed to successes such as the eradication of smallpox, the widespread vaccination of children against other diseases, and support for family planning that helped slow population growth.

The hope has often arisen that a turn to the numbers would shed light on the contentious questions of whether and when aid works. Starting in 1970, cross-country empirics were brought to bear to examine how aid affects macro variables such as investment, growth, and poverty. In the view of Hansen and Tarp (2000), this literature has gone through three generations. The first generation essentially spanned 1970–72, and mainly investigated the aid-*savings* link. Influenced by the Harrod-Domar model, in which savings is the binding constraint on growth, aid-induced

saving was assumed to lead directly to investment, thence to growth via a fixed incremental capital-output ratio. Not considered were the realistic possibilities that some aid is consumed and that some invested aid has low, zero, even negative returns. The second generation of studies, which ran from the early 1970s to the early 1990s, avoided the simplistic assumptions about savings by directly investigating the relationship between aid on the one hand and investment and growth on the other.

Hansen and Tarp argue that the preponderance of the evidence from these first two generations says that 1) aid increases total savings but less than one-to-one, partly substituting, partly complementing other sources of savings; and 2) that aid increases investment and growth. They suggest that studies with more pessimistic results, such as Mosley *et al.* (1987), have gained disproportionate attention despite being in the minority precisely because they are contrarian.

The third generation commenced with Boone (1994) and continues to this day. It has brought several innovations. The data sets cover more countries and years. Reflecting the influence of the new growth theory, regressors are typically included to represent the economic and institutional environment (sometimes together called the "policy environment"). It has become the norm to address methodologically the potential endogeneity of aid and policy to growth, usually with two-stage least squares. And, finally, the aid-growth relationship is allowed to be non-linear, through incorporation of such regressors as aid^2 and aid×policy. (Hansen and Tarp, 2000) The data sets are almost always panels. The studies I test for robustness all belong to the third generation.

Burnside and Dollar disseminated their first results on aid effectiveness in 1997, as a World Bank working paper that eventually appeared in the *American Economic Review* (2000).

They test whether an interaction term of aid and an index of recipient economic policies is significantly associated with growth. Their data set is a panel drawn from developing countries outside the former Eastern bloc, covering the six four-year periods in 1970–93. They incorporate some controls found significant in the general growth literature, namely: initial income (log real GDP/capita) to capture convergence; ethno-linguistic fractionalization (Easterly and Levine, 1997), assassinations/capita, and the product thereof; the Knack-Keefer (1995) institutional quality variable, called "ICRGE"; M2/GDP, l to indicate financial depth, lagged one period to avoid endogeneity (King and Levine, 1993); dummies for sub-Saharan Africa and fast-growing East Asia; and period dummies.

Burnside and Dollar use a measure of aid called Effective Development Assistance (EDA), as computed by Chang *et al.* (1998). EDA differs in two major respects from the usual net Overseas Development Assistance measure (net ODA) tabulated by the Paris-based Development Assistance Committee (DAC). First, EDA excludes technical assistance, on the grounds that it funds not so much recipient governments as consultants. Second, it differs in its treatment of loans. Net ODA counts disbursements of concessional (low-interest) loans only, but at full face value.[1] It nets out principal but not interest payments on old loans because it is built on a capital flow concept. In contrast, EDA includes all disbursements of all development loans regardless of how concessional they are (for example, near-commercial rate loans by the World Bank to middle-income countries such as Brazil), but counts only their grant element, that is, their net present value. To express EDA as a share of recipient GDP, Burnside and Dollar convert it to constant 1985 dollars using the IMF's world Export Unit Value Index, then divide into real GDP from Penn World Tables 5.5 (Summers and Heston, 1991).

[1] DAC considers a loan concessional if it has a grant element of at least 25 percent of the loan value, using a 10 percent discount rate.

To represent recipient economic policies with a single variable, Burnside and Dollar first run a growth regression without aid terms, but with all controls and three indicators of economic policy—log (1+inflation), budget balance/GDP, and the Sachs-Warner (1995) variable measuring openness to trade. The coefficients of all three policy variables differ from 0 at the 0.05 level in this regression, so Burnside and Dollar form a linear combination of the three using the coefficients as weights. This is their policy index.[2]

When Burnside and Dollar run their base specification—containing the controls, the policy index, aid, and aid×policy—on their full data set, the term of central interest, aid×policy, does not in fact enter significantly, whether in OLS or in 2SLS with aid and aid×policy instrumented. However, they find that it becomes significant after either of two possible changes. Five outlier observations can be excluded (Burnside and Dollar's preferred specification). Or a cubic term can be added—aid^2×policy, in which case both aid×policy and aid^2×policy appear significant, the first with positive sign, the second negative. Burnside and Dollar famously conclude that aid raises growth in a good policy environment, but with diminishing returns.

Burnside and Dollar's work triggered responses in the literature, some critical, some supportive. Hansen and Tarp (2001) make one prominent attack. They modify the Burnside and Dollar 2SLS regressions in several ways, most importantly by adding an aid^2 term. Aid×policy is not significant in their results, but aid and aid^2 are, the first positive and the second negative. The implication is that aid is effective on average, but with diminishing returns—regardless of recipients' policies as far as the evidence goes. Hansen and Tarp then criticize both the Burnside and Dollar regressions and their own 2SLS regressions for failing to handle several econometric problems. There may be unobserved country-level effects that correlate with both policies and growth. Failing to purge or control for all such effects could give spurious explanatory power to

[2] They also add a constant term to the index, but this has no effect on the regression results of interest here.

policies and aid×policy. Also, variables other than aid and its interaction terms, such as fiscal balance, could be endogenous and ought to be instrumented. Thus they deploy the Arellano-Bond GMM estimator. This estimator runs in first differences, which purges country effects, and it allows the instrumentation of many variables, using their own lags. In switching to this "difference GMM" estimator, Hansen and Tarp also modify the regressor list. They drop fiscal balance as a policy variable. And they add Δaid and $\Delta(\text{aid}^2)$. Hansen and Tarp's key results on aid and aid^2 hold. And Δaid and $\Delta(\text{aid}^2)$ are significant too, again the first with positive sign and the second negative.

Guillaumont and Chauvet (2001) tell a third story of aid effectiveness. They hypothesize that the economic vulnerability of a country influences aid effectiveness. They call economic vulnerability the "environment," not to be confused with Burnside and Dollar's "policy environment." In this story, aid flows stabilize countries that are particularly buffeted by terms of trade difficulties, other sorts of external shocks, or natural disasters. In the spirit of Burnside and Dollar, Guillaumont and Chauvet build an environment index out of four variables: volatility of agricultural value added (to proxy for natural disasters), volatility of export earnings, long-term terms of trade trend, and log of population (small countries being more vulnerable to external forces). Their specification is unique in this literature in using 12-year periods, and in its set of controls, which are partly inspired by Barro (1991) and Mankiw *et al.* (1992). They control for population growth, mean years of secondary school education among adults, a Barro-Lee measures of political instability based on assassinations and revolutions, ethno-linguistic fractionalization, and lagged M2/GDP. In their OLS and 2SLS regressions, aid×environment appears with the predicted negative sign, indicating that aid works better in countries with worse environments. The term also drives out the significance of aid×policy.

Other papers have reached conclusions similar to those of Burnside and Dollar. Collier and Dollar (2002) corroborate them with a quite different data set and specification. They use OLS only. They include former Eastern bloc countries, the Bahamas, and Singapore. They use net ODA rather than EDA. They study 1974–97 instead of 1970–93. They drop all Burnside and Dollar controls except log initial GDP/capita, ICRGE, and period dummies. They add region dummies.[3] And they define policy as the overall score from the World Bank's Country Policy and Institutional Assessment (CPIA), which is a composite rating of countries on 20 aspects of policies and institutions.[4]

In a paper that takes the Collier and Dollar core regression as its starting point, Collier and Hoeffler (2002) analyze how aid effectiveness is influenced by whether a country has recently ended a civil war. Sticking to the four-year panel arrangement, they create three dummies to indicate how recently civil war ended. The dummy for "peace-onset" is 1 in the period when a country goes from civil war to peace. "Post-conflict 1" takes a value of 1 in the following period, and "post-conflict 2" in the period after that—assuming civil war does not recur. In Collier and Hoeffler's preferred (OLS) specification, aid×policy×post-conflict 1 is significantly different from 0: aid works particularly well in a good policy environment a few years after conflict has ended.

Also corroborating Burnside and Dollar, Collier and Dehn (2001) hew more closely to the Burnside and Dollar specification and data set, and tell a story that synthesizes elements from Burnside and Dollar and Guillaumont and Chauvet. They find that adding variables incorporating information on export shocks renders Burnside and Dollar's preferred specification—the one

[3] The regions are Europe and Central Asia, Middle East and North Africa, Southern Asia, East Asia and Pacific, Sub-Saharan Africa, and Latin America and the Caribbean, as defined by the World Bank.
[4] Collier and Hoeffler (2002) make a small correction to the Collier and Dollar data set, excluding 5 observations where a missing value had been treated as 0. I test the Collier and Hoeffler version of the Collier and Dollar regres-

with aid×policy but not aid^2×policy—more robust to the inclusion of Burnside and Dollar's five outliers. First, they add two variables indicating the magnitude of any positive or negative commodity export price shocks. Aid×policy is then significant at 0.01 for a regression on the full sample. The negative-shock variable is too, with the expected minus sign. Then Collier and Dehn add four aid-shock interaction terms: lagged aid×positive shock, lagged aid×negative shock, Δaid×positive shock, and Δaid×negative shock. The first and last prove positive and significant in OLS, and the last, Δaid×negative shock proves particularly robust in their testing. But including the four raised the significance level of aid×policy to 0.08.[5] Still, the study provides some buttressing for Burnside and Dollar, and suggests that well-timed aid increases ameliorate negative export shocks. This matches the Guillaumont and Chauvet result in spirit. But where Guillaumont and Chauvet interact the *amount of aid* with the *stand*ar*d deviation* of an index of export *volume* and other variables, Collier and Dehn's significant interaction term involves the *change* in aid and the *change* in export *prices*.

Most recently in the peer-reviewed literature, Dalgaard *et al.* (2004) tell a novel aid-growth story. They focus on the share of a country's area that is in the tropics as a determinant of both growth and the influence of aid on growth. This variable is indisputably free of the endogeneity worries that beset other variables advocated as determinants of aid effectiveness, such as inflation and export volume volatility. It surfaces as a growth determinant in the work of Jeffrey Sachs and others (Bloom and Sachs, 1998; Gallup and Sachs, 1999; Sachs, 2001, 2003). And the causal links between tropical location and growth may include institutions and economic policies (Acemoglu *et al.*, 2001; Easterly and Levine, 2003). Dalgaard *et al.* thus see tropical area as an

---

sion.

[5] I am able to reproduce exactly the results of the specification with the four interaction terms, in which aid×policy is marginally significant. That is the one tested below. But I am not able to reproduce their results for the variant including only positive- and negative-shock terms. In my results, aid×policy has a t statistic of only 0.42. Yet the R^2

exogenous "deep determinant" of growth. In OLS, 2SLS, Arellano-Bond "difference GMM," and Blundell-Bond "system GMM" regressions[6], aid and aid×tropical area fraction are quite significant, the first with positive sign, the second with negative sign and similar magnitude. For countries in the tropics, the derivative of growth with respect to aid (the sum of the coefficients on aid and aid×tropical area fraction) is statistically indistinguishable from 0. Thus, on average, aid seems to work outside the tropics but not in them. The authors report that their new interaction term drives out both aid×policy and aid^2.

There are other third-generation studies. Hadjimichael *et al.* (1995), Durbarry *et al.* (1998), and Lensink and White (2001), all find evidence for Hansen and Tarp's story of aid effectiveness. Svensson (1999) finds a positive interaction between aid and recipients' level of democracy. Chauvet and Guillaumont (2002) enrich their earlier analysis, in draft form. In a working paper, Burnside and Dollar (2004) defend their earlier analysis with a radically different specification, a cross-section of countries rich and poor in the 1990s with minimal controls and a different definition of policy.

In the present paper I focus on the studies already highlighted as being among the most influential and, with two exceptions, having been published in the peer-reviewed literature. The two exceptions are Collier and Dehn (2001) and Collier and Hoeffler (2002), which are pillars of the peer-reviewed Collier and Dollar (2004). From 6 of these 7 studies, I test the single regression that appears to be the authors' preferred one. From Hansen and Tarp (2001), I take two regressions. One, an Arellano-Bond GMM regression, seems to be their preferred regression. I in-

---

and sample size match theirs exactly.

[6] Blundell-Bond "system GMM" augments Arellano-Bond "difference GMM" by creating a system of equations, half in first differences, half in levels. In the difference equations, predetermined and endogenous variables are instrumented with lags of their own levels while in the levels equations they are instrumented with lags of their first differences. See Blundell and Bond (1998).

clude one of their 2SLS regressions too because it is most akin to the core specification in an important new study by Clemens *et al.* (2004).

II. The Test Suite

A. The Tests

There is some robustness testing in the recent literature on aid-growth connections, albeit focusing almost exclusively on Burnside and Dollar (2000). Lu and Ram (2001) introduce fixed effects into the Burnside and Dollar regressions. Ram (2004) splits the aid variable into the components coming from bilateral and multilateral donors, and also tests alternative definitions of policy. Dalgaard and Hansen (2001) modify the choice of excluded outliers. Easterly *et al.* (2004) extend the data set to additional countries and an additional period, 1994–97. All these tests eliminate the key Burnside and Dollar (2000) result.

The present study expands Easterly *et al.* work along two dimensions. It applies more tests. And it tests more studies. Table 1 details the regressions chosen for testing from the publications highlighted above. The regressions vary in type of estimator, control set, countries sampled, length of periods within the panel, overall study timeframe, definitions of aid and policy, and treatment of outliers. They do appear to contain "whimsy."

Table 1. Regressions tested

Regression	Estimator	Former East bloc?	Controls	Study period	Years/ period	Definition of Aid	Definition of Policy	Outliers out?	Key significant term(s)
Burnside & Dollar 5/OLS	OLS	No	LGDP, ETHNF, ASSAS, ETHNF×ASSAS, ICRGE, M2, SSA, EASIA, period dummies	1970–93	4	EDA/real GDP	BB, INFL, SACW	Yes	aid×policy
Collier & Dehn 3.4	" "	" "	" "	1974–93	" "	" "	" "	No	aid×policy, Δaid×negative shock
Collier & Dollar 1.2[1]	OLS	Yes	LGDP, ICRGE, policy, period and region dummies	1974–97	" "	ODA/real GDP	CPIA	" "	aid×policy, aid^2
Collier & Hoeffler 3.4	OLS	" "	" "	" "	" "	" "	" "	" "	aid×policy× post-conflict 1
Hansen & Tarp 1.2	2SLS	No	LGDP, ETHNF, ASSAS, ETHNF×ASSAS, ICRGE, M2, SSA, EASIA, period dummies	" "	" "	ODA/ exchange rate GDP	BB, INFL, SACW	" "	aid, aid^2
Hansen & Tarp 3.2	Difference GMM	" "	LGDP, ASSAS, ETHNF×ASSAS, ICRGE, M2, period dummies	1978–93	" "	" "	INFL, SACW	" "	aid, aid^2, Δaid, Δaid^2
Dalgaard, *et al.*	System GMM	Yes	LGDP, policy, period dummies	1970–97	" "	EDA/real GDP^2	BB, INFL, SACW	" "	aid, aid×tropical area fraction
Guillaumont & Chauvet 2.4	2SLS	No	LGDP, ENV, SYR, POPG, M2, PINSTAB, ETHNF, period dummy	1970–93	12	ODA/ exchange rate GDP	" "	" "	aid, aid×environment

[1]As revised in Collier & Hoeffler 1.1. [2]As extrapolated to 1970–74 and 1996–97 in Easterly *et al.* (2004).
Abbreviations: LGDP=log initial real GDP/capita; ETHNF=ethno-linguistic fractionalization, 1960; ASSAS=assassinations/capita; ICRGE=composite of International Country Risk Guide governance indicators; M2=M2/GDP, lagged; SSA=Sub-Saharan Africa dummy; EASIA=fast-growing East Asia dummy; ENV=Guillaumont & Chauvet "environment" variable; SYR=mean years of secondary schooling among adults; PINSTAB=average of ASSAS and revolutions/year; BB=budget balance/GDP; INFL=log(1+inflation); SACW=Sachs-Warner openness; EDA=Effective Development Assistance; ODA=Net Overseas Development Assistance.

The tests applied to these regressions constitute a systematic sampling of a larger region of specification space than has hitherto been examined in the aid-growth literature. To limit complexity and avoid whimsy, each test ideally involves changing just one aspect of the estima-

tions at a time. The tests are summarized in Table 2. The first four groups of tests, relating to the controls, the definition of aid and policy, and period length, transfer one specification's "whimsical" choices to the others. Next are the tests relating to the definition of export shock, which are motivated not by differences among studies but concern about the appropriateness of the original authors' definition. Last are tests that modify the sample by dropping outliers (copying Burnside and Dollar's "whimsical" choice) and/or expanding to new countries and periods.

Following are more detailed descriptions of the tests:

1. *Changing the control set.* In his discourse on whimsy, the specification choice that Leamer worries most about is the choice of regressors. The first set of tests is in this spirit. The studies examined use a variety of control sets, by which I mean regressors other than aid, the variables it is interacted with, and the interaction terms themselves. Burnside and Dollar, Collier and Dehn, and Hansen and Tarp control for initial log real GDP/capita; ethno-linguistic fractionalization, assassinations/capita, and the product thereof; the ICRGE governance variable; M2/GDP, lagged; being in sub-Saharan Africa or fast-growing East Asia; and fixed period effects. Dalgaard *et al.* control for about half of those—log real GDP/capita, ICRGE, and sub-Saharan Africa, East Asia, and period dummies.[7] Guillaumont and Chauvet control for a different but equally rich set of variables, namely initial log real GDP/capita, mean years of secondary schooling, population growth, ethno-linguistic fractionalization, the Barro-Lee political instability variable, and period effects. Collier and Dollar opt for a simpler set: initial log real GDP/capita, ICRGE, period effects, and—uniquely—region effects.

   These four sets of choices give rise to four robustness tests. Each test substitutes a given

[7] Dalgaard *et al.* actually use the listed control set for their OLS and 2SLS regressions. For their system GMM regression, they use only log real GDP/capita and period dummies. But the fuller set is more appropriate for testing the regressions from other papers because all but one is OLS or 2SLS, and that one—the Hansen and Tarp GMM regression—is a pure first difference regression in which the additional controls all drop out because they are constant

control set for the original one and examines the effect on the significance of key terms.

One important comment is in order, which applies to other robustness tests as well. I always use all complete observations available for developing countries (including the countries of Eastern Europe). Because different variables are available for different subsets of countries, changing the regressor set changes the regression sample. One could perform variants of the tests that are restricted to the intersections of the old and new samples in an attempt to distinguish the effects of changing sample and changing variables. But this would add to the complexity, and the approach I take aims to answer the hypothetical, "What would the results have been if the original authors had used alternative controls?" The authors almost certainly would have used all available observations.

over time. The slimmer control set causes widespread autocorrelation.

Table 2. Robustness Tests

Test	Description
Changing aid definition	
EDA/real GDP	Effective Development Assistance/real GDP, as in Burnside & Dollar, Collier & Dehn, Dalgaard *et al.*
ODA/real GDP	Net Overseas Development Assistance/real GDP, as in Collier & Dollar, Collier & Hoeffler
ODA/exchange rate GDP	Net Overseas Development Assistance/exchange rate GDP, as in Hansen & Tarp, Guillaumont & Chauvet
Changing policy definition	
INFL, BB, SACW	Inflation, budget balance, and Sachs-Warner openness, as in Burnside & Dollar, Collier & Dehn, Hansen & Tarp 2SLS
INFL, SACW	Inflation and Sachs-Warner, as in Hansen & Tarp GMM
CPIA	Country Policy and Institutional Assessment, as in Collier & Dollar, Collier & Hoeffler
Changing controls	
BD controls	Control for LGDP, ETHNF, ASSAS, ETHNF×ASSAS, ICRGE, M2, SSA, EASIA, period effects, as in Burnside & Dollar, Collier & Dehn, Hansen & Tarp
CD controls	Control for LGDP, ICRGE, period and region effects, as in Collier & Dollar, Collier & Hoeffler
GC controls	Control for LGDP, ENV, SYR, POPG, M2, PINSTAB, ETHNF, period effects, as in Guillaumont & Chauvet
DHT controls	Control for LGDP, ICRGE, SSA, EASIA, period effects, as in Dalgaard *et al.*
Changing period length	
12-year	Aggregate over 12-year periods, as in Guillaumont & Chauvet
Changing shock definition	
Pooled distribution	Use one threshold for all countries, rather than country-specific thresholds, that the unpredicted component of commodity export price index change must exceed to be a shock
Shock/GDP	Express shock magnitude as share of GDP rather than price change
Shock/GDP, pooled distribution	Combine above two changes
Changing sample and data set	
No outliers	Remove Hadi outliers in the partial scatter of the dependent variable and the independent variable of greatest interest
Expanded sample	New data set. Carried to 2001, except shocks data end in 1997 and Guillaumont & Chauvet environment variable not updated
Expanded sample, no outliers	Combine above two changes
Expanded sample, AR-robust	Use new data; use 5-year periods to eliminate higher-order autocorrelation; instrument almost all variables with one-period lags; cluster standard errors by country
Expanded sample, AR-robust, no outliers	Same as previous but removing Hadi outliers

Abbreviations as in Table 1.

2. *Redefining aid.* All the studies take total aid received as a share of recipient GDP. But there are differences in defining both the numerator and denominator of the ratio. Burnside and Dollar, Collier and Dehn, and Dalgaard *et al.* use Effective Development Assistance in the numerator while the rest use net ODA. On the choice of denominator, there is also a split. Hansen and Tarp and Guillaumont and Chauvet use GDP converted to dollars using market exchange rates. The others use real GDP from the Penn World Tables. A country's relative price level strongly correlates with income per head, with the poorest countries having a price levels 20–25% that of the United States. Thus, using purchasing power parities instead of exchange rates will cause the GDPs of the poorest countries to be measured as relatively larger and aid to them as relatively smaller as a share of GDP. It is hard to know a priori what effect this change could have on regression coefficients for aid and its interactions, but it could be significant.

   The use of exchange rates seems more appropriate. One way to see this is to ask what the local-currency cost is to the recipient of *not* receiving a quantum of aid. If the aid would be spent entirely on tradables, which cost about the same everywhere, then the cost to Sri Lanka of not receiving $1 million in aid is the exchange-rate equivalent in rupees. The opposite extreme would occur if the aid would be spent on the representative consumption basket for the country, including tradables and nontradables, *with all goods and services purchased at donor-economy prices*. In this case, the recipient government could purchase the same goods and services for far less, so that the value of aid would be greatly inflated going by exchange rates, and the PPP adjustment would be appropriate. This case, however, seems quite unlikely. By definition, donors must purchase nontradables locally. They may pay above-market, but probably do not pay close to the donor-economy prices that can be 400% higher

or more.[8] In fact, studies of the "tying" of aid—which is when donors require that it be spent into the their own economy, on their own tractors or consultants—put the cost increase at only 15–30%, not 400% or more (Jepma, 1991). This argues for exchange rates.

At any rate, with two options each for measuring aid and GDP, there are four possible combinations for aid/GDP. The literature includes three of them (all but EDA/exchange rate GDP), and these are the bases for three tests.[9] In fact, EDA/real GDP and ODA/real GDP are highly correlated (Dalgaard and Hansen, 2001), so switching from one to the other may not stress results much. Switching between real and nominal GDP can be expected to impose a more difficult test, because the correlations are lower. (See Table 3.)

Table 3. Simple correlations of aid measures, four-year periods, all available observations

	EDA/real GDP	ODA/real GDP	ODA/ exchange rate GDP
EDA/real GDP	1.00		
ODA/real GDP	0.97	1.00	
ODA/exchange rate GDP	0.78	0.82	1.00

3. *Redefining good policy.* Three sets of "good policy" variables appear among the tested regressions: 1) Burnside and Dollar's famous combination of budget balance, inflation, and Sachs-Warner openness; 2) inflation and Sachs-Warner only (Hansen and Tarp GMM); and 3) CPIA alone (Collier and Dollar, Collier and Hoeffler). These generate three robustness tests. Using Burnside and Dollar's coefficients to form policy indexes (6.85 for budget balance, –1.40 for inflation, and 2.16 for Sachs-Warner), I find that the first two policy definitions are highly correlated, but the third varies more distinctly. (See Table 4.) Switching between the first two definitions is probably a mild test, but switching between them and CPIA a more severe one. To apply the tests, in each case I rerun the Burnside-Dollar–style index-

[8] If the overall price level in a donor country is five times that of the recipient, but the price level for tradeables is about the same in both countries, then the ratio for nontradables must be well above five.
[9] The published EDA data (Chang *et al.*, 1998) cover only 1975–95. I extrapolate EDA to the rest of 1970–2001 via

forming regression, which includes all regressors except aid and its interaction terms, then use the coefficients on the policy variables to make the index.[10] I include all candidate policy variables in the index regardless of their significance in this regression.

Table 4. Simple correlations of "good policy" measures

	Inflation, budget balance, Sachs-Warner	Inflation, Sachs-Warner	CPIA
Inflation, budget balance, Sachs-Warner	1.00		
Inflation, Sachs-Warner	0.98	1.00	
CPIA	0.53	0.52	1.00

4. *Changing periodization.* All but one of the studies use four-year periods, the exception being Guillaumont and Chauvet's, which uses 12 years. Since I lack higher-frequency observations of the Guillaumont and Chauvet environment variable, I cannot test their regression on a 4-year-period panel. But I test all the other regressions on 12-year panels. Notably, key studies in the broader growth literature use periods of 10–25 years despite the small samples that result[11] (Barro, 1991; Mankiw *et al.,* 1992; Sachs and Warner, 1995).

---

a regression of EDA on net ODA, which is available for the whole period.

[10] I compute the constant term in the policy index in the same manner as BD. It is the predicted growth rate in the model when the policy variables and the period dummies are zero, and all other variables take their sample-average values.

[11] The question of whether the effects measured in four-year regressions are "short-term" is not straightforward to answer. On the one hand, very little of the aid in a given *P*-year period is disbursed right at the start of the period and very little of the growth record occurs right at the end. In other words, assuming that aid disbursements are spread evenly over time within the period, the average *positive* lag between two point in time within the period—the first when aid is disbursed the second when a quantum of growth experience occurs—works out to be:

$$\frac{\int_0^P \int_t^P (s-t)\,ds\,dt}{\int_0^P \int_t^P ds\,dt} = \frac{P}{3}.$$

For *P*=4 years, this is 16 months. On the other hand, aid is significantly autocorrelated. I find that the three aid variables used in this literature have a serial correlation of 0.80–0.85 with 4-year periods. Thus, information about current-period aid flows conveys a good deal about previous-period aid. Among the tested regressions, only the Hansen and Tarp GMM regression controls for previous-period aid, so the coefficients in all the other regressions actually

5. *Changing the definition of shock.* This test applies to the Collier and Dehn regression only. Their positive and negative shock variables are built from a data set of commodity prices for 113 developing countries (Dehn, 2000). The definition has two parts: a rule for determining *whether* a change in a country's aggregate commodity export price index is large and unexpected enough to constitute a shock, and a formula for measuring the *size* of the shock. To determine which price movements are shocks, Collier and Dehn examine the residuals from a regression based on the following forecasting model:

$$\Delta y_{it} = \alpha_0 + \alpha_1 t + \beta_1 \Delta y_{i,t-1} + \beta_2 y_{i,t-2} + \varepsilon_{it}$$

where y_{it} is the commodity export price index for country i in time t and ε_{it} is the forecasting error. The regression is run separately for each country over 1957–97, and those forecasting errors at least 1.96 standard deviations from the mean are considered shocks. The *magnitude* of a shock is then the full price change Δy_{it}.

Two aspects of this definition deserve comment. First, the threshold that the unpredicted component of a price movement must exceed to be a "shock" varies by country. This arguably makes sense in that an "unexpected" 20% price change might be routine in one country and once-a-century in another. Countries that experience large price changes more frequently may adapt to them, so that the effects on growth and poverty would be systematically different. On the other hand, this leads to the odd result that countries that are in fact more shock-prone (in lay terms) do not necessarily experience more "shocks." To take two extreme ex-

---

reflect effects of past aid too. One can estimate the true average aid-growth lag in these regressions as a weighted average involving not just $P/3$ but terms such as P, $2P$, $3P$…, with weights ρ, ρ^2, ρ^3…, where ρ is the autocorrelation coefficient. It can be shown that the average lag is then:

$$\frac{\frac{P}{3} + P\rho + 2P\rho^2 + 3P\rho^3 + \ldots}{1 + \rho + \rho^2 + \rho^3 + \ldots} = \frac{P}{3}\frac{1+\rho+\rho^2}{1-\rho}.$$

For P=4 and ρ=0.81–0.85, this equals 17.3–22.9 years.

amples, in South Africa, where the standard deviation of the forecasting error was 4.8% during 1957–97, the 11.0% price index increase in 1988 (with a 13.2% forecasting error) is considered a positive shock. But in Laos, where the standard deviation of forecasting error was 24.9%, the 32.7% percent drop in 1987 (forecasting error of –48.1%) is not considered a shock.

Second, the magnitude of the shocks as defined by Collier and Dehn does not take into account the economic significance of commodities exports for a country. The growth impact of a commodities price shock is probably related to both its magnitude and the share of commodity exports in GDP. But the Collier-Dehn definition is purely a measure of price change. In contrast, Easterly and Rebello (1993), for example, gauge terms-of-trade shocks relative to GDP.

I test the Collier and Dehn regressions with two changes that modify these aspects of the shock definition—first separately, then together, for a total of three robustness checks. To address the concern about the country dependence of shock definition, I pool the forecasting errors for all 117 countries in the Dehn (2000) data set, and choose a single 1.96-standard-error shock threshold, which turns out to be a forecasting error of ±29.5%. To address the concern about economic significance, I express the price changes as shares of GDP, using 1990 data on commodities exports, total exports, and total exports/GDP from Dehn (2000).[12]

6. *Removing outliers.* The Burnside and Dollar specification I test excludes five observations that are a) outliers in aid×policy.and b) highly influential with respect to the coefficient on that term. This raises a general question about the extent to which significant results in this

[12] This implicitly assumes that the share of exports in GDP remained fixed at 1990 levels throughout the study period. This assumption is not completely realistic, but it allows use of Dehn's data, with its broad coverage, and re-

literature are driven by outliers. To investigate the role of outliers, I rerun the reproductions of the original regressions after excluding outliers. I do the same for the expanded-sample versions of the regressions. (See below.) The process for picking outliers is different from Burnside and Dollar's, but less subjective. I apply the Hadi (1992) procedure for identifying multiple outliers to the partial scatter of growth with a regressor of particular interest, such as aid×policy, using 0.05 as the cut-off significance level.[13] In instrumented regressions, I use the variable of interest after projection onto the instruments.

I even run this test on the Burnside and Dollar 5/OLS regressions, from which one set of outliers is already excluded. Fundamentally, regardless of the genesis of these regressions' results, it is still interesting to determine whether they are driven by a few observations in the remaining sample.

Outliers are not synonymous with *influential* observations. Why then focus on outliers? For one, even outliers that do not have a high influence on coefficients of interest—what standard tests for influence measure—can substantially affect the *standard errors* of the coefficients, which is at least as important for the present study. In addition, outliers are the observations most likely to signal measurement problems or structural breaks beyond which the core model does not hold—both of which seem better reasons than high influence for exclusion. That said, outliers do not necessarily signal measurement problems or structural break. This is especially possible when the variable of interest is highly non-normal, such as the Collier and Dehn export price shock variable, which is usually 0 but occasionally takes large values. In such cases, outliers may contain valuable information about the development process under rare combinations of circumstances. Still, on balance, regression results driven by a

---

duces any endogeneity of the modified shock variable to growth.

[13] Applying the Hadi procedure directly to a full, many-dimensioned regression data set typically identified 20% or

few outliers need to be interpreted with care.

The method I use for identifying outliers does not extend straightforwardly to GMM because the two-dimensional partial scatterplot is a construct without a perfect analogue in the GMM setting. Moreover, in system GMM, with parallel equations in levels and differences, the notion of an observation itself becomes fuzzy. For these reasons, I do not attempt to identify and remove outliers from difference and system GMM regressions. Ideally, the combination of extensive instrumentation and GLS-style reweighting trims and deemphasizes outliers.

7. *Expanding the sample.* Easterly *et al.* (2004) develop a new dataset that extends that of Burnside and Dollar from 1970–93 to 1970–97 and adds six countries. For the present study, that data set has been extended to 2001, pushed back to 1950, expanded to include variables in other studies and the literature, and given better coverage of post-Communist Eastern Europe. (See Appendix.) I use pre-1970 data only for lagged regressors and instruments, or for computing first differences for 1970–73. This data set allows a net expansion in both years and countries for all but Guillaumont and Chauvet regression, whose 12-year periods and unusual environment variable make extension difficult. (See Table 5.) For example, in the case of the Hansen and Tarp 2SLS regression, although 16 of the 231 observations complete in the original data set are incomplete in the new one, the new set has 194 novel observations, for a total of 409. More sample size information is in Table 14.

---

more of observations as outliers.

Table 5. Overview of differences in regression samples, original and expanded data sets

Regression	Countries unique to the regression sample: Original set	Countries unique to the regression sample: Expanded set	Study period: Original	Study period: Expanded
Burnside & Dollar	Somalia, Tanzania	Burkina Faso, Bulgaria, China, Cyprus, Hungary, Iran, Jordan, Myanmar, Papua New Guinea, Poland, Republic of Congo, Romania, Singapore, South Africa, Uganda	1970–93	1970–2001
Collier & Dehn	" "	" "	1974–93	1974–97
Collier & Dollar, Collier & Hoeffler	None	Angola, Burkina Faso, Bulgaria, China, Cyprus, Czech Republic, Guinea, Guinea-Bissau, Iran, Jordan, Liberia, Mongolia, Mozambique, Myanmar, Namibia, Oman, Papua New Guinea, Poland, Republic of Congo, Romania, Russia, Somalia, Suriname, Uganda	1974–97	1974–2001
Hansen & Tarp 2SLS	Guyana, Somalia, Tanzania	Burkina Faso, Bulgaria, China, Côte d'Ivoire, Cyprus, Hungary, Iran, Jordan, Mali, Myanmar, Papua New Guinea, Poland, Republic of Congo, Romania, Singapore, South Africa, Uganda	1974–93	1970–2001
Hansen & Tarp GMM	Somalia	Angola, Burundi, Benin, Burkina Faso, Barbados, Bulgaria, Central African Republic, Chad, China, Cyprus, Hungary, Iran, Jordan, Mauritania, Mauritius, Mozambique, Myanmar, Nepal, Papua New Guinea, Poland, Republic of Congo, Romania, Rwanda, Singapore, South Africa, Uganda	1978–93	" "
Dalgaard *et al.*	None	Bangladesh, Czech Republic, Guinea-Bissau, Russia, Singapore, South Africa	1974–97	" "
Guillaumont & Chauvet	None	None	1970–93	1970–93

B. Testing the validity of the modified regressions

I apply up to three specification tests to the modified regressions. The first, which applies to all the regressions, is for autocorrelation in the residuals. Autocorrelation can bias coefficients as well as standard errors. In the OLS and 2SLS regressions I test, many of the regressors that are treated as exogenous—inflation, governance quality, etc.—are potentially affected by past growth (predetermined), which would render them endogenous in the presence of autocorrelation and bias results. In the GMM regressions, few regressors are treated as exogenous; they are instrumented instead. But the instruments are lags of regressors in levels or first differences, which can themselves be rendered invalid by autocorrelation of certain orders. To check for autocorrelation, I use the Arellano-Bond (1991) test for AR(), which is a flexible test designed for linear

GMM regression on panel data with arbitrary patterns of autocorrelation and heteroskedasticity.[14] Since OLS and 2SLS with robust standard errors are special cases of linear GMM, the Arellano-Bond test is appropriate for all the regressions I run. For the GMM regressions, I follow the usual practice of applying the test to first difference residuals, which can detect autocorrelation in the error component that is purged of fixed country effects (Arellano and Bond, 1991). I use a significance threshold of 0.05.

The second specification test, which applies to all the instrumented regressions, is the Hansen *J* test of overidentifying restrictions—also using a significance level of 0.05. Failure on this test indicates that an excluded instrument may in fact be a significant growth determinant and that its exclusion may be causing omitted variable bias. Unlike the Sargan test of overidentifying restrictions, the Hansen test is consistent in the presence of heteroskedasticity and autocorrelation.

The third specification "test" is a check on whether the number of instruments in instrumented regressions is large. As the instruments become numerous relative to the sample size, they can overfit the instrumented variables, biasing the results toward those of OLS (in the case of 2SLS) or feasible GLS (in the case of GMM). In the extreme case, where instruments match or outnumber observations, the results will match those of OLS or FGLS. However, the literature provides little guidance on how many instruments is too many (Ruud, 2000 (p. 515)). Proliferation of collinear or nearly collinear instruments can also cause the two-step estimate of the covariance of the moments, which is the basis for the optimal twos-step GMM weighting matrix, to be singular.[15]

[14] I apply the test to OLS and 2SLS regressions using my "abar" module for Stata.
[15] The two-step weighting matrix is $(\mathbf{Z'HZ})^{-1}$, where $\mathbf{Z}$ is the instrument matrix and $\mathbf{H}$ is a robust proxy for covariance of the errors. Normally in Arellano-Bond/Blundell-Bond estimation rank($\mathbf{H}$)=N, the number of individuals in the panel. This is because $\mathbf{H}$ is block diagonal, with one block for each individual. Each block is an outer product

Indeed, this appears to have happened in the original Dalgaard *et al.* regression, and this "test" is only relevant for that regression. In Arellano-Bond and Blundell-Bond GMM regressions, one "GMM-style" instrument is generated for each instrument variable, time period, and available lag (Holtz-Eakin *et al.*, 1988; Arellano and Bond, 1991). Instruments can easily number in the hundreds. Where the Hansen and Tarp GMM specification covers 4 time periods and uses at most 1 or 2 lags of 6 instrumenting variables, the newer Dalgaard *et al.* one covers 6 periods, imposes no lag limits, and uses 12 instrumenting variables, leading to 303 instruments for 371 observations in my reproduction.[16]

In performing the "test" of excessive instruments, the rule of thumb I use is that the number of instruments should not exceed the number of countries in the regression.[17] Typically there is an average of 6 observations per country in the variants of the Dalgaard *et al.* regression that I run, so this rule is equivalent to requiring that the ratio of instruments to observations not exceed about 0.17.

If a modified regression fails one of these specification tests, the question then is what to do about it. One possible response to an invalid specification is to simply not report the results. Another is to report the results but document that they come from an invalid specification. A third is to make judicious additional changes to the failing specification so that it passes. I sometimes take the second approach, sometimes the third. On the one hand, searching for additional modifications that can allow the specifications to pass tests can rapidly raise the complexity of the testing and open the door to whimsy. On the other hand, minimal changes sometimes can

---

$\mathbf{e}_i\mathbf{e}_i'$, where $\mathbf{e}_i$ is the column vector of one-step residuals for individual *i*. These blocks have rank 1. As the number of columns in **Z** approaches rank(**H**), the risk of singularity of **Z'HZ** rises.

[16] In their own paper, the authors note this problem. But they report that limiting the number of lags does not affect their results significantly and prefer to report the results from the unlimited regression as simpler, apparently not noticing the singular matrix problem.

[17] I thank Henrik Hansen for this rule of thumb.

make invalid tests much more meaningful. For example, it useful and relatively un-whimsical to limit the number of instruments in variants of the Dalgaard *et al.* regression. And autocorrelation of various orders turns out to be endemic in the results from the sample-expansion test; some effort at overcoming it makes the results from this important test more useful.

To be precise, I make three additional changes in order to reduce the number of modified regressions that fail specification tests. First, I add the log of initial population as regressor in all robustness tests, except for the test that simply drops outliers from the original regression. All the 2SLS regressions I test use this variable as an instrument for aid because of the well-known bias among donors toward small countries. But perhaps because of the economic performance of China and India since the early 1990s the variable has picked up significance for growth too, especially under the sample-expansion test, which extends most of the regression samples to 2001.[18] This change not only allows many of the modified 2SLS regressions to pass the Hansen *J* test, it also eliminates some occurrences of AR(1).

Second, I perform a version of the sample-expansion test that incorporates several changes to remove or accommodate autocorrelation. I call this the "AR-robust" sample-expansion test. Autocorrelation of up to order 4 occurs in many of the expanded-sample regressions. For example, the Arellano-Bond test results for AR() of various orders in the Burnside and Dollar regression on the enlarged sample are:

Order	*z*	*p*
1	2.64	0.008
2	1.65	0.098
3	–1.77	0.076
4	2.32	0.021

where *z* is the (normally distributed) test statistic and *p* is the corresponding two-tailed probabil-

[18] Except that the data set is only extended to 1997 or the Colliar and Dehn regression, for lack of the shock variable after 1997, and not extended at all for the Guillaumont and Chauvet regression, for lack of their environment variable after 1993.

ity. The source of this autocorrelation is not obvious. It may be a statistical artifact. Indeed, the first change I make to deal with autocorrelation in the sample-expansion test, switching from four- to five-year periods, eliminates all the autocorrelation of concern, except for AR(1) in the OLS and 2SLS regressions. Having restricted the problem to AR(1) in OLS and 2SLS, I then instrument almost all right-hand-side variables in these regressions with their one-period lags in order to make the regressions consistent even if these variables are predetermined. I exempt only the Collier and Dehn price shock variables, the Collier and Hoeffler post-conflict 1 variable, and interaction terms involving them. The shock and post-conflict variables are quite stochastic and hard to instrument, and at least the price shock variable should not have a large predetermined component. Finally, I report standard errors that are "clustered" on the country identifier, making them robust to autocorrelation.[19]

The third change I make to pass specification tests is to limit the number of instruments in the Dalgaard *et al.* regression. In standard difference and system GMM, sets of "GMM-style" instruments (Holtz-Eakin *et al.*, 1988; Arellano and Bond, 1991) embody the expectations:

$$\sum_i w_{i,t-l}\Delta e_{it} = 0 \text{ for each } t \text{ and } l, L_1 \le l \le L_2$$

where w is an instrumenting variable, e, is the residual vector, i is the panel index, t the time index, and L_1 and L_2 define the range over which lags can be taken. (Typically L_1=0 or 1.) Thus there is one column in the instrument matrix for each w, t, and l. In testing the Dalgaard *et al.* regression, I find that even setting L_2=1 allows too many instruments. I therefore group together columns of the instrument matrix that are for the same w and lag distance and combine them by addition in order to generate a smaller set of moment conditions, which work out to imply the expectations:

[19] An alternative approach would be to increase the period length until all autocorrelation disappears—in practice

$$\sum_{i,t} w_{i,t-l} \Delta e_{it} = 0 \text{ for each } l,\ L_1 \le l \le L_2.$$

I find that when instruments are combined in this way, L_2 can take a value of up to 3 without violating the rule of thumb. So in all tests of the Dalgaard *et al.* regression, I "collapse" the instruments and set L_2=3.[20]

C. Theoretical issues in interpreting the results

If Leamer's (1983) extreme bounds analysis is applied to the results of this testing, then a coefficient will be deemed robustly different from 0 only if it is significantly different from 0 in *every* test. However, as Sali-I-Martin (1997) argues, this definition of robustness may be too extreme. For example, I could test the robustness of a regression by averaging together all observations for each global region, generating a sample of some 6 observations. Presumably almost no regression would pass this test. One could argue that this test is "unfair," which is to say, too demanding to generate meaningful results. But there is no bright line dividing fair tests from unfair ones. Indeed, in this test suite, the 12-year-period test destroys every regression it can be applied to. It is not obvious *a priori* whether this means the test is too strong or the regressions too weak. For this reason, as Sali-I-Martin suggests, robustness should be thought of a continuous rather than dichotomous concept. Results can be more or less robust.

Sala-I-Martin offers his own procedure for assessing robustness. In essence, he estimates the cumulative distribution function for a coefficient of interest by running a large number of variants of the regression it comes from. The *robustness* of a coefficient is then the fraction of the cumulative distribution that is on one or the other side of zero. For example, a coefficient would be robustly different from 0 at the conventional level of significance if 95% of its cdf is above or below zero.

---

about eight years. But this would destroy more information than the approach I use.

The validity of this inference is based on the assumption, however informal, that the set of regressions actually run is a *representative sample of all possible variants of the original regression*. For the collection of tests assembled here, that assumption does not seem valid. Consider the set of tests that involves varying how total aid receipts are measured. As described above, two of the three aid definitions are highly correlated in practice. So regressions that originally used the third definition are subjected to two nearly identical aid-related tests. Sala-i-Martin's algorithm would treat these two tests as distinct, giving them nearly equal weight, rather than as what is essentially an unrepresentative oversampling of one test. Similarly, one subset of tests, those expanding the sample, cannot be applied to the Guillaumont and Chauvet regression. It does not seem plausible that the test suite is representative both with and without this important subset of tests.

In sum, the sampling of specification space that is made here is *minimally arbitrary*, but cannot be assumed to be *representative* of all possible tests. Thus while Leamer's definition of robustness may be to harsh for this context, Sali-I-Martin's has its own limitations. This would be true even if I performed every possible combination of tests in the suite rather than just one at a time. In the end, it seems that human judgment applied to the full set of results must substitute for mechanical definitions of robustness. This in turn argues for keeping the number of tests small enough for the human mind to embrace.

III. Results

In the first step of the testing process, I use the authors' data sets to reproduce their original results. (See Tables 6–10. The dependent variable in all the regressions is average annual real GDP/capita growth.) All of the reproductions exhibit the same pattern of results as the original

[20] I use the "collapse" sub-option of my xtabond2 module for Stata.

and all but one have the same sample size.[21] The Burnside and Dollar, Collier and Dehn, and Hansen and Tarp reproductions are perfect. Since the purpose of the paper is to test robustness, the inexact matches are not a major problem. If the results from the tested regressions are robust, they should withstand whatever minor changes in data or specification cause the discrepancies in my reproductions.

I discovered two important specification issues in the original regressions. One was a multicolinearity problem in the Collier and Hoeffler regression. In their preliminary regression 3.1 (not regression 3.4, which is reproduced here), they include the variables post-conflict 1, post-conflict 1×policy, and post-conflict 1×aid^2, along with post-conflict 1×aid ×policy. These four terms involving post-conflict 1 all have absolute *t* statistics of 0.5 or less. Collier and Hoeffler then search for a specification iteratively, by dropping the least significant of these terms one-by-one and rerunning the regression. But in my reproduction of 3.1, post-conflict 1×aid×policy has a partial correlation of 0.985 with post-conflict 1×aid, making the two statistically indistinguishable. Thus the Collier and Hoeffler results ought to be interpreted as pertaining to *either* post-conflict 1×aid ×policy *or* post-conflict 1×aid. Occam's razor argues for the latter.

The second has already been mentioned. In the Dalgaard *et al.* regression, the two-step estimated covariance matrix of the moments turns out to be singular.[22] I avoid this problem in my reproduction by reducing the number of instruments exactly as described above. My reproduction has their key terms still quite significant, and with larger magnitudes than in the original.

[21] The Dalgaard *et al.* regression was executed with the DPD for Ox package. It turns out that an undocumented limitation in this software—incomplete observations that create gaps in the time series must always be included in the data file rather than deleted—led to a slight mishandling of the data. I use my xtabond2 module for Stata, which does not have this limitation. This explains the difference in samples.
[22] The DPD for Ox package computes a generalized inverse in such cases.

Table 6. Reproduction of Burnside and Dollar, Collier and Dehn results

Regression	Burnside & Dollar 5/OLS	Collier & Dehn 3.4 (OLS)
Log initial real GDP/capita	–0.60	–0.77
	(–1.02)	(–1.32)
Ethno-linguistic fractionalization, 1960	–0.42	–0.35
	(–0.57)	(–0.45)
Assassinations/capita	–0.45	–0.37
	(–1.68)	(–1.27)
Ethno-linguistic fractionalization × Assassinations/capita	0.79	0.63
	(1.74)	(1.31)
Sub-Saharan Africa	**–1.87**	**–2.08**
	(–2.41)	**(–2.83)**
Fast-growing East Asia	**1.31**	1.21
	(2.19)	(1.90)
Institutional quality (ICRGE)	**0.69**	**0.67**
	(3.90)	**(3.57)**
M2/GDP, lagged	0.01	0.02
	(0.84)	(1.12)
Policy	**0.71**	**0.86**
	(3.63)	**(4.23)**
Aid/GDP	–0.02	–0.16
	(–0.13)	(–1.23)
Aid/GDP×policy	0.19	0.10
	(2.61)	(1.70)
Positive shock		0.00
		(0.18)
Negative shock		**–0.03**
		(–2.38)
ΔAid/GDP ×negative shock		**0.04**
		(3.17)
ΔAid/GDP ×positive shock		0.00
		(–0.01)
Aid/GDP, lagged×negative shock		0.01
		(1.23)
Aid/GDP, lagged×positive shock		**0.02**
		(2.40)
Observations	270	234
R^2	0.39	0.46
Arellano-Bond AR(1) test (*p* value)	0.54	0.88

Period dummies and constant term not reported. Heteroskedasticity-robust *t* statistics in parenthesis. Entries significant at 0.05 in bold.

Table 7. Reproduction of Collier and Dollar, Collier and Hoeffler results

	Collier & Dollar 1.2[1]	Collier & Hoeffler 3.4
Log initial real GDP/capita	0.70	0.69
	1.13	1.11
Institutional quality (ICRGE)	0.12	0.14
	0.73	0.93
CPIA	**1.16**	**1.15**
	2.89	**2.88**
Aid/GDP×policy	**0.14**	0.12
	2.15	1.88
$(Aid/GDP)^2$	**–0.03**	**–0.03**
	2.20	**2.20**
Aid×policy×post-conflict 1		**0.18**
		3.92
Observations	344	344
R^2	0.37	0.38
Arellano-Bond AR(1) test (*p* value)	0.63	0.61

Period and region dummies and constant term not reported. Heteroskedasticity-robust *t* statistics in parenthesis. Entries significant at 0.05 in bold. Reproduction is not exact, but exhibits the same pattern and matches in sample size.

[1]As revised in Collier and Hoeffler (2002), regression 1.1, where five erroneous observations were deleted.

Table 8. Reproduction of Hansen and Tarp results

Regression	1.2 (2SLS)	3.2 (GMM)
Average annual GDP/capita growth, lagged		**–0.37**
		(–7.09)
Log initial real GDP/capita	0.09	–3.56
	(0.14)	(–1.05)
Ethno-linguistic fractionalization, 1960	0.11	
	(0.12)	
Assassinations/capita	**–0.46**	**–0.53**
	(–2.02)	**(–2.31)**
Ethno-linguistic fractionalization × Assassinations/capita	**0.92**	**1.00**
	(2.17)	**(2.75)**
Sub-Saharan Africa	**–2.25**	
	(–2.98)	
Fast-growing E. Asia	**1.52**	
	(2.42)	
Institutional quality (ICRGE)	**0.81**	
	(4.57)	
M2/GDP, lagged	0.01	
	(0.55)	
Budget surplus/GDP	**9.12**	
	(2.49)	
Log (1 + inflation)	**–1.13**	–0.19
	(–2.30)	(–0.30)
Sachs-Warner	**1.70**	**2.83**
	(3.36)	**(4.37)**
Aid/GDP	**0.24**	**0.90**
	(2.34)	**(4.22)**
$(\text{Aid/GDP})^2$	**–0.01**	**–0.02**
	(–2.38)	**(–3.83)**
$\Delta(\text{Aid/GDP})$		**–0.70**
		(–4.91)
$\Delta(\text{Aid/GDP})^2$		**0.01**
		(3.64)
Observations	231	213
R^2	0.38	
Arellano-Bond test for AR(1) (*p* value)	0.76	0.12
Hansen J (*p* value)	0.81	0.39

Period dummies and constant term not reported. Robust *t* statistics in parenthesis. For GMM, *t* statistics are autocorrelation-robust too, and Arellano-Bond test is for AR(2) in first differences. Some coefficients differ from original by a factor of 100 because of scaling changes. Entries significant at 0.05 in bold.

Table 9. Reproduction of Dalgaard *et al.* results

Log initial real GDP/capita	2.03
	(1.67)
Budget surplus/GDP	0.09
	(0.55)
Log (1 + inflation)	**–2.98**
	(–2.88)
Sachs-Warner	1.57
	(1.80)
Aid/GDP	**1.47**
	(4.02)
Aid/GDP × tropical area fraction	**–1.33**
	(–2.62)
Observations	371
Arellano-Bond test for AR(2) in first differences (*p* value)	0.37
Hansen J (*p* value)	0.21

Period dummy and constant term not reported. Unlike in original, to limit the number of instruments, at most 3 lags of instrumenting variables are used, and sets of instruments are combined by addition, as described in text. *t* statistics (in parentheses) are heteroskedasticity- and autocorrelation-robust and include the Windmeijer (2000) finite-sample correction. Entries significant at 0.05 in bold.

Table 10. Reproduction of Guillaumont and Chauvet 2SLS results

Log initial real GDP/capita	**–2.51**
	(–3.11)
Mean years secondary schooling among those over 25	0.93
	(1.16)
Population growth	**–0.83**
	(–2.64)
M2/GDP, lagged	**0.07**
	(2.64)
Barro-Lee political instability, lagged	–2.03
	(–1.08)
Ethno-linguistic fractionalization, 1960	**–2.13**
	(–2.04)
Environment/low vulnerability	0.53
	(1.91)
Aid/GDP	0.78
	(1.47)
Aid/GDP×environment	–0.15
	(–1.68)
Observations	68
R^2	0.63
Arellano-Bond test for AR(1) (*p* value)	0.08
Hansen J (*p* value)	0.98

Period dummy and constant term not reported. Heteroskedasticity-robust *t* statistics in parenthesis. Entries significant at 0.05 in bold.

I run 86 robustness checks in all. Full results are at <www.cgdev.org>. Here, I report full details only for the Burnside and Dollar regressions, as an illustration, and results on key coefficients for the rest.

Table 11 shows the full Burnside and Dollar test results. Note first that the testing is not nihilistic: some regressors survive all or most of the tests. Dummies for Sub-Saharan Africa and fast-growing East Asia, the governance and policy indexes, and the newly added population variable are usually significant at or near the 0.05 level, frequently enough that these do appear to be true regularities in the data. But aid×policy, the essential term for Burnside and Dollar, is more fragile. Switching to the stripped-down control set of Collier and Dollar raises the significance level of aid×policy above 0.05. Adopting the Guillaumont and Chauvet controls depresses it further, albeit with AR(1). Only the Dalgaard *et al.* control set, not so different from Burnside and Dollar's original, leaves the key term significant. Changing the definition of aid does reduce the coefficient and, though not dramatically, its significance. Switching to the Collier and Dollar definition of policy as CPIA completely eliminates the result.[23] Going to 12-year periods also erodes the *t* statistic.

In a counterpoint to the focus of Leamer (1983), Levine and Renelt (1992), and Sali-I-Martin (1997) on the choice of regressors as a source of fragility, it appears that modifying the sample affects results much more than modifying the regressor set. Strikingly, removing seven additional outliers (Botswana 1978–81, 1982–85, and 1986–89, Gabon 1974–77, Mali 1986–89, and Zambia 1986–89 and 1990–93) completely eliminates the result, sending the coefficient on aid×policy slightly negative. Expanding the dataset to 2001 and to new countries also gives aid×policy a negative but insignificant coefficient. (Easterly *et al.* (2004) get the same result when extending to 1997.) Excluding outliers from the expanded-sample regression actually

strengthens this contrary result, so that result is not itself driven by a few outliers. Unfortunately, the expanded-sample regressions exhibit autocorrelation of order up to 4. So the final two columns show the results after switching to five-year periods to expunge the higher-order autocorrelation, and after instrumenting all variables with their lags and clustering the standard errors by country to make the results robust to autocorrelation. Aid×policy remains insignificant.

The results from the original regression and the "AR-robust," expanded sample regression are illustrated in Figure 1, with and without the outliers picked by the Hadi procedure. Each graph in the figure shows the partial scatter plot of GDP/capita growth versus aid×policy in the 5/OLS regression. The two digits in the data point labels indicate the first year of the period for each observation. Outliers are marked separately, and two partial regression lines are shown, one for the full sample, one for the sample excluding outliers. Note that the second line is not the best fit to the non-outlier data points as plotted. Deleting observations causes the estimated coefficients to shift and all remaining data points in the partial scatter to move. The second line therefore is the best fit to the data points in their post-exclusion positions, which are not shown.

Figure 1 shows that in Burnside and Dollar's original regression, Botswana had very high aid×policy and fairly high growth throughout the 1980s, controlling for the other regressors, while Zambia was opposite on both counts in the late 1980s and early 1990s. Their experiences seem to drive the original 5/OLS result on aid×policy. Moreover it seems quite possible that the Botswana observations are a case where reverse causality plays a role, with high growth (and perhaps good policies) attracting more aid. When aid and aid×policy are instrumented they cease to be outliers. (Results not shown.)

Table 12, Table 13, and Table 14 report results on key terms in all the tested regressions. Because not all tests were applicable to all regressions, some cells are blank. The test involving

[23] Ram (2003) performs a similar test.

the definition of aid as EDA/real GDP, for example, is not applicable to regressions that originally use it. Using 12-year periods does not work for the Collier and Hoeffler regression, because the definition of their post-conflict 1 variable assumes 4-year periods. Some regressions do not have a policy variable and so were exempt from policy-related tests.[24] Lack of higher-frequency data for Guillaumont and Chauvet's environment variable prevents short-period tests.

Table 12 has results for most of the tests inspired by "whimsical" differences among the original regressions. It turns out that the first test, switching to the Burnside and Dollar control set, generally reinforces regressions that did not use it in the first place. The Dalgaard *et al.* subset of the Burnside and Dollar controls has a similar effect, except that it undermines Collier and Dehn. The slim Collier and Dollar control set is more destructive, leaving only the Hansen and Tarp 2SLS and Dalgaard *et al.* results significant at 0.05. These, along with Collier and Hoeffler and Hansen and Tarp GMM pass the Guillaumont and Chauvet control test. Overall, the Collier and Dehn results are most fragile to control changes.

Altering the definitions of aid proves to be a mild test. As expected, in the first four tested regressions, which use real GDP in the denominator of aid, switching between ODA and EDA in the aid/GDP numerator has little effect. Changing the denominator turns out not to be a much harder test. But changing the definition of policy proves tougher. The Collier and Dehn, Collier and Dollar, and Hansen and Tarp regressions are all vulnerable to it to various degrees. Aggregating variables over 12 years is uniformly destructive. The drastic reductions in sample size may be the reason, and may make this test "unfair"; on the other hand, several regressors not involving aid pass this test easily in the Burnside and Dollar test results in Table 11.

Overall, the Collier and Hoeffler result on post-conflict 1×aid×policy (or the collinear post-conflict 1×aid), the Hansen and Tarp 2SLS coefficients on aid and aid^2, and the Dalgaard *et*

[24] However, I do take license to stretch their definition of post-conflict 1 to 5-year periods in the "AR-robust" tests.

al. results for aid and aid×tropical area fraction persist most under tests inspired by "whimsy" in the literature. They are the most whimsy-robust, one could say. Interestingly, except for the Hansen and Tarp 2SLS coefficients, all of these relatively strong results center on highly non-normal variables. The Collier and Hoeffler post-conflict 1 dummy is 1 for only 13 of the 344 observations in their original sample. Only 38 of the 234 Collier and Dehn observations have negative shocks. In the Dalgaard *et al.* sample, 233 of the 371 observations have tropical area fraction=1 and 68 have it 0, leaving 70 in between. Evidently regularities involving such variables are more resilient to specification changes.

The Guillaumont and Chauvet result on aid×environment is also persistent, but is frequently marred by autocorrelation in the tests. (Indeed, the *p* value for the Arellano-Bond test for AR(1) in my reproduction of the original regression is 0.08, as shown in Table 10.)

Only Collier and Dehn's regression included export shocks per se, so only it was subjected to the sensitivity tests relating to that variable. (See Table 13.) The result on Δaid×negative shock is robust to the two modifications of the shock definition whether applied separately or together.

Results from sample-modifying tests appear in Table 14. The first two results columns are based on regressions on the original authors' datasets—first for their full sample, second for the sample excluding outliers picked by the Hadi algorithm on partial scatterplots. The next pair of columns are analogous, but for the expanded data set. Autocorrelation is prevalent in the expanded-sample results. So the final pair of columns shows the results from making expanded-sample regressions "AR-robust."

Except for the Guillaumont and Chauvet result, all the original OLS and 2SLS results depend on outliers for some or all of their significance. The dependence is particularly heavy for

the regressions involving aid×policy. On the other hand, the *lack* of significance of most of the coefficients under the sample-expansion test is not driven by new, wayward outliers. Excluding them does not restore significance. (See Figures 1–8.)

The weakest results again are those on aid×policy, from Burnside and Dollar, Collier and Dollar, and Collier and Dehn. The Collier and Dehn coefficient on Δaid×negative shock is stronger—arguably stronger than a glance at the table suggests. The coefficient is dramatically reversed by the exclusion of outliers from the original sample. So it would seem that, leaving aside extreme observations, increasing aid to countries undergoing such shocks slows growth. However, shocks are by definition outlier events. The 10 outliers identified in the Collier and Dehn original-sample regression include 8 of the 38 negative shock episodes in the sample. It may be inconsistent therefore to draw conclusions about the role of shocks in growth having excluded many of the most dramatic examples. On the other hand, 30 shock episodes remain even after excluding outliers.

Two regressions without such stochastic terms go relatively unscathed in the first four columns of Table 14. The Guillaumont and Chauvet result goes relatively unscathed but also relatively untested because I am not able to expand its sample. I do nearly double the sample of the Hansen and Tarp 2SLS regression, and it does comparatively well in the first four columns. The coefficients on aid and aid^2 achieve similar values and *t* statistics in the expanded sample as in the original, and the significance is not dependent on outliers.

However, all these expanded-sample results are in some doubt because all fail autocorrelation tests at 0.05 or 0.06. In the face of specification problems that can bias results, one can err on the side of inclusion—taking the regressions as they are, caveat emptor—or exclusion—purging the data of biasing information, but along with it, probably, information that is both rel-

vant and non-biasing. Table 14 does both. Specifically, its final pair of columns shows the results of making the sample-expansion regressions "AR-robust"—using 5-year periods and autocorrelation-robust standard errors, and instrumenting most variables with their lags. All the OSL and 2SLS results that can be put to this test fail it. But again, though tough, the test is not inherently nihilistic. In the Burnside and Dollar regression, both the Sub-Saharan Africa dummy and the institutional quality index survived this test.

Meanwhile, the two GMM regressions are, if anything, revivified by the AR-robust sample-expansion test. Of course for these regressions, which already instrument most right-hand-side variables and have autocorrelation-robust standard errors, the AR-robust test consists only in switching from 4- to 5-year periods in order to expunge higher-order autocorrelation that can render instruments invalid. So the test for these regressions is arguably less severe. Nevertheless, the overall pattern is striking. The most methodologically conservative regressions tested produce coefficients in the expanded sample that, though smaller, are significant at or near 0.05 and consistent with the authors' original results. Of these, the Dalgaard *et al.* results also pass almost all the "whimsy"-inspired tests (Table 12), the only exception being the 12-year test.

Table 11. Robustness tests of Burnside and Dollar regression

		Changing controls			Changing aid		Changing policy		Changing periods	Changing sample				
											New data		New data, AR-robust	
	Original	CD	GC	DHT	ODA/ PPP GDP	ODA/ XR GDP	INFL, SACW	CPIA	12-year	Original, no outliers	Full sample[1]	No outliers[1]	Full sample[1]	No outliers
Log initial real GDP/capita	–0.60	–0.37	–0.05	–0.43	–0.30	–0.32	–0.39	–0.50	–0.61	–1.01	–0.39	–0.49	–1.16	–0.90
	(–1.02)	(–0.65)	(–0.11)	(–0.78)	(–0.52)	(–0.57)	(–0.69)	(–0.90)	(–1.06)	(–1.90)	(–1.14)	(–1.39)	(–1.90)	(–1.75)
Ethno-linguistic fractionalization	–0.42		–1.10		–0.86	–0.93	–0.93	–0.83	–0.36	–0.54	–0.80	–0.72	–0.51	–0.41
	(–0.57)		(–1.62)		(–1.12)	(–1.18)	(–1.21)	(–1.15)	(–0.34)	(–0.73)	(–1.19)	(–1.09)	(–0.47)	(–0.39)
Assassinations/ capita	–0.45				–0.50	–0.47	–0.47	–0.36	–0.30	–0.40	–0.39	–0.39	0.19	0.05
	(–1.68)				(–1.88)	(–1.74)	(–1.74)	(–1.22)	(–0.99)	(–1.41)	(–1.65)	(–1.62)	(0.44)	(0.14)
Ethnic × Assas.	0.79				**0.87**	0.82	0.77	0.51	0.49	0.69	0.26	0.24	–0.84	–0.44
	(1.74)				**(1.98)**	(1.85)	(1.68)	(0.98)	(0.68)	(1.50)	(0.41)	(0.38)	(–0.77)	(–0.51)
Sub–Saharan Africa	**–1.87**			**–2.28**	**–1.68**	**–1.48**	**–1.80**	**–1.81**	**–1.71**	**–2.28**	**–1.20**	**–1.37**	**–1.83**	**–1.77**
	(–2.41)			**(–3.58)**	**(–2.38)**	**(–2.06)**	**(–2.55)**	**(–2.76)**	**(–2.00)**	**(–3.26)**	**(–2.20)**	**(–2.56)**	**(–2.74)**	**(–2.73)**
Fast–growing E. Asia	**1.31**			0.89	0.78	0.82	0.84	**1.55**	1.09	1.04	0.74	0.65	1.64	1.37
	(2.19)			(1.53)	(1.31)	(1.36)	(1.37)	**(2.51)**	(1.48)	(1.81)	(1.45)	(1.27)	(1.69)	(1.56)
Institutional quality (ICRGE)	**0.69**	**0.67**		0.71	**0.63**	**0.63**		0.15	**0.64**	**0.67**	**0.38**	**0.36**	**0.44**	**0.43**
	(3.90)	**(3.71)**		(0.71)	**(3.48)**	**(3.41)**		(0.74)	**(3.30)**	**(3.74)**	**(3.56)**	**(3.38)**	**(3.02)**	**(3.21)**
M2/GDP, lagged	0.01		**0.04**		0.01	0.01	0.01	0.01	0.04	0.01	0.01	0.01	0.02	0.01
	(0.84)		**(2.68)**		(1.01)	(1.01)	(0.75)	(0.69)	(1.84)	(0.62)	(0.54)	(0.86)	(1.03)	(0.99)
Policy	**0.71**	**0.91**	**0.92**	**0.80**	**0.81**	**0.80**	**0.79**		**0.78**	**0.89**	**1.34**	**1.40**	0.45	0.77
	(3.63)	**(4.66)**	**(7.57)**	**(3.93)**	**(4.16)**	**(3.84)**	**(4.31)**		**(3.07)**	**(4.53)**	**(5.65)**	**(5.92)**	(0.74)	(1.56)
Mean years secondary schooling			–0.01											
			(–0.04)											
Population growth			–0.01											
			(–0.06)											
Political instability, lagged			–0.61											
			(–0.94)											
Aid	–0.02	0.17	0.01	0.16	0.16	0.04	–0.24	0.07	–0.12	0.06	0.31	0.41	–0.89	0.14
	(–0.13)	(0.88)	(0.07)	(0.73)	(0.80)	(0.72)	(–0.73)	(0.46)	(–0.52)	(0.26)	(1.15)	(1.40)	(–0.98)	(0.18)
Aid × policy	**0.19**	0.13	0.06	**0.19**	0.14	0.05	**0.15**	0.00	0.12	–0.05	–0.15	–0.28	0.39	–0.19
	(2.61)	(1.73)	(1.06)	**(2.27)**	(1.95)	(1.63)	**(2.32)**	(0.01)	(1.31)	(–0.45)	(–1.04)	(–1.54)	(0.97)	(–0.45)
Log population		0.18	**0.49**	**0.35**	**0.39**	**0.36**	**0.40**	0.27	0.17		**0.40**	**0.40**	0.10	0.22
		(1.18)	**(3.73)**	**(2.28)**	**(2.50)**	**(2.44)**	**(2.52)**	(1.67)	(0.85)		**(3.19)**	**(3.25)**	(0.41)	(1.06)
Observations	270	278	262	278	275	272	268	264	97	263	430	420	287	276
R^2	0.39	0.43	0.41	0.42	0.46	0.46	0.43	0.38	0.61	0.41	0.37	0.38	0.39	0.40
AR(1) (*p* value)	0.54	0.48	**0.05**	0.24	0.34	0.32	0.36	0.92	0.31	0.19	**0.01**	0.08	**0.04**	0.13

[1]Regression fails Arellano-Bond test for AR(1) at 0.05 for either the full regression sample or that excluding residual–lagged residual outliers.
Period dummies and constant term not reported. All *t* statistics are heteroskedasticity-robust; those in last two columns also robust to arbitrary patterns of autocorrelation. Entries significant at 0.05 in bold.

Table 12. Coefficients on key terms under specification-modifying tests (original data set)

Specification	Key term	Original	Changing controls: BD	CD	GC	DHT	Changing aid: EDA/ real GDP	ODA/ real GDP	ODA/ XR GDP	Changing policy: INFL, BB, SACW	INFL, SACW	CPIA	Changing periods: 12-year
Burnside & Dollar	Aid × policy	**0.19**		0.13	0.06[1]	**0.19**		0.14	0.05		**0.26**	0.00	0.12
		(2.61)		(1.73)	(1.06)	**(2.27)**		(1.95)	(1.63)		**(3.02)**	(0.01)	(1.31)
		270		278	262	278		275	272		296	264	97
Collier & Dehn	Aid × policy	0.10		0.03	0.02	0.04		0.06	0.02		0.09	0.15	–0.02
		(1.76)		(0.46)	(0.51)	(0.61)		(1.02)	(1.33)		(1.65)	(1.07)	(–0.16)
	ΔAid × negative shock	**0.04**		0.02	0.02	0.02		**0.03**	**0.01**		0.03	0.02	0.01
		(3.17)		(1.11)	(1.21)	(0.93)		**(2.83)**	**(4.41)**		(1.93)	(0.83)	(1.03)
		234		242	227	242		281	281		256	268	51
Collier & Dollar	Aid × policy	**0.14**	**0.21**		–0.01	**0.20**	0.17		0.04	0.07	0.06		0.19
		(2.15)	**(2.87)**		(–0.24)	**(2.71)**	(1.72)		(1.65)	(1.44)	(1.19)		(1.64)
		344	337		374	349	349		347	351	388		119
Collier & Hoeffler	Post-conflict 1 × aid × policy	**0.18**	**0.16**		**0.11**[1]	**0.16**[1]	**0.30**		**0.04**	**0.15**	**0.18**[1]		
		(3.92)	**(3.54)**		**(2.70)**	**(3.46)**	**(3.64)**		**(3.93)**	**(3.03)**	**(4.13)**		
		344	337		374	349	349		347	351	388		
Hansen & Tarp 2SLS	Aid	**0.24**		**0.32**	**0.38**	**0.39**[1]	0.74	**1.00**			**0.24**	0.11[3]	0.20
		(2.34)		**(2.75)**	**(3.81)**	**(3.27)**	(1.85)	**(2.99)**			**(2.61)**	(1.45)	(1.13)
	Aid^2	**–0.01**		**–0.01**	**–0.01**	**–0.01**[1]	–0.08	**–0.09**			**–0.01**	0.00[3]	–0.01
		(–2.38)		**(–2.60)**	**(–2.94)**	**(–3.07)**	(–1.16)	**(–2.33)**			**(–2.51)**	(–1.31)	(–1.14)
		231		232	206	232	231	231			264	216	50
Hansen & Tarp GMM	Aid	**0.90**		1.04	**1.12**	1.04	–0.05	2.20		**0.91**		1.01	
		(4.22)		(1.64)	**(3.04)**	(1.64)	(–0.03)	(1.59)		**(2.01)**		(1.46)	
	Aid^2	**–0.02**		–0.02	**–0.02**	–0.02	0.06	–0.13		–0.02		–0.02	
		(–3.83)		(–1.07)	**(–2.41)**	(–1.07)	(0.16)	(–1.04)		(–1.45)		(–0.49)	
	ΔAid	**–0.70**		–0.52	–0.71	–0.52	–0.83	–1.79		**–0.80**		–0.86	
		(–4.91)		(–1.50)	(–1.93)	(–1.50)	(–0.96)	(–1.92)		**(–2.51)**		(–1.88)	
	$\Delta(Aid^2)$	**0.01**		0.01	0.02	0.01	0.03	0.11		0.01		0.01	
		(3.64)		(0.97)	(1.62)	(0.97)	(0.10)	(1.15)		(1.87)		(1.00)	
		213		213	181	213	214	214		213		215	
Dalgaard *et al.* GMM	Aid	**1.47**	**1.37**	**1.47**	**1.34**			**0.76**	**0.29**				0.15
		(4.02)	**(3.50)**	**(2.99)**	**(2.69)**			**(4.20)**	**(3.32)**				(0.02)
	Aid × tropical area fraction	**-1.33**	**–1.73**	**–1.70**	**–1.55**			**–0.87**	**–0.23**				–2.30
		(-2.62)	**(–4.10)**	**(–2.80)**	**(–3.15)**			**(–4.27)**	**(–2.37)**				(–0.14)
		371	354	371	315			371	365				116
Guillaumont & Chauvet	Aid × environment	–0.15	**–0.16**	–0.15		–0.12	**–0.49**[1]	–0.35[1]		–0.16[1]	**–0.15**[1]	**–0.14**	
		(–1.79)	**(–2.01)**	(–1.88)		(–1.82)	**(–2.03)**	(–1.96)		(–1.91)	**(–2.30)**	**(–2.38)**	
		68	71	73		73	66	66		66	68	69	

[1]Regression fails Arellano-Bond test for AR(1) at 0.05 significance level for either the full regression sample or that excluding residual–lagged residual outliers. [2]Regression fails Arellano-Bond test for AR(2) at in first differences at 0.05. [3]Regression fails Hansen *J* test of overidentifying restrictions at 0.05.
All *t* statistics are heteroskedasticity-robust; those for GMM regressions also autocorrelation-robust. Except for original Hansen and Tarp regression, all GMM standard errors include the Windmeijer (2000) finite-sample correction. Entries significant at 0.05 in bold.

Table 13. Coefficients on key terms in Collier and Dehn regression under tests of shock definition

Distribution of forecasting errors used for shock definition:	Country-specific	Country-specific	Pooled	Pooled
Shock measurement:	Price change	Share of GDP	Price change	Share of GDP
Aid × policy	0.10	0.07	0.06	0.04
	(1.76)	(1.06)	(1.00)	(0.91)
∆Aid × negative shock	**0.04**	0.18	**0.03**	**0.50**
	(3.17)	(1.96)	**(2.35)**	**(4.25)**
Observations	234	234	234	234

Heteroskedasticity-robust *t* statistics in parenthesis. Those significant at 0.05 in bold. First results column is for original specification.

Table 14. Coefficients on key terms under data set–modifying tests

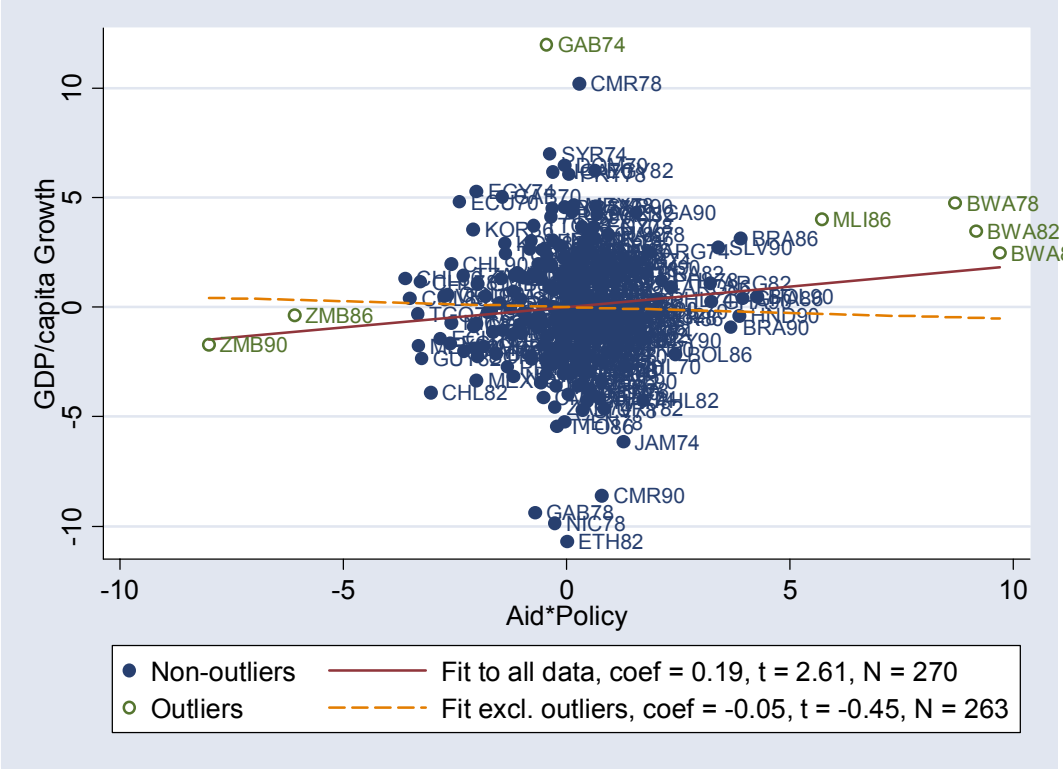
Specification	key term	Original		Expanded sample		Expanded sample, AR-robust	
		Full sample	No outliers	Full sample	No outliers	Full sample	No outliers
B&D 5/OLS	Aid × policy	**0.19** **(2.61)**	–0.05 (–0.45)	–0.15[1] (–1.04)	–0.28[1] (–1.54)	0.39[1] (0.97)	–0.19 (–0.45)
	Observations	270	263	430	420	287	276
Collier & Dehn	Aid × policy	0.10 (1.70)	0.11 (1.11)	0.03[1] (0.58)	0.01[1] (0.05)	0.16 (1.39)	0.35 (1.81)
	ΔAid × negative shock	**0.04** **(3.17)**	–0.06 (–1.33)	**0.03**[1] **(2.54)**	–0.16[1] (–1.75)	0.01 (0.64)	0.03 (0.32)
	Observations	234	224	388	364	297	279
Collier & Dollar	Aid × policy	**0.14** **(2.15)**	0.07 (1.06)	0.00[1] (0.06)	–0.02[2] (–0.30)	–0.03 (–0.43)	–0.03 (–0.58)
	Observations	344	341	521	506	296	291
Collier & Hoeffler	Post-conflict 1 × aid × policy	**0.18** **(3.92)**	**1.18** **(2.12)**	**0.08**[1] **(2.12)**	–0.10[2] (–0.29)	0.09 (0.96)	0.22 (0.43)
	Observations	344	333	521	495	296	287
Hansen & Tarp 2SLS	Aid	**0.24** **(2.34)**	0.15 (1.82)	**0.18**[1] **(2.35)**	**0.18**[1] **(2.13)**	–0.01 (–0.07)	–0.02 (–0.20)
	Aid^2	**–0.01** **(–2.38)**	–0.005 (–1.90)	**–0.01**[1] **(–2.28)**	–0.01[1] (–1.43)	0.00 (0.27)	0.00 (0.38)
	Observations	231	229	409	406	294	293
Hansen & Tarp GMM	Aid	**0.90** **(4.22)**		0.16[3] (0.81)		0.29 (1.86)	
	Aid^2	**–0.02** **(–3.83)**		–0.002[3] (–1.01)		–0.01 (–1.95)	
	ΔAid	**–0.70** **(–4.91)**		–0.18[3] (–1.12)		**–0.30** **(–2.43)**	
	$\Delta(Aid^2)$	**0.01** **(3.64)**		0.002[3] (1.08)		**0.01** **(2.19)**	
	Observations	213		516		381	
Dalgaard *et al.* GMM	Aid	**1.47** **(4.02)**		**0.19**[4] **(2.60)**		**0.41** **(4.39)**	
	Aid × tropical area fraction	**-1.33** **(-2.62)**		–0.10[4] (–1.30)		**–0.39** **(–3.73)**	
	Observations	371		450		343	
Guillaumont & Chauvet	Aid × environment	–0.15 (–1.79)	–0.11 (–1.96)				
	Observations	68	67				

[1]Regression fails Arellano-Bond test for AR(1) at 0.05 for either the full regression sample or that excluding residual–lagged residual outliers. [2]Fails same test at 0.06. [3]Regression fails test for AR(3) in first-differences at 0.002, rendering several instruments invalid. [4]Regression fails test for AR(3) and AR(4) in first-differences at 0.05, rendering several instruments invalid.

All *t* statistics are heteroskedasticity-robust; those in last two columns or in GMM regressions also autocorrelation-robust. Except for original Hansen and Tarp regression, all GMM standard errors include the Windmeijer (2000) finite-sample correction. Entries significant at 0.05 in bold.

Figure 1. B&D regression 5/OLS: Partial scatter of GDP/capita growth versus aid×policy

Original data



Expanded sample, AR-robust

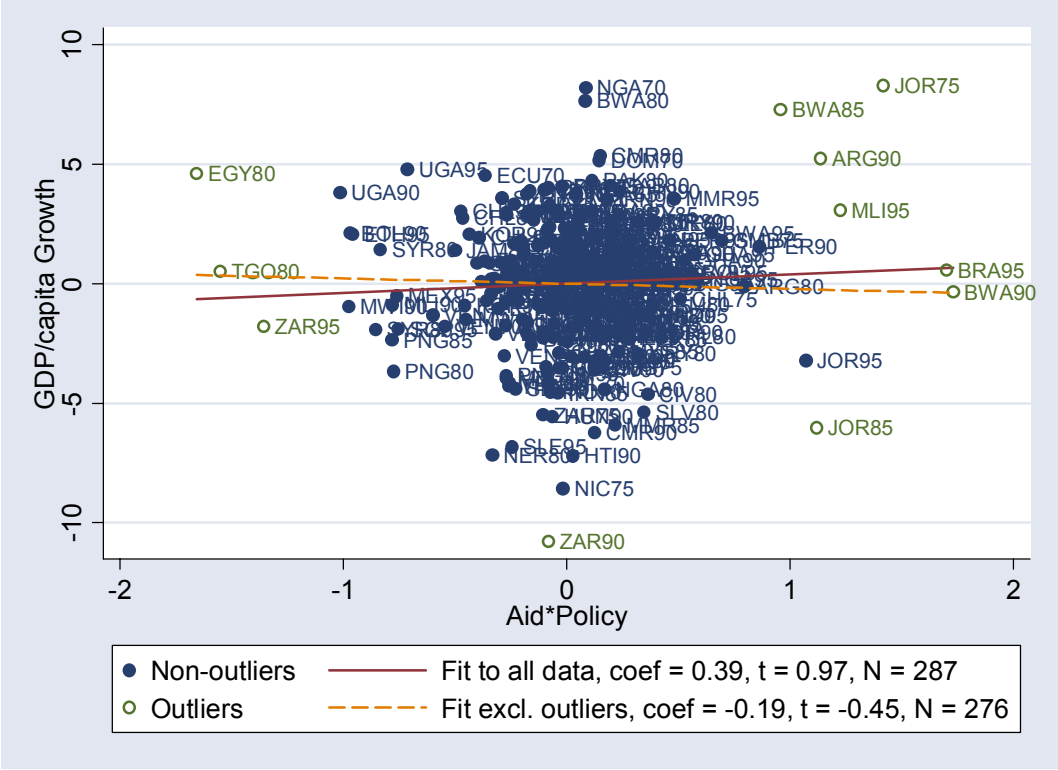
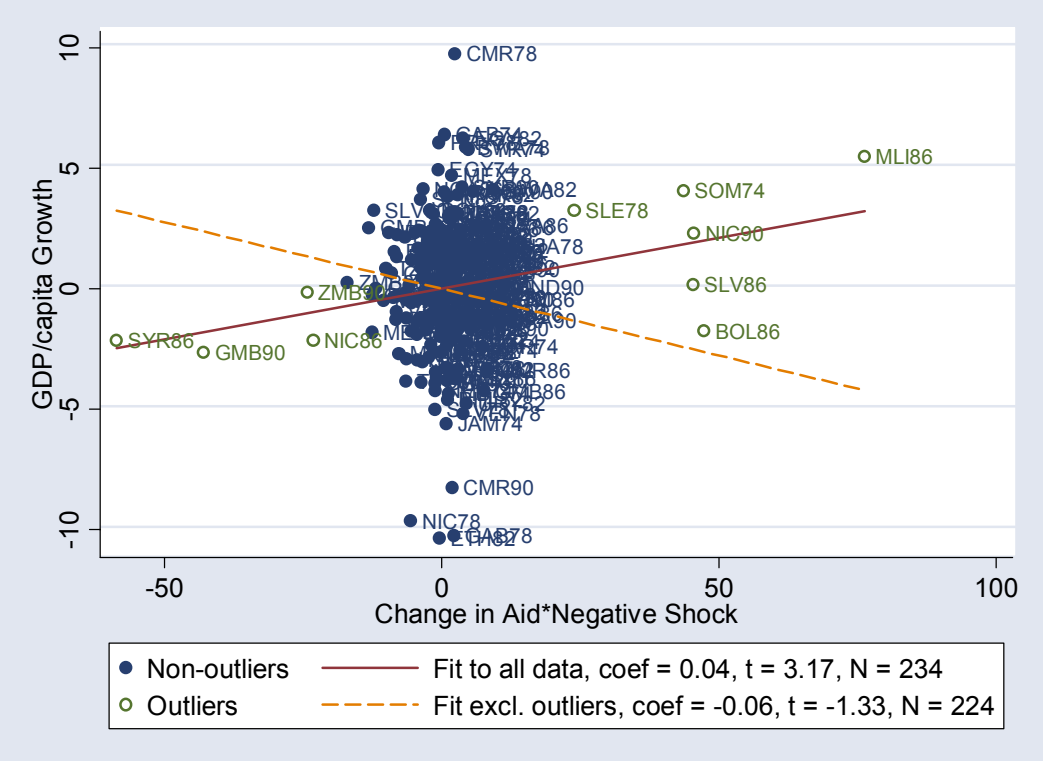


Figure 2. Collier and Dehn regression: Partial scatter of GDP/capita growth vs. Δaid×negative shock

Original data



Expanded sample, AR-robust

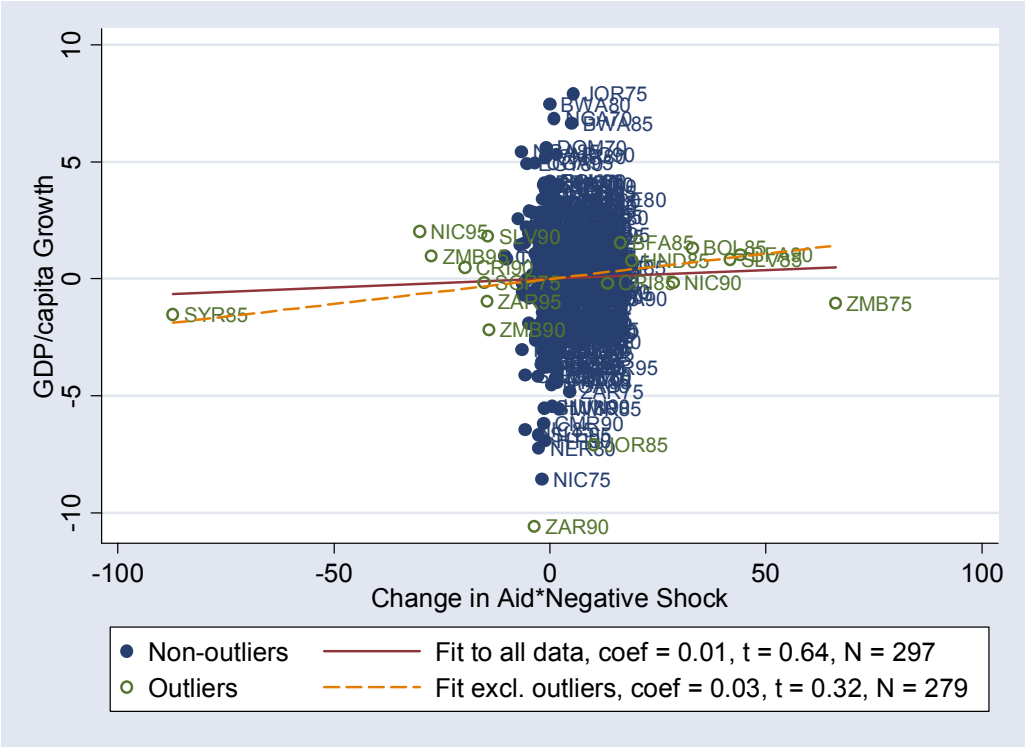
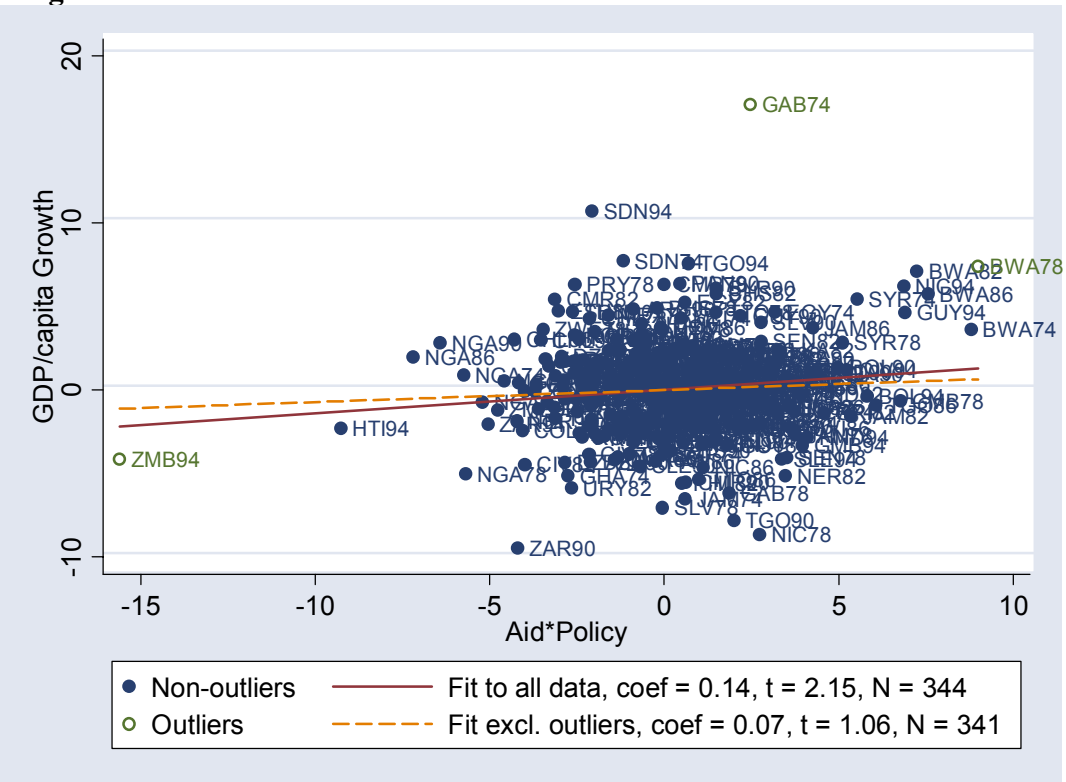


Figure 3. Collier & Dollar regression: Partial scatter of GDP/capita growth vs. aid×policy

Original data



Expanded sample, AR-robust

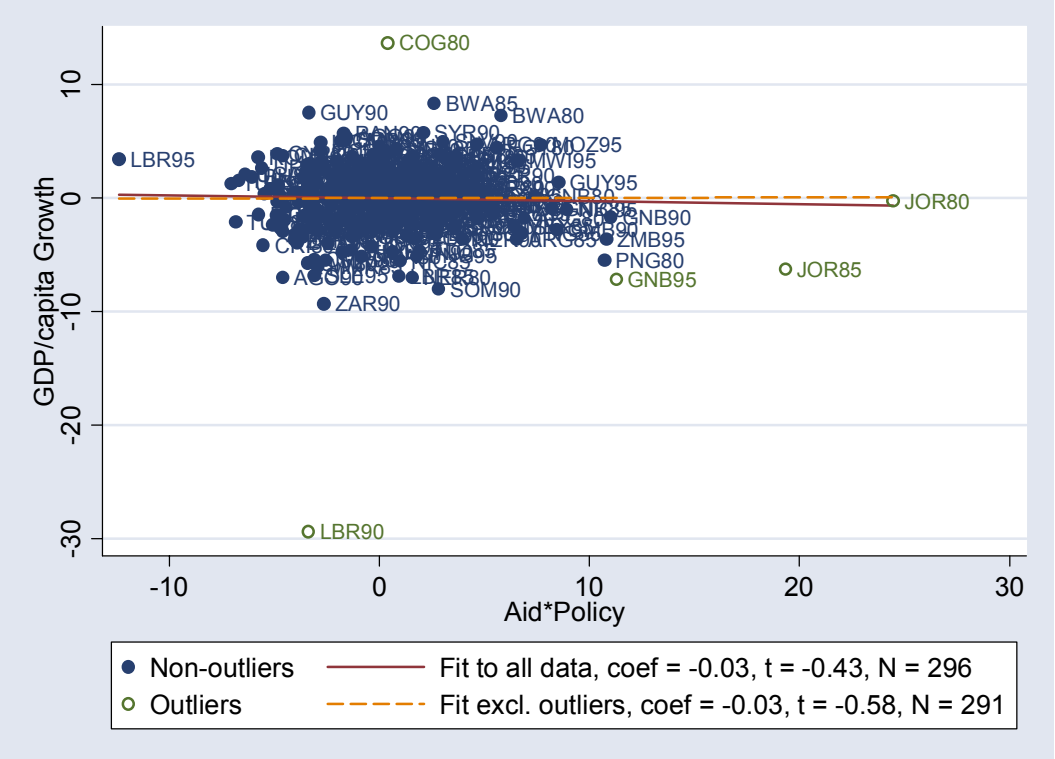
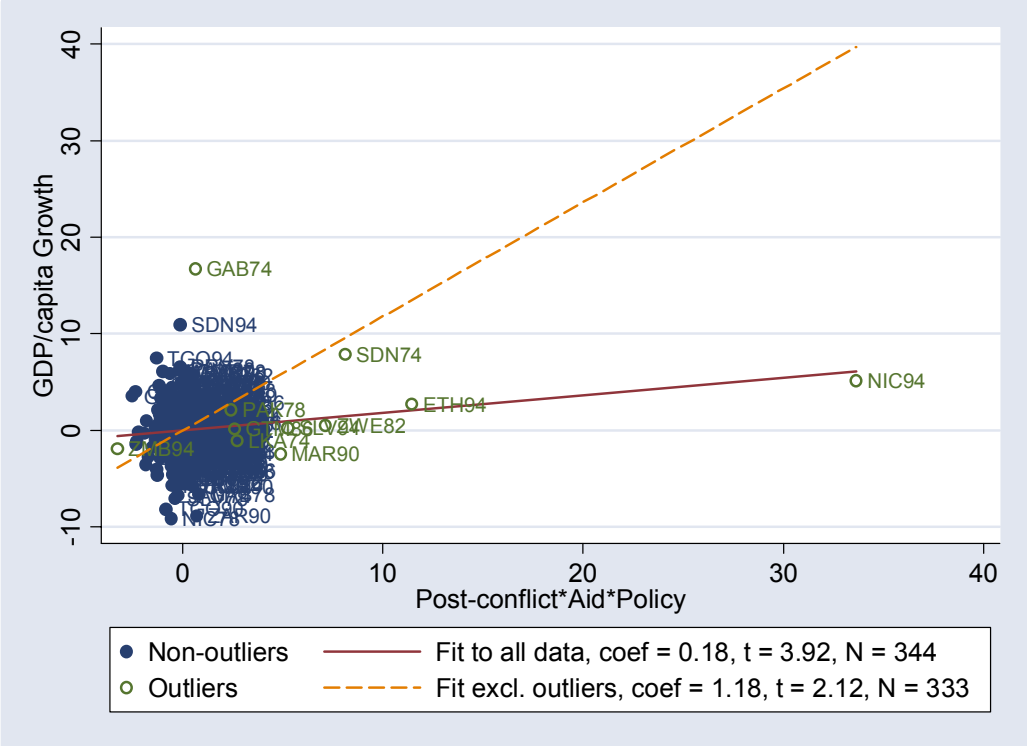


Figure 4. Collier & Hoeffler regression: Partial scatter of GDP/capita growth vs. Post-conflict 1×aid ×policy

Original data



Expanded sample, AR-robust

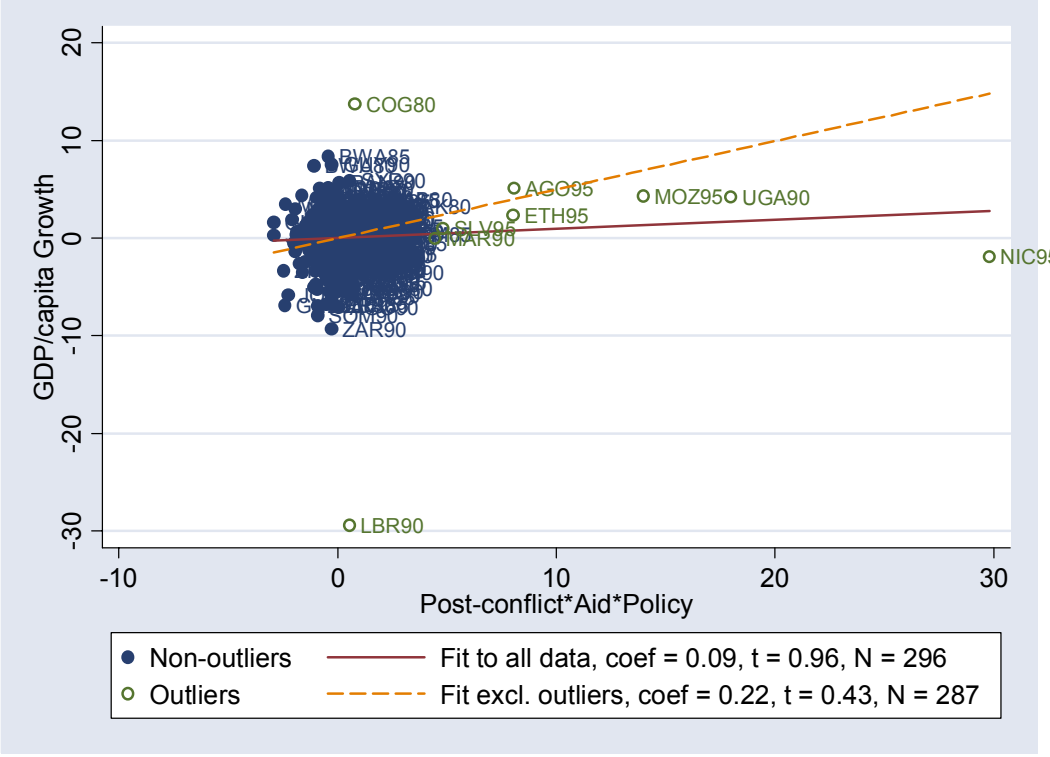
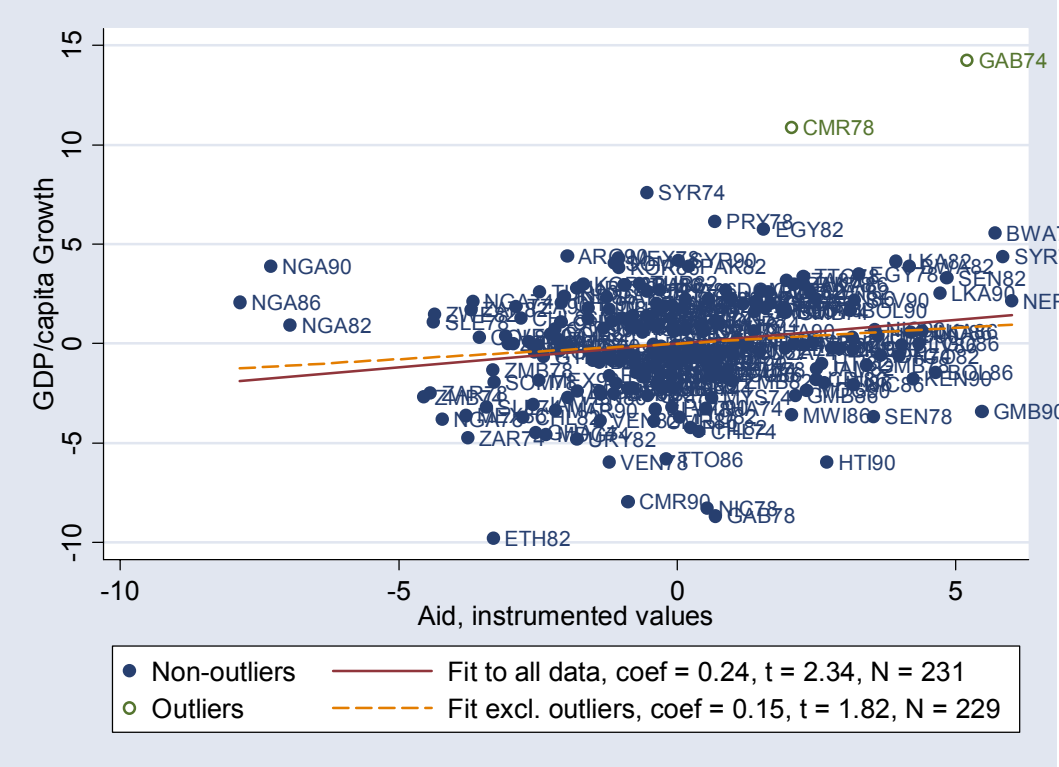


Figure 5. Hansen & Tarp 2SLS regression: Partial scatter of GDP/capita growth vs. aid

Original data



Expanded sample, AR-robust

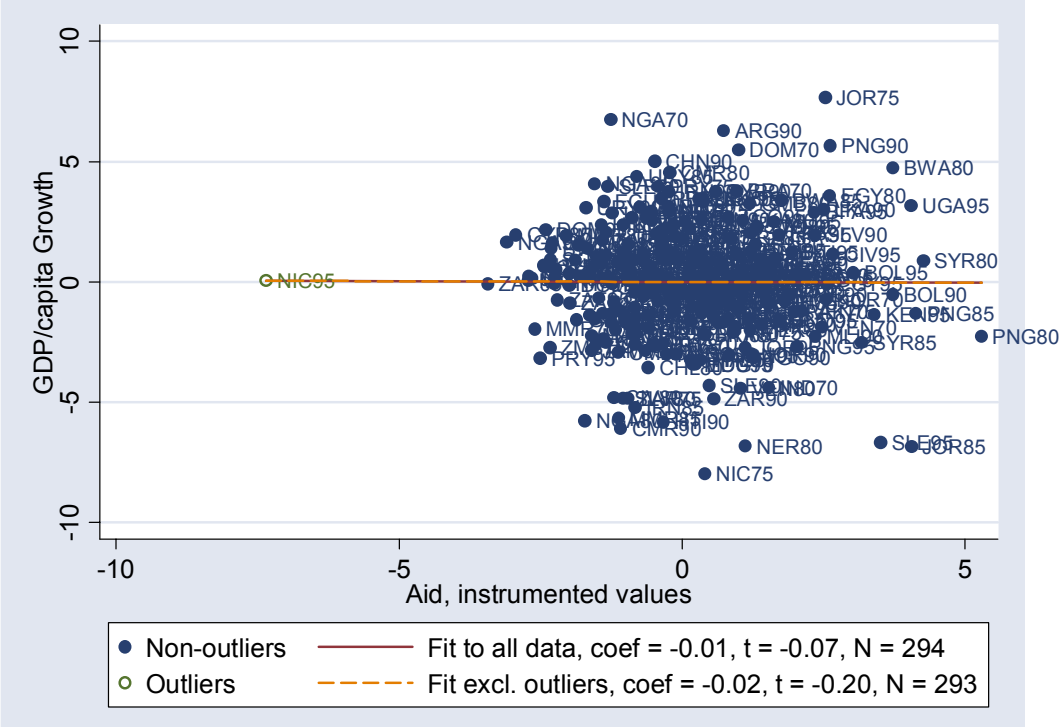
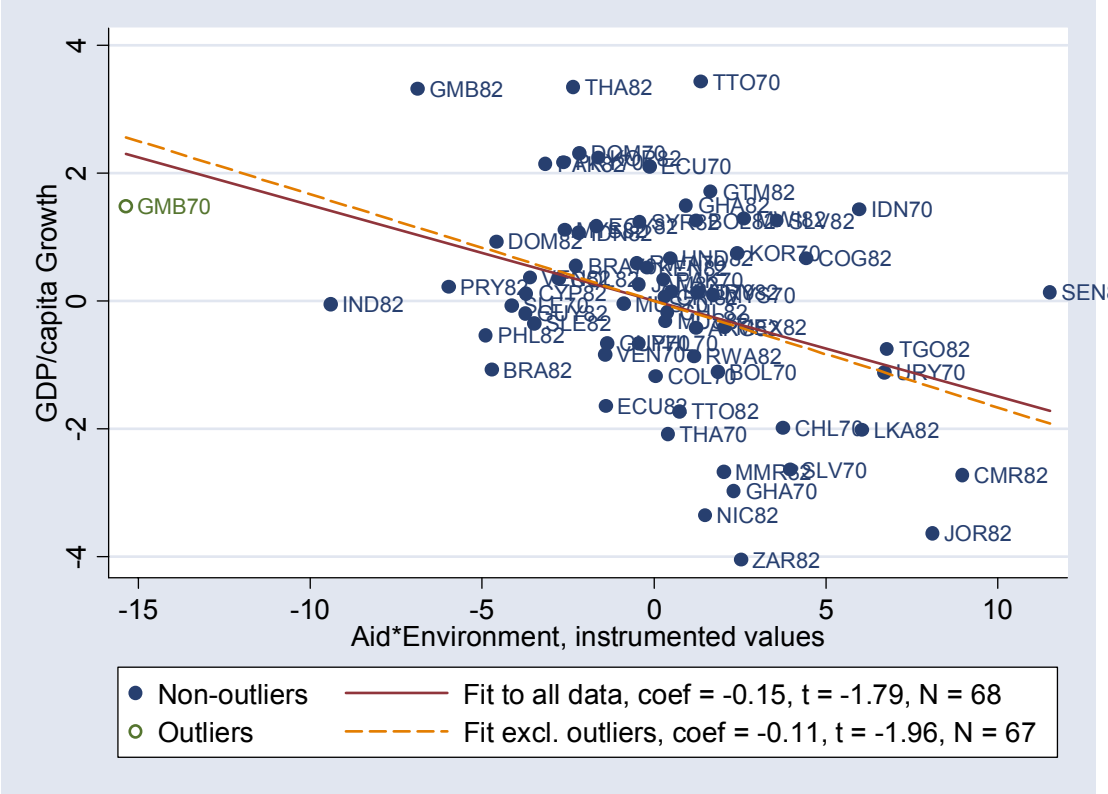


Figure 6. Guillaumont and Chauvet Regression 1.2: Partial Scatter of GDP/Capita Growth versus Aid×Environment

Original data



IV. Conclusion

Each of the papers examined here embodies a set of choices about model specification and data. Aid is measured a certain way. A certain epoch is studied. Periods have a certain length. And so on. Some of these choices imply certain assumptions about the world, such as, say, that aid is not endogenous to growth. All limit the scope of a strict interpretation of the results.

To wit, the Burnside and Dollar results can be stated more precisely as follows:

Aid was associated with higher GDP growth in a good policy environment during 1970–93, on average, in countries and periods where the necessary data was collected, except for outliers (unless aid^2 ×policy is included to allow for diminishing returns to aid), when aid is defined as "Effective Development Assistance" as a share of real GDP and polices are defined by inflation, budget deficit, and the complex Sachs-Warner "openness" variable, controlling for log of initial real GDP/capita, assassinations per capita, ethno-linguistic fractionalization, the product of those two, money supply/GDP, and period effects, assuming that no unobserved country-specific effects simultaneously and substantially influence aid, policies, and growth, and no variables other than aid aid×policy are endogenous to growth.

This is not quite "aid has a positive effect on growth in a good policy environment."

In fairness, such qualifiers could be tacked on to the conclusion of any study in the cross-country literature on aid and development, or in econometrics more generally (though perhaps not always quite so many). Moreover, Burnside and Dollar did test some of their assumptions, such as the exogeneity of the policy variables. Nevertheless, a question of great scientific and practical importance remains, and it is how many of such implied qualifiers in studies of aid and growth can be dropped without harming the conclusions. This study attempts to contribute to answering that question.

The test results reported here suggest that the fragility found in Easterly *et al.* (2004) for the case of Burnside and Dollar is common in the cross-country aid effectiveness literature.

In surveying the results, it is tempting to ask which results are robust and which are not. But the test data are best seen in shades of grey. The results tested here break roughly into five groups, ranging from weakest to strongest. The weakest group consists of the results on aid×policy in the Burnside and Dollar, Collier and Dollar, and Collier and Dehn regressions, which lose significance at 0.05 in all but the weakest tests. In the second division I put the Collier and Dehn result on Δaid×negative shock, which passes more tests but is quite sensitive to changes in the control set, as well as to removal of a minority of the negative shocks in the sample. The Collier and Hoeffler result on post-conflict 1×aid×policy (or the collinear post-conflict 1×aid), the Hansen and Tarp results on aid and aid^2, and Guillaumont and Chauvet result on aid×environment seem stronger. They generally pass the "whimsy"-based tests at or near 0.05. Most survive the sample-expansion test, but with autocorrelation—and then fail the AR-robust test meant to address this problem. The Guillaumont and Chauvet result could not be put

to the sample-expansion tests, so the degree of its robustness is less certain and I add it to this middle category.

In the fourth and fifth categories are the GMM results of Hansen and Tarp and Dalgaard *et al.*, respectively. Both fare well under the sample expansion. The Hansen and Tarp GMM results, especially those on aid and lagged aid, generally persist through the test suite, though not always significantly at 0.05. The only test that completely eliminates the Hansen and Tarp GMM results is that of defining aid as EDA/real GDP, but this is a misleading measure of aid. As for the Dalgaard *et al.* results, they come through powerfully in all the tests but the 12-year aggregation that reduces the sample size to 116.

Does this mean that the statistically weaker stories of aid effectiveness should be dismissed? Are recipient policies, exogenous economic factors, and post-conflict status irrelevant to aid effectiveness? No. There can be no doubt that aid sometimes finances investment (Hansen and Tarp, 2001), and that domestic policies, governance, external conditions, and historical circumstances influence the productivity of investment. Why then do such stories of aid effectiveness not shine more clearly through the numbers? The reasons are several. Aid is probably not a fundamentally decisive factor for development, not as important as, say, domestic savings, inequality, and governance. Moreover, foreign assistance is not homogenous. It consists of everything from in-kind food aid to famine-struck countries and technical advice on building judiciaries to loans for paving roads. And some aid is poorly used. Thus the statistical noise nearly drowns out the signal.

If there is one strong conclusion from this literature it is that *on average* aid works well outside the tropics but not in them. But just as it would be a mistake to conclude that the other stories of aid effectiveness contain no truth, it would also be mistaken to conclude that this result

is the wholly, simply true. Indeed, the Dalgaard *et al.* result is more of a question than an answer. Presumably distance from the poles is not a direct determinant of aid effectiveness. Rather, the causal pathways are complex, and so it cannot be assumed that *no* kind of aid will work well in the tropics. Much the same can be said about the more optimistic, but somewhat less robust, Hansen and Tarp result on the overall positive effect of aid on growth. Even accepting it as true, it gives little guidance about where aid ought to be sent, and in what forms.

Perhaps further econometric work will disaggregate aid by types of program and recipient and unearth more robust answers to the fundamental questions of aid policy. Or perhaps researchers have hit the limits of what cross-country empirics can reveal about aid. The search for truth may need to rely more on the particularistic case study approach. Van de Walle and Johnston (1996), for example, synthesize conclusions from case studies on the use and effects of aid in seven African countries, each jointly conducted by researchers from donor and recipient countries. Of course, the lessons that emerge from case studies are particular to the country studied. But they can be generalized. Killick (1998) provides an excellent example by conducting a systematic survey of case studies that feeds into a trenchant analysis of the effects of IMF conditionality. Nevertheless, robust generalizations will not come easily.

Appendix. Data set construction

The new data set used in this study is based heavily on that for Easterly *et al.* (2004). Some variables in that set have been slightly revised. Others have been added to match the data sets of the tested regressions. The period of coverage has been pushed back to 1958 and extended forward to 2001 for most variables. All data were collected from standard cross-country sources, expect for countries' export price indexes, which were provided by Jan Dehn (see Dehn (2000)). (See Table A–1.)

Following are notes on the data set construction:

(a) Revisions since Easterly et al. (2004)

- Some observations for inflation were completed by using wholesale inflation where consumer price inflation was unavailable.
- The update of the Sachs-Warner variable was slightly revised under the influence of the independent update by Wacziarg and Welch (2002). The full update will be described and published separately.
- Some missing values for Effective Development Assistance during 1975–95, the period of the EDA data set, were filled in in the same manner as missing values outside this period already were, via a regression of EDA on net ODA.
- ICRGE now varies over time, rather than taking 1982 values throughout. Observations before 1982 were assigned 1982 values. In addition, the variable was revised in order to extend it beyond 1997. In 1998, the PRS Group stopped reporting two of ICRGE's original components, Expropriation Risk and Repudiation of Government Contracts. So these were dropped entirely from ICRGE, leaving Corruption, Bureaucratic Quality, and Rule of Law. On annual data, the revised ICRGE has a 0.97 correlation with the original.
- Missing values for ethnolinguistic fractionalization were filled in from Roeder (2001).

(b) Expansion of period

- Data was collected for all available years in 1958–2001.
- However, the Collier and Dehn shocks variables were only updated to 1997 because the underlying data on export prices from Dehn (2000) cease in 1997.
- The Guillaumont and Chauvet environment variable was not updated at all, for lack of underlying data on its four components.
- The 1998–2001 values for the updated Sachs-Warner variable are based on 1998 data only. Currency Data International, the long-time source of black market premium data, which is one component of Sachs-Warner, shut down in 1999.

Table A–1. Construction of data set

Variable	Code	Data source	Notes[1]
Per-capita GDP growth	GDPG	World Bank, 2003	
Initial GDP per capita	LGDP	Summers and Heston, 1991, updated using GDPG	Natural logarithm of GDP/capita for first year of period; constant 1985 dol-

			lars
Ethno-linguistic fractionalization, 1960	ETHNF	Roeder, 2001	Probability that two individuals will belong to different ethnic groups
Assassinations/ capita	ASSAS	Banks, 2002	Assassinations/capita
Political instability, lagged	PINSTAB–1	Banks, 2002	Simple average of ASSAS and revolutions/year
Institutional quality	ICRGE	PRS Group's IRIS III data set (see Knack and Keefer, 1995)	Revised version of variable. Computed as the average of the three components still reported after 1997.
M2/GDP, lagged one period	M2–1	World Bank, 2003	
Sub-Saharan Africa	SSA	World Bank, 2003	Codes nations in the southern Sahara as sub-Saharan
East Asia	EASIA		Dummy for China, Indonesia, South Korea, Malaysia, Philippines, and Thailand, following Burnside and Dollar
Budget surplus	BB	World Bank, 2003; IMF, 2003	World Bank primary data source. Additional values extrapolated from IMF, using series 80 and 99b (local-currency budget surplus and GDP)
Inflation	INFL	World Bank, 2003; IMF, 2003	Natural logarithm of 1 + inflation rate. World Bank primary data source. Wholesale price inflation from IMF used where consumer price data unavailable
Sachs-Warner, updated	SACW	Sachs and Warner, 1995; Easterly *et al.,* 2004; Wacziarg and Welch, 2002	Extended to 1998. Slightly revised pre-1993. Full description will be published separately
Positive (and negative) shock	POSSHOCK NEGSHOCK	Dehn, 2000	Shocks are % price index changes. "Shock" threshold country-specific. Reconstructed based on underlying index data for 1957–97
Positive (and negative) shock/GDP	POSSHOCKGDP NEGSHOCKGDP		Shocks are shares of GDP, computed using 1990 data on commodity share of exports and exports share of

			GDP from Dehn 2000. "Shock" threshold country-specific
Positive (and negative) shock, pooled basis	POSSHOCKPOOLED NEGSHOCKPOOLED		Shocks are % price changes. "Shock" threshold universal
Positive (and negative) shock/GDP, pooled basis	POSSHOCKGDPPOOLED NEGSHOCKGDPPOOLED		Shocks are shares of GDP. "Shock" threshold universal
Effective Development Assistance/ real GDP	AID	Chang *et al.*, 1998; DAC, 2002; IMF, 2003; World Bank, 2003; Summer and Heston, 1991	Available values for 1975–95 from Chang *et al.* Missing values extrapolated based on regression of EDA on Net ODA. Converted to 1985 dollars with World Import Unit Value index from IMF, series 75. GDP computed like LGDP above
Net Overseas Development Assistance/real GDP	ODAPPPGDP	DAC, 2002; IMF, 2003; World Bank, 2003; Summer and Heston, 1991	Like AID exception using ODA from DAC
Net Overseas Development Assistance/nominal GDP	ODAXRGDP	DAC, 2002; World Bank, 2003	.
Dummy for end of civil conflict in previous period	POSTCONFLICT1	Collier and Hoeffler, 2002	
Tropical area fraction	TROPICAR	Gallup and Sachs, 1999	
Population	LPOP	World Bank, 2003	Natural logarithm
Population growth	POPG	World Bank, 2003	
Mean years of secondary schooling among those over 25	SYR	Barro and Lee, 2000	
Arms imports/total imports lagged	ARMS-1	U.S. Department of State, various years	

[1]All variables aggregated over time using arithmetic averages.

References

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001). 'The colonial origins of comparative development: An empirical investigation', *American Economic Review*, vol. 91, pp. 1369–1401.

Anderson, T.W. and Hsiao, C. (1982). 'Formulation and Estimation of Dynamic Models Using Panel Data', *Journal of Econometrics*, vol. 18(1) (January), pp. 47–82.

Arellano, M. and Bond, S. (1991). 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *The Review of Economic Studies*, vol. 58(2) (April), pp. 277–97.

Banks, A. (2002). *Cross-National Time-Series Data* Ar*chive,* Bronx, NY: Databanks International.

Barro, R. J. (1991). 'Economic Growth in a Cross Section of Countries', *American Economic Review*, vol. 106(2) (May), pp. 407–43.

Barro, R. J. and Lee, J. (2000). 'International Data on Educational Attainment: Updates and Implications', Working Paper 7911, National Bureau of Economic Research, Cambridge, MA (September).

Bauer, P.T. (1976). *Dissent on Development,* Cambridge, MA: Harvard University Press.

Bloom, D. and Sachs, J. D. (1998). 'Geography, demography and economic growth in Africa', *Brookings Papers on Economic Activity*, vol. 2, pp. 207–73.

Blundell, R. and Bond, S. (1998). 'Initial conditions and moment restrictions in dynamic panel data models', *Journal of Econometrics,* vol. 87(1) (November), pp. 115–43.

Boone, P. (1994). 'Aid and growth', mimeo, London School of Economics.

Burnside, C. and Dollar, D. (2000). 'Aid, Policies, and Growth', *American Economic Review*, vol. 90(4) (September), pp. 847–68.

Burnside, C. and Dollar, D. (2004). 'Aid, Policies, and Growth: Revisiting the Evidence', Policy Research Paper O-2834, The World Bank, Washington, DC (March).

Chang, C. C., Fernandez-Arias, E. and Serven, L. (1998). 'Measuring Aid Flows: A New Approach', Working Paper No. 387, Inter-American Development Bank, Washington, DC (December).

Chauvet, L. and Guillaumont, P. (2002). 'Aid and Growth Revisited : Policy, Economic Vulnerability and Political Instability', paper present at the Annual Bank Conference on Development Economics: Towards Pro-poor Policies, Oslo (June).

Clemens, M., Radelet, S. and Bhavnani, R. (2004). 'Counting Chickens When They Hatch: The

Short-Term Effect of Aid on Growth', Working Paper 44, Center for Global Development, Washington, DC.

Collier, P. and Dehn, J. (2001). 'Aid, Shocks, and Growth', Working Paper 2688, The World Bank, Washington, D.C. (October).

Collier, P. and Dollar, D. (2002). 'Aid Allocation and Poverty Reduction', *European Economic Review*, vol. 45(1) (September), pp. 1–26.

Collier, P. and Dollar, D. (2004). 'Development Effectiveness: What Have We Learnt?', *The Economic Journal*, vol. 114(496) (June), pp. F244–71.

Collier, P. and Hoeffler, A. (2002). 'Aid, Policy and Growth in Post-Conflict Societies', Policy Research Working Paper 2902, The World Bank, Washington, D.C.

Dalgaard, C. and Hansen, H. (2001). 'On Aid, Growth and Good Policies', *Journal of Development Studies*, vol. 37(6) (August), pp. 17–41.

Dalgaard, C., Hansen, H. and Tarp, F. (2004). 'On the Empirics of Foreign Aid and Growth', *The Economic Journal*, vol. 114(496) (June), pp. F191–F216.

Dehn, J. (2000). 'Commodity Price Uncertainty in Developing Countries', Working Paper 12, Centre for the Study of African Economies, Oxford (May).

Development Assistance Committee (DAC) (2002). *Development Assistance Committee Online,* Paris.

Durbarry, R., Gemmell, N. and Greenaway, D. (1998). 'New Evidence on the Impact of Foreign Aid on Economic Growth', CREDIT Research Paper 98r8, University of Nottingham.

Easterly, W. (2001). *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics*, Cambridge, MA: MIT Press.

Easterly, W. and Levine, R. (1997). 'Africa's Growth Tragedy: Policies and Ethnic Divisions', *Quarterly Journal of Economics*, vol. 112(4) (November), pp. 1203–50.

Easterly, W. and Levine, R. (2003). 'Tropics, Germs, and Crops: How Endowments Influence Economic Development', *Journal of Monet*ary *Economics*, vol. 50(1) (January), pp. 3–39.

Easterly, W., Levine, R. and Roodman, D. (2004). 'New Data, New Doubts: A Comment on Burnside and Dollar's "Aid, Policies, and Growth" (2000)', *American Economic Review*, vol. 94(2) (June).

Easterly, W. and Rebelo, S. (1993). 'Fiscal Policy and Economic Growth: An Empirical Investigation', *Journal of Monet*ary *Economics*, vol. 32(3) (December), pp. 417–58.

Gallup, J. L., and Sachs, J. D. (1999). 'Geography and economic development', in (B. Pleskovic and J. E. Stiglitz, eds.), *Annual World Bank Conference on Development Economics, 1998*

Proceedings, pp. 127–78, Washington, DC: The World Bank.

Griffin, K. B. and Enos, J. L. (1970). 'Foreign assistance: Objectives and consequences', *Economic Development and Cultural Change*, vol. 18(3), pp. 313–27.

Guillaumont, P. and Chauvet, L. (2001). 'Aid and Performance: A Reassessment', *Journal of Development Studies*, vol. 37(6) (August), pp. 66–92.

Hadi, A. S. (1992). 'Identifying multiple outliers in multivariate data', *Journal of the Royal Statistical Society*, Series B 54, pp. 761-777.

Hadjimichael, M. T., Ghura, D., Muhleisen, M., Nord, R. and Ucer, E. M. (1995).'Sub-Saharan Africa: Growth, Savings, and Investment, 1986–93', Occasional Paper 118, International Monetary Fund, Washington, DC.

Hancock, G. (1989). *Lords of Poverty: The Power, Prestige, and Corruption of the International Aid Business,* New York: Atlantic Monthly Press.

Hansen, H. and Tarp, F. (2000). 'Aid Effectiveness Disputed', *Journal of International Development*, vol. 12(3) (April), pp. 375–98.

Hansen, H. and Tarp, F. (2001). 'Aid and Growth Regressions', *Journal of Development Economics*, vol. 64(2) (April), pp. 547–70.

Holtz-Eakin, D., Newey, W. and Rosen, H. S. (1988). 'Estimating vector autoregressions with panel data', *Econometrica*, vol. 56(6) (November), pp. 1371–95.

International Monetary Fund (IMF*)* (2003). *International Financial Statistics* database, Washington, DC (July).

Jepma, C. J. (1991). *The Tying of Aid,* Paris: OECD Development Centre.

Kaufmann, D., Kraay A. and Mastruzzi, M. (2003). 'Governance Matters III: Governance Indicators for 1996–2002', Policy Research Working Paper 3106, The World Bank, Washington, DC (August).

Killick, T. (1998). *Aid and the Political Economy of Policy Change,* London: Routledge.

King, R. G. and Levine, R. (1993). 'Finance and Growth: Schumpeter Might be Right', *American Economic Review*, vol. 108(3) (August), pp. 717–37.

Knack, S., and Keefer, P. (1995). 'Institutions and Economic Performance: Cross-Country Tests Using Alternative Institutional Measures', *Economics and Politics*, vol. 7(3) (November), pp. 207–27.

Leamer, E. E. (1983). 'Let's Take the Con out of Econometrics', *American Economic Review*, vol. 73(1) (March), pp. 31–43.

Levine, R. and Renelt, D. (1992). 'A Sensitivity Analysis of Cross-Country Growth Regres-

sions', *American Economic Review*, vol. 82(4) (September), pp. 942–63.

Lensink, R. and White, H. (2001). 'Are There Negative Returns to Aid?', *Journal of Development Studies*, vol. 37(6) (August), pp. 42–65.

Lu, S. and Ram, R. (2001). 'Foreign Aid, Government Policies, and Economic Growth: Further Evidence from Cross-country Panel Data for 1970–93', *Economia Internazionale*, vol. 54(1) (February), pp. 15–29.

Mankiw, N. G., Romer, D. and Weil, D. N. (1992). 'A Contribution to the Empirics of Economic Growth', *American Economic Review*, vol. 107(2) (May), pp. 407–37.

Mosley, P., Hudson, J. and Horrell, S. (1987). 'Aid, the Public Sector and the Market in Less Developed Countries', *Economic Journal*, vol. 97(387) (September), pp. 616–41.

Ram, R. (2004). 'Recipient Country's 'Policies' and the Effect of Foreign Aid on Economic Growth in Developing Countries: Additional Evidence', *Journal of International Development*, vol. 16(2) (March), pp. 201–11.

Reusse, E. (2002). *The Ills of Aid: An Analysis of Third World Development Policies*, Chicago: University of Chicago Press.

Rodríguez, F. and Rodrik, D. (2001). 'Trade Policy and Economic Growth: A Skeptic's Guide to the Cross-National Evidence', in (B. Bernanke and K. S. Rogoff, eds.), *NBER Macroeconomics Annual*, Cambridge, MA: MIT Press.

Roeder, P.G. (2001). 'Ethnolinguistic Fractionalization (ELF) Indices, 1961 and 1985', <http//:weber.ucsd.edu\~proeder\elf.htm>, accessed May 2004.

Ruud, P. A. (2000). *Classical Econometrics,* New York: Oxford University Press.

Sachs, J. D. (2001). 'Tropical underdevelopment', Working Paper W8119, National Bureau of Economic Research, Cambridge, MA.

Sachs, J. D. (2003). 'Institutions don't rule: direct effects of geography on per capita income', Working Paper W9490, National Bureau of Economic Research, Cambridge, MA.

Sachs, J. D. and Warner, A. (1995). 'Economic reform and the process of global integration', in *Brookings Papers on Economic Activity*, pp. 1–118, Washington, DC: The Brookings Institution.

Sala-I-Martin, X. X. (1997). 'I Just Ran Two Million Regressions', *American Economic Review*, vol. 87(2) (May), pp. 178–83.

Summers, R. and Heston, A. (1991). 'The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950–88', *Qu*a*rterly Journal of Economics*, vol. 106(2) (May), pp. 327–68.

Svensson, J. (1999). 'Aid, Growth and Democracy', *Economics and Politics*, vol. 11(3) (November), pp.275–97.

Van de Walle, N. and Johnston, T. A. (1996). *Improving Aid to Africa,* Policy Essay No. 11, Washington, DC: Overseas Development Council.

U.K. Department for International Development (DFID) (2000). *Eliminating World Poverty: Making Globalisation Work for the Poor*, White Paper on International Development Presented to Parliament by the Secretary of State for International Development by Command of Her Majesty, London (December).

U.S. Department of State (various years). *World Milit*ar*y Expenditures and* Ar*ms Transfers,* Washington, DC.

Wacziarg, R. and Welch, K. H. (2002). 'Trade Liberalization and Growth: New Evidence', mimeo, Stanford University (November).

Windmeijer, F. (2000). 'A Finite Sample Correction for the Variance of Linear Two-step GMM Estimators', Working Paper 00/19, Institute for Fiscal Studies, London.

World Bank (2003). *World Development Indicators 2003* database, Washington, DC.