

Where Policy Experiments are Conducted in Economics and Political Science: The Missing Autocracies*

Graeme Blair[†] Radha K. Iyengar[‡] Jacob N. Shapiro[§]

This version May 20, 2013.

*The authors thank Eli Berman, Jim Berry, Chris Blattman, Jon Bendor, Michael Callen, Allan Dafoe, Jishnu Das, Angus Deaton, Thad Dunning, James Fearon, Don Green, Macartan Humphreys, Dean Karlan, Phil Keefer, David Laitin, Patrick Lam, David Nickerson, Dan Posner, Dustin Tingley, and Jeremy Weinstein for helpful conversations and feedback. Rajiv Lalla did a tremendous job compiling our initial data set, and Morris Breitbart provided excellent research assistance. All errors are our own. This research was supported, in part, by the U.S. Department of Defense Minerva Research Initiative through the Air Force Office of Scientific Research, grant #FA9550-09-1-0314 to Shapiro.

[†]Ph.D. candidate, Department of Politics, Princeton University, Princeton, NJ 08544; Email: gblair@princeton.edu, URL: <http://www.princeton.edu/~gblair>.

[‡]Senior Economist, RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202; E-mail: riyengar@rand.edu.

[§]Assistant Professor, Woodrow Wilson School and Department of Politics, Princeton University, Princeton, NJ 08544; Phone: 609-258-2256; Fax 609-258-0482, Email: gblair@princeton.edu, URL: <http://www.princeton.edu/~jns>. Corresponding author.

Abstract

Over the last 20 years economists and political scientists have rapidly adopted the experimental method. Randomized control trials (RCTs) are being used to study everything from election monitoring in Afghanistan to vocational training programs. This shift has paid substantial dividends in understanding how to address economic, political, and social challenges, but that knowledge has been developed primarily in places with certain governance characteristics. Using data from a nearly-complete enumeration of RCTs in political science and economics conducted since 1995, we show that the sample of RCT locations in non-OECD countries is skewed towards countries that are poorer than average, but that are also relatively democratic and spend more on citizen welfare. Only 20% of RCTs have been conducted in autocratic states, the kinds of places where roughly 35% of the worlds population lives. The RCT endeavor in economics and political science, in other words, provides a great deal of evidence about what works in poor countries that these fields theories predict will do well without intervention and has relatively less to say regarding how policies will work in poorly-governed, non-democratic environments.

INTRODUCTION

The movement in the social sciences to embrace the experimental methods of the natural sciences has expanded rapidly since the mid 1990s, well beyond the earlier exclusive reliance on laboratory experiments to address basic social research issues such as cooperation in groups and risk aversion. Scholars in this new tradition aim to address complex questions such as how to reduce poverty, prevent disease, improve government services, and avoid large-scale violence. To do so they use experiments to study actual programs designed to address these challenges. By introducing random assignment to the process of allocating benefits, they develop causal estimates of program effects which can be used to (1) make precise cost-benefit calculations, (2) help with subsequent decisions about how to better design similar programs, and (3) provide evidence for theories about social behavior. As is well known, these kinds of impact evaluations accumulate most readily into a corpus of evidence on how to address social problems for contexts similar in key respects to the ones in which they were conducted (Cartwright and Hardie, 2012).¹ While concerns with external validity are present for experiments of all kinds, the complexity of social phenomena — the interactions between political, economic, and psychological forces — conspires to diminish the likelihood that a policy found to be effective in one context will succeed in another. In this essay, we demonstrate that, in aggregate, the social scientists studying development draw their experimental locations not from the population of poor places (which includes many poor, unstable, undemocratic, and war-torn countries), but from the places easiest to work in — the poor places that are stable and have cooperative, democratic governments — who already spend a relatively large amount on pro-poor policies.²

The benefits of the experimental tradition to social science research are clear. Based on the experimental design, estimates of the causal effect of a policy change can be obtained for the subjects in a study, or a population from which a random sample of subjects is drawn. Scholars using this approach can dispense with some of the herculean statistical assumptions that are common in

¹Experiments can, of course, be informative for settings distinct from those where they are conducted if local conditions are not critical for mediating the treatment's impact. This may be particularly true of theory-based experiments that aim to test fundamental behavioral factors that may (or may not) vary across cultural contexts (Callen et al., Forthcoming).

²On the related fact that RCTs tend to be conducted with higher-quality partners within countries and its importance for interpreting studies see (Allcott and Mullainathan, 2012).

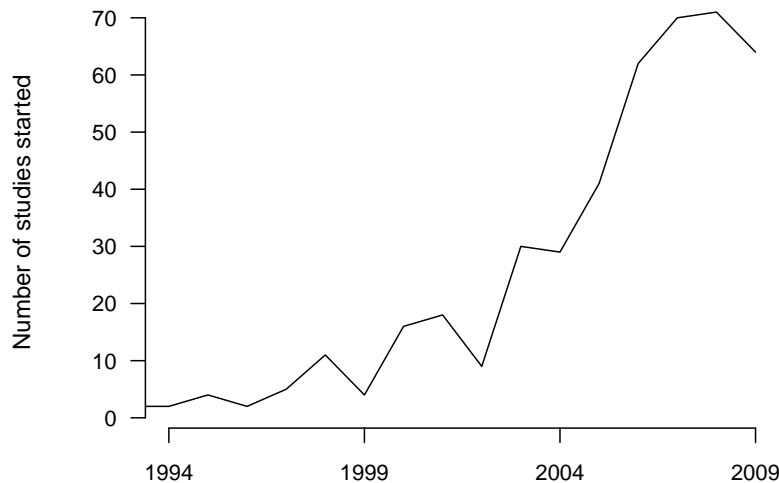


Figure 1: Number of randomized-control trials published in the top-25 economics and political science journals or listed on the websites of major organizations supporting RCTs in development economics by reported study start date.

analysis of observational data and are also freed from the problems of missing data which plague historical analyses of less developed countries. These virtues prompted the enormous growth in the use of randomized control trials across the social sciences between 1995 and 2009 depicted in Figure 1, which shows the number of policy experiment-based studies started in non-OECD countries by year.³ The rate increased from 10 per year in 2000 and earlier to over 70 per year in 2009, and covers questions ranging from how to monitor elections in Afghanistan (Callen and Long, 2012) to the impact of community-driven development programs on public-mindedness in Liberia (Fearon, Humphreys and Weinstein, 2009), and how to increase firm productivity in India (Bloom et al., 2011).⁴

Despite this uptake there is substantial disagreement among social scientists about the magnitude of the contribution field experiments make to the body of social scientific knowledge, with criticisms originating with James Heckman (1991), but renewed recently and forcefully by An-

³Our sample, described in detail below, records 640 RCTs published in top-25 economics and political science journals, reported in working papers on the National Bureau of Economic Research website, or described on the websites of CEGA, IGC, IPA, J-PAL, and 3iE (the five major organizations supporting RCTs in development economics).

⁴The decrease in the rate of studies in the data from 2008 to 2009 is not because the actual rate of studies starting slowed, but because studies often start several years before they make it into the some of our data sources.

gus Deaton (2009). Critics issue three challenges. The first is the extent to which a small scale RCT can be useful for understanding a scaled up program or larger policy (LaLonde, 1986; Iyengar, 2010). In particular, specific implementors or conditions that allow small scale successes may not be generalizable to larger, broader programs. A second concern is the extent to which RCTs conducted on a limited sample can isolate differential treatment effects across the population. The concern is that heterogeneity in treatment effects may be particularly important if groups most likely to take part in the program once it is scaled up differ significantly from those included in the RCT. This is analogous to the problem in medical trials of patients self-selecting into trials (Malani, 2008) and non-random subject recruitment (Robinson et al., 1996). A third concern is the extent to which estimated effects can be generalized to other settings such as different countries or time periods (Humphreys and Weinstein, 2009). These three concerns are often in combination referred to as the “external validity” of a study, meaning the study’s applicability outside of the particular program and sample.

The counter-argument of some field experiment promoters is that with sufficient replication the reliably-estimated causal effects of programs can be extrapolated beyond the contexts of individual studies to the relevant populations of vulnerable places and people (Imbens, 2009). Two of the most prominent proponents of this movement write,

A single experiment does not provide a final answer on whether a program would universally “work.” But we can conduct a series of experiments, differing in either the kind of location in which they are conducted or the exact intervention being tested (or both). Together, this allows us to both verify the robustness of our conclusions (Does what works in Kenya also work in Madagascar?) and narrow the set of theories that can explain the data. (Banerjee and Duflo, 2011, p. 14)

Thus although any individual study is subject to the skeptics critiques, proponents claim, the literature as a whole can provide invaluable generalizable insights. We agree, subject to the assumption that experiments are conducted across the relevant set of places.

It is in this respect that one central critique of RCTs has some merit. The set of places RCTs are being done differs dramatically from the universe of cases afflicted by the economic and governance maladies being studied.⁵ Only 20% of all RCTs in our data were conducted in autocracies (measured

⁵This problem is not unique to RCTs, the sample of all studies in economics similarly departs from being representative of the world population or of the population of states; it is biased towards wealthier countries with publications in top journals being strongly biased towards the United States (Das and Do, 2012).

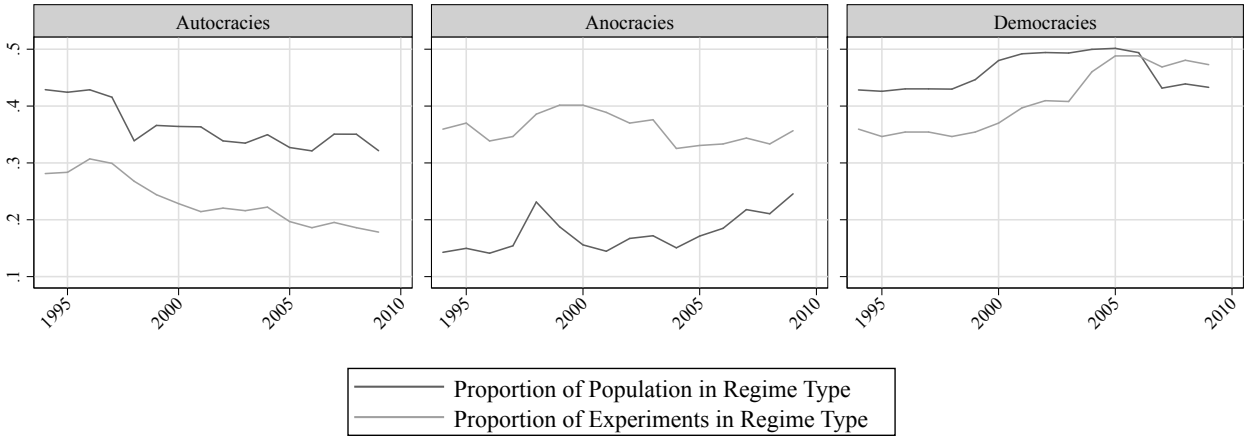


Figure 2: Proportion of randomized-control trials in each regime type by reported study start date plotted with proportion of world population living in each regime type. Following standard practice autocracies are places scoring below -5 on the Polity IV autocracy-democracy scale, democracies are places above 5, and anocracies are all places in between.

using the standard Polity IV coding), yet 35% of the world’s population lives in such countries. This difference has been quite consistent as the world has democratized, as figure 2 shows by plotting the proportion of the world’s population living in different regime types by year from 1995 through 2009 against the proportion of the RCTs started in that year conducted in that type of regime.⁶ Moreover, there is a striking concentration of research on specific topics within countries. Zambia, for example, has 10 RCTs on public health and one on education while Sri Lanka has 9 on banking and micro-enterprise and one on education. If we believe that governance quality has a strong interactive relationship with a broad range of potential policy interventions, and that governance status is correlated with where RCTs are being done, then we face the problem of interactive confounders described in Leamer (2010), for which randomization is not a useful solution.

The fact that the sample of social science RCT studies departs from a random sample of countries should not come as a surprise and is not necessarily a problem. The differences in how policies work in autocratic vs. democratic states is surely more important for some policies than for others and from both inferential and practical standpoints, there are economies of scale from working in a known environment repeatedly over time. Moreover, many RCTs require either the

⁶Critically, the autocracy gap has remained at about 15% since 1995, which equates to 40% underrepresentation, suggesting that the natural growth of the RCT endeavor is not ameliorating the potential problem we identify.

support or direct involvement of the government entities.⁷ Getting such support is difficult if not impossible in settings with weak governance structures or highly volatile politics and is much easier in settings where public servants are committed to helping their constituents. The macro-implications of the rational project selection by individual researchers, however, have not been previously documented.

The goal of this article is therefore to highlight the sampling traits of the RCT endeavor in economics and political science. In some ways, this is a natural extension of the discussion *Political Analysis* recently hosted about using various forms of pre-publication research registration to improve empirical work (Humphreys, de la Sierra and van der Windt, 2013; Monogan III, 2013; Laitin, 2013). While those scholars focused on how to minimize the bias in presenting results conditional on having done research, we seek to draw attention to potential issues arising from how researchers select study locations.

To diagnose the sample of places being studied, we collected data on nearly all RCT studies in these fields and match these to social, economic, and political conditions at the country-level. We find that the sample of RCTs in less developed countries (defined as all countries that are not part of the Organization for Economic Co-operation and Development - OECD) is skewed towards places that are poorer than average but that are also relatively well-governed given their level of economic and social development.

One way to think about whether this fact is a problem is to note that most of the time what scholars studying policies care about is not the absolute performance of different programmatic alternatives, but their ranking. Clearly almost any kind of reform in an unstable, conflict-ridden society will work worse *ceteris paribus* than the same reform in a well-governed place. But, for policies where either (a) the absolute magnitude of effects are important or (b) theory leads one to expect the returns to differ substantially across levels of governance, the current distribution is potentially problematic.

⁷For reasons why cooperation from government officials is important and example of how political changes can make conducting RCTs hard, see Briceño, Cuesta and Attanasio (2011).

EMPIRICAL APPROACH

We collected basic data on 640 RCTs conducted outside the United States that were either (1) published in top-25 economics or political science journals since 1995 as well as the National Bureau of Economic Research working paper series or (2) listed on the websites of CEGA, J-PAL, IPA, 3iE, or the World Bank. While surely an undercount of the set of RCTs in progress, the results were quite striking. The distribution is illustrated graphically in figure 3. Of these 30% occurred in just 3 countries—India (91), Kenya (56), and Mexico (45)—and among non-OECD countries, almost 60% occurred in the 13 countries which had 15 or more RCTs.

We also collected data on a range of social, economic, and political dimensions in all non-OECD countries so that we could compare the distribution of key variables in developing countries against that in the RCT study sample. We start by comparing without weights to check whether the RCT sample is representative of countries around the world, and then weight by share of total Official Development Assistance (ODA), to check if RCTs are representative for the distribution of ODA; share of world population, to see if they are representative population-wise; and share of people making less than \$2/day, to see if they are representative of the world's poor. To motivate the dimensions of comparison we identified five major topics which account for over 90% of RCTs: economic initiatives (35.8%), health programs (25.31%), education programs (20.0%), governance programs (5.0%), and agriculture/environmental programs (2.97%). Of the 35.8 % of economic RCTs, the topics include banking and microfinance (75%), and poverty alleviation (15%), and labor and employment programs (10%). We therefore collected data on income (measured in terms of PPP-adjusted GDP per captial) from the Penn World Tables Version 7.0 Heston, Summers and Aten (2011), health (measured by infant mortality and HIV prevalence 15-49) from the World Bank Development Indicators Group (2012), education spending as a share of government revenues Group (2012), and governance using the Polity IV measure of regime type Marshall, Jaggers and Gurr (2010) and the proportion of government spending on salaries vs. that on welfare spending Group (2012).

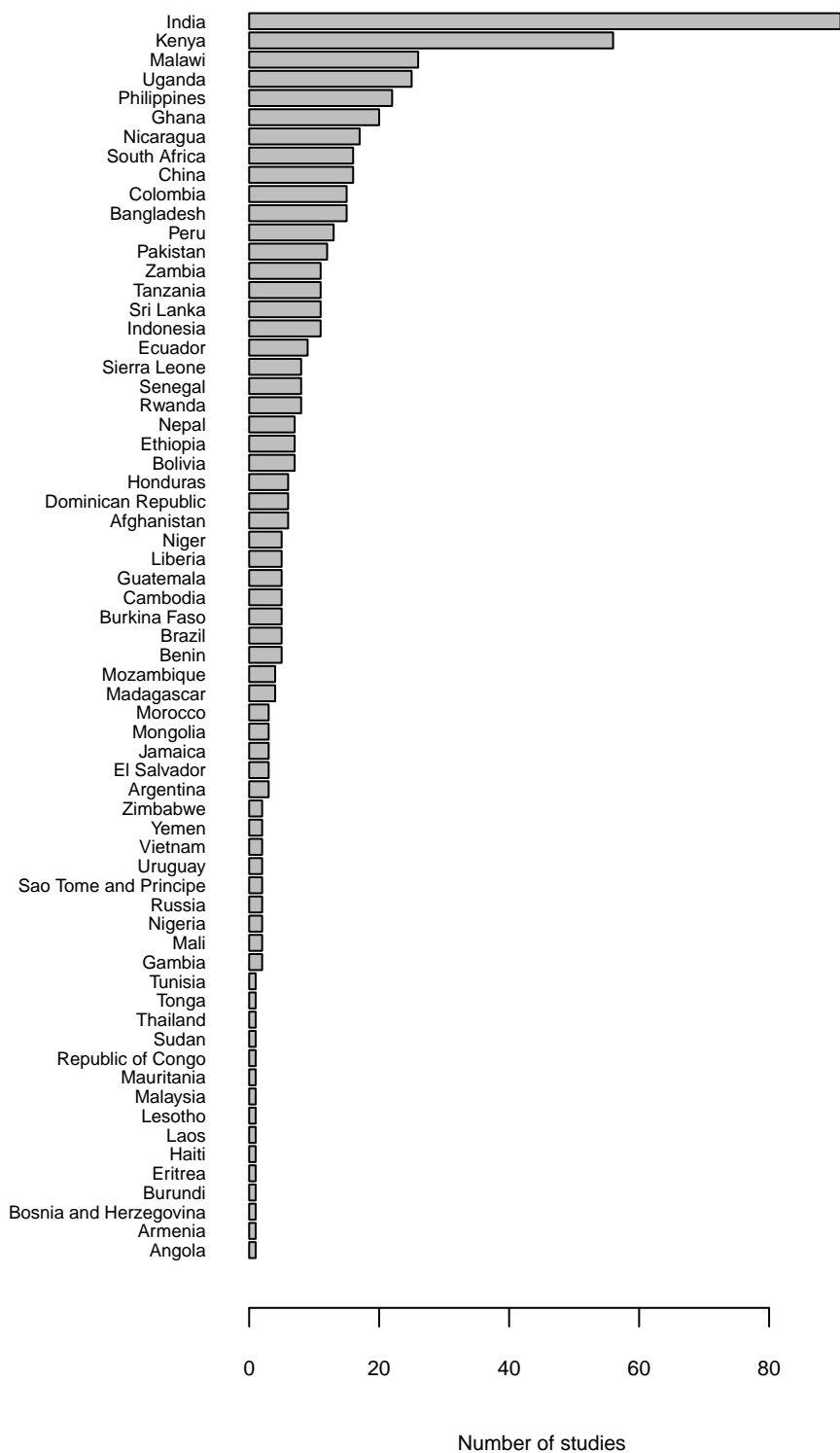


Figure 3: Number of RCTs per non-OECD country published in the top-25 economics and political science journals or listed on the websites of major organizations supporting RCTs in development economics. Fully 27% of RCTs were conducted in either India or Kenya.

FINDINGS

We report our key findings using a series of figures. The left panel of each figure helps visualize whether the RCTs started in any given year are drawn from a representative distribution of that variable worldwide in that year. An RCT in country A started in year Y is a dot at the value of that variable in that year. The inter-quartile range of that variable for all non-OECD countries in that year is depicted with a grey band. The right panel of each plot shows the probability density function of that variable in 2008 for all non-OECD countries (dashed line) compared to the distribution of that variable for all country-years in which RCTs were started (solid line) and all observational studies in economics journals (dotted line). Because it does not make sense that one should have the same number of RCTs in Gambia as in India (the latter has a large population than all of sub-Saharan Africa), we weight the density plots by country population averaged over 1994 to 2009.⁸

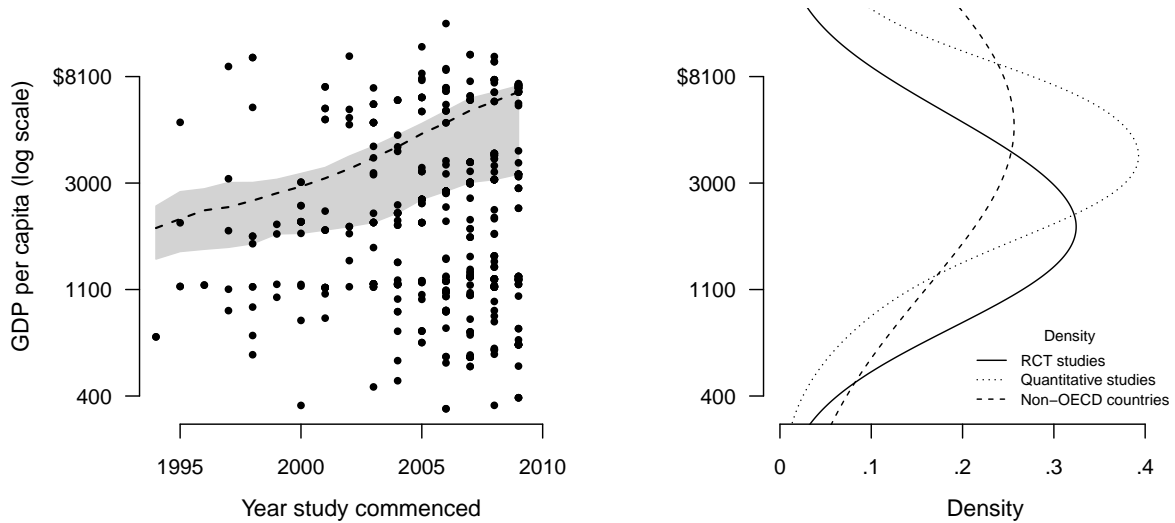


Figure 4: Left panel shows inflation-adjusted GDP per capita by RCT each year with the grey band showing the interquartile range of GDP per-capita in all non-OECD countries. Right panel shows the distribution of inflation-adjusted GDP per-capita in 2009 for all non-OECD countries vs. the distribution in the RCT sample and that in the non-RCT study sample.

⁸In supplementary materials, we demonstrate that running the analysis without weighting or with weighting by share of total ODA or share of the world's poor do not substantially change the results discussed here.

Three facts are apparent in these plots. First, as figure ?? shows, RCTs are concentrated in countries that are poorer than the average non-OECD country. The median GDP per capita of non-OECD countries is \$5003 per person per year while the median for RCT countries is approximately \$2838, the 35th percentile of the full non-OECD distribution. Second, RCTs have somewhat higher levels of infant mortality and HIV than non-OECD countries as a whole (figure S10 and figure S11).⁹

Third, and most notably, RCTs are concentrated in countries that are more democratic than the average non-OECD country, especially given their level of income, as figure 5 shows. The polity score is striking because unlike for health or income, where one might think that oversampling from the lower tail of the distribution is consistent with the use of RCTs to inform the development project, oversampling from the upper tail of the governance distribution means the majority of studies are being conducted in places that most current theories would predict should do quite well in the long-run (see e.g. (Acemoglu and Robinson, 2012)). This bias is less present in non-RCT studies which under-sample the middle of the polity distribution but do a better job of covering autocracies.

This bias towards democratic places becomes even more striking if we look at where RCTs are being done as a function of both GDP and Polity Score. As table 1 shows, RCTs in Economics and Political Science are under-sampling autocracies of all kinds and are dramatically oversampling poor democracies. Each panel report the distribution of studies across income quartiles for one of the three standard regime types used by political scientists: autocracies with Polity IV scores below -5; anocracies with Polity IV scores in $[-5, 5]$; and democracies with Polity IV scores above 5. Within each panel, the first column reports the proportion of all RCTs being done in that type of country, the second column reports the proportion of all non-OECD countries that fit in that cell, and the third column reports the ratio of these two, which can be taken as a measure of the sample bias in the administration of RCTs. If the sample were representative of all states, for example, the ratios would all be close to one. In the bottom two cells of the last column, however, we see that democracies in the bottom two income quartiles are overrepresented by 204 and 326% respectively.

⁹If RCT countries are drawn from the lower tail of the health distribution they may overestimate the challenges to solving health problems and the suggested solutions may thus entail a misallocation of resources.

Table 1: Administration of RCTs by Regime Type and GDP

Panel A: Autocracies			
GDP quartile	Proportion of all RCTs	Proportion of all non-OECD Countries	RCT Sample Bias
75th pctile - Max	0.03	0.07	0.39
Median - 75th pctile	0.03	0.07	0.39
25th pctile - Median	0.02	0.06	0.28
Min to 25th pctile	0.00	0.03	0.14
Panel B: Anocracies			
GDP quartile	Proportion of all RCTs	Proportion of all non-OECD Countries	RCT Sample Bias
75th pctile - Max	0.00	0.04	0.10
Median - 75th pctile	0.00	0.04	0.09
25th pctile - Median	0.03	0.09	0.27
Min to 25th pctile	0.14	0.13	1.09
Panel C: Democracies			
GDP quartile	Proportion of all RCTs	Proportion of all non-OECD Countries	RCT Sample Bias
75th pctile - Max	0.07	0.09	0.81
Median - 75th pctile	0.17	0.14	1.17
25th pctile - Median	0.28	0.09	3.26
Min to 25th pctile	0.19	0.09	2.04

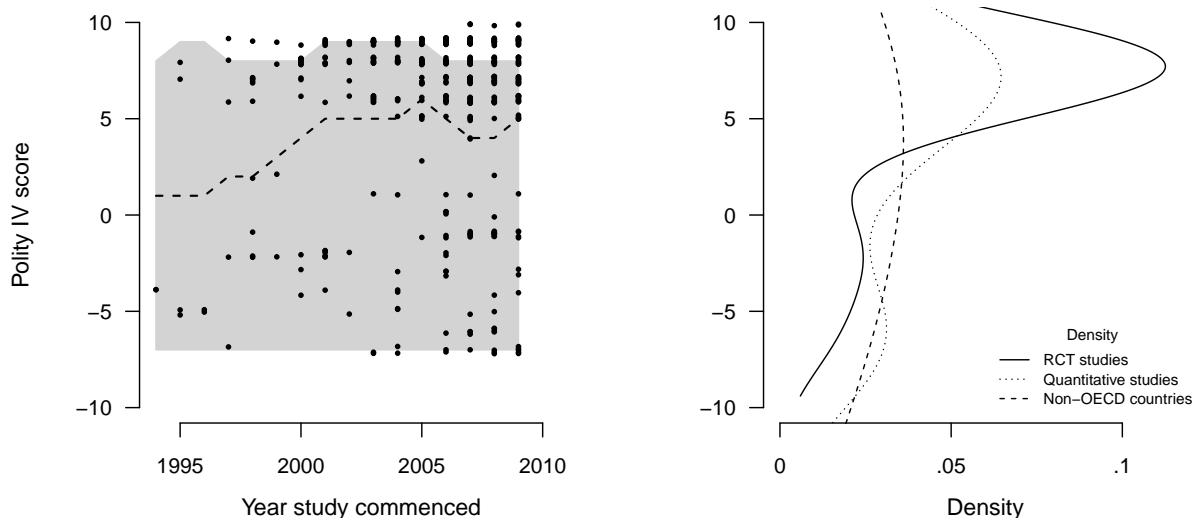


Figure 5: Left panel shows Polity IV scores by RCT each year with the grey band showing the interquartile range of all non-OECD countries. Right panel shows the distribution of Polity IV scores in 2009 for all non-OECD countries vs. the distribution in the RCT sample and that in the non-RCT study sample.

In summary, RCTs are being conducted in places that score higher on standard measures of democratic governance than the average OECD country but have lower levels of per-capita GDP and slightly worse health outcomes. That RCTs are more likely to be done in places with stable, relatively well-functioning governments means we are learning about behavior under a particular set of parameter values with loss of generality to more poorly governed settings. This sampling pattern makes it difficult to use the cumulated evidence from RCTs to draw conclusions about what can be done to improve economic and political outcomes in poor countries with poor governance.

Further, within the 10 countries that have more than 15 RCTs, the distribution of RCTs across topics clearly diverges from a representative one. Figure 6 shows this by plotting the number of studies in each of the six topics that comprise more than 90% of RCTs: banking and micro-enterprise; public health; education; poverty alleviation; governance; labor and employment; elections, and agriculture. For any sampling scheme to rationalize this distribution, it would have to be one that sought to learn about specific topics in specific countries. Doing so is valuable for helping to choose the right policies for that place, but leads to an aggregate distribution that is hard to square with the larger RCT endeavor.

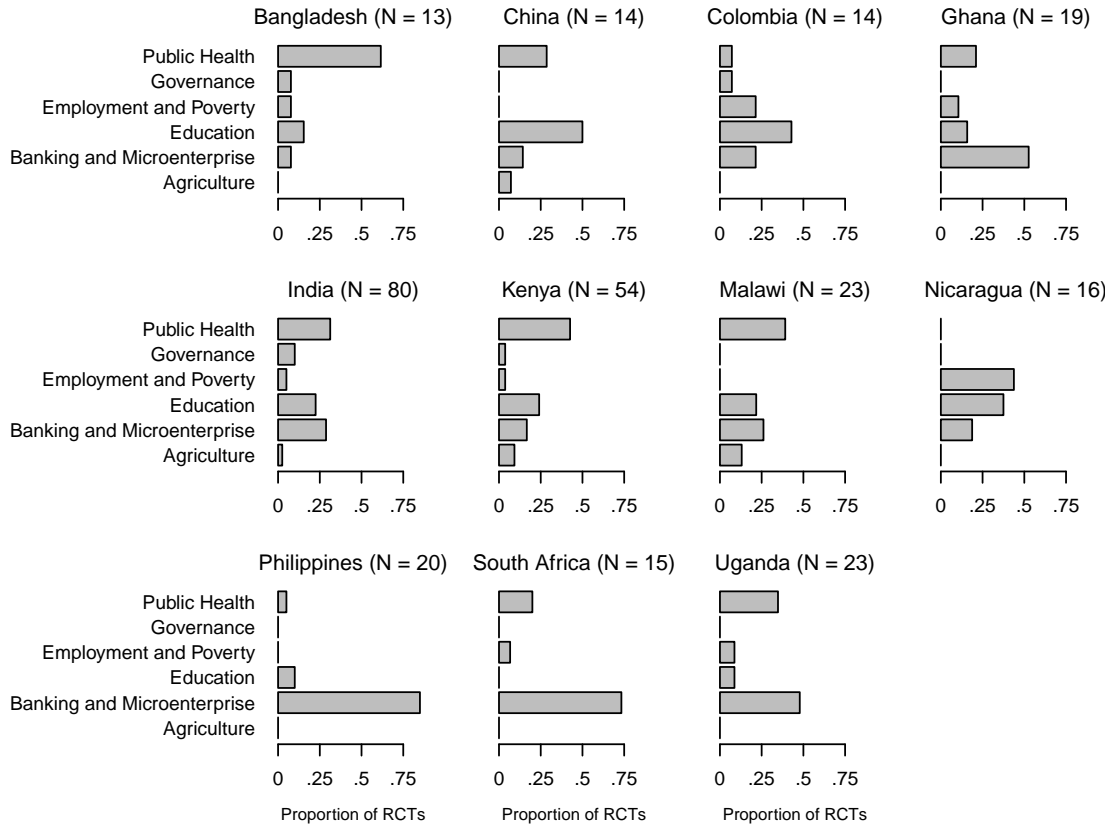


Figure 6: Each panel shows the distribution of experiments by topic in one country for the top-10 countries by total number of RCTs.

IMPLICATIONS

These issues with RCTs in policy research are similar to those seen in using controlled experiments to guide the introduction of new medical technologies.¹⁰ In the medical research a similar sampling problem arises due to comorbidity factors in study participants Aapro et al. (2005); Fortin et al. (2006); Van Spall et al. (2007).¹¹ Many trials exclude those with health problems other than that for which the treatment being studied is intended (Masoudi et al., 2003; Van Spall et al., 2007). This can obviously lead to a range of inferential problems, particularly when the population that is likely to receive the treatment in regular practice tends to suffers from multiple ailments and those ailments, or the medications given for them, influence the efficacy of the treatment in question.

¹⁰We thank Angus Deaton for this analogy.

¹¹The most famous example of this came to light with the voluntary recall of the selective COX-II inhibitor, VIOXX. In that study the researchers failed to account for differential attrition due to the side-effects of the drug, which is related but not exactly the phenomenon of interest here. See supporting materials for more details.

In the context of social science research poor governance can be thought of as a comorbidity factor for poor economic conditions and low growth potential. Poorly governed places are often excluded from trials for perfectly sensible reasons; officials and NGO partners in these settings can be quite unpredictable when it comes to teaming with researchers. There is modest evidence that RCTs in economics and political science are in fact being done in places where governments are more committed than average to the public well-being. Using World Bank data we find that countries with RCTs spend less on government salaries than the average non-OECD country, but spend more on welfare and education. Figure 7 highlights the fact that few RCTs are being done in places that spend little on education.¹²

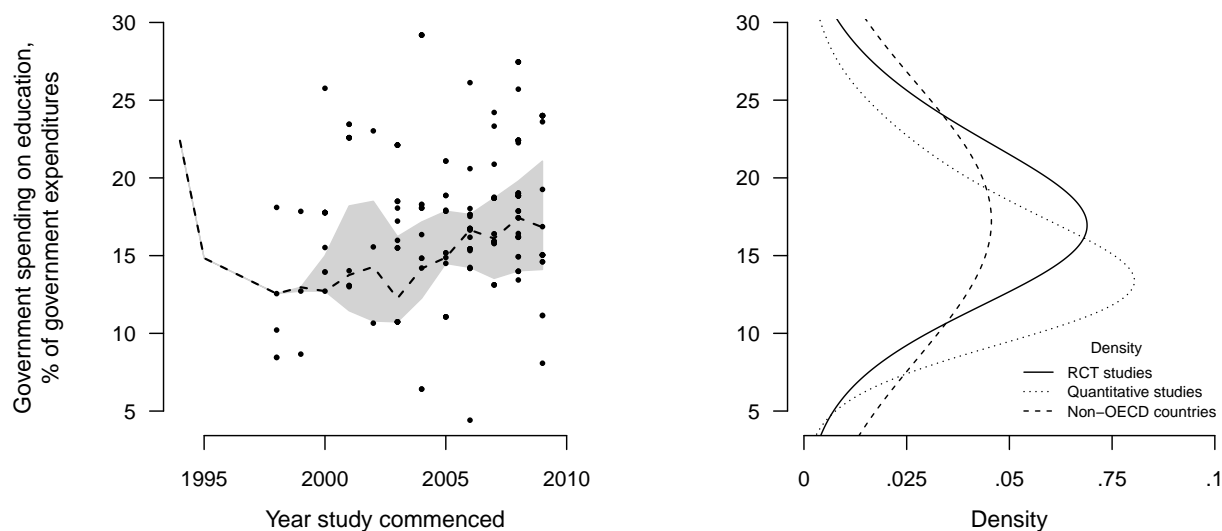


Figure 7: Left panel shows government spending on education as a proportion of all government spending in RCT countries by year with the grey band showing the interquartile range of all non-OECD countries. Right panel shows the distribution of education spending in 2009 for all non-OECD countries vs. the distribution in the RCT sample and that in the non-RCT study sample.

The potential consequences of this the pattern we illustrate can be understood with an example of a program that worked well in one place and then failed in another (Cartwright, 2010; Cartwright and Hardie, 2012). Beginning in 1980 the Tamil Nadu Integrated Nutrition Project (TINP) took a novel approach to improving the health of poor children by integrating modest food support

¹²For government spending on salaries see appendix S13 and on welfare appendix S14.

to children with education for pregnant mothers. The project was led to a major reduction in malnutrition in Tamil Nadu. When the program was replicated in Bangladesh, however, it had no impact on malnutrition. Researchers think this is because mothers of children were the main decision-makers when it came to food in Tamil Nadu, but not in Bangladesh where mother-in-laws tended to make those decisions and men tended to do the shopping (White, 2009). Taken together, that meant that providing nutritional education to pregnant mothers was not giving the information to those who could take advantage of it, as it had in India. One can easily imagine similar subtle differences meaning that programs which work in public-goods providing democracies would fail when applied in more authoritarian settings.

CONCLUSION

Overall, the move into the experimental method in economics and political science has yielded tremendous benefits. These fields are on firmer empirical footing now than ever before and in many countries policies are being implemented more effectively because of feedback developed through RCTs. For the experimental move in social science to realize its full potential, though, we believe more attention needs to be paid to where studies are being done. The most appropriate analogy is to the ongoing debate in the medical field on populations included in the RCTs for new drugs and technologies. As discussed above, drug trials typically exclude participants with co-morbidity factors as well as limiting their samples by age and sometimes gender or ethnicity. This is perfectly reasonable sample if the goal is to produce clean estimates of an effect as a first step in understanding the mechanism by which a drug works or in verifying that it can be useful, the analogues to the theory-based or proof-of-concept RCT in social science. But if few of the patients who will eventually need these drugs fall into such a pool (most sick people are older and have multiple problems), then the research is not estimating the effect of the drug or technology on the populations for whom it matters most. The same problem applies, we argue, to how social science RCTs are being conducted.

Baring the ability of social scientists to find an institutional solution which has so far eluded clinical scientists, we argue they should more explicitly look for consilience, or lack thereof, between RCTs and well-documented observational studies. Ironically, two of the most prominent supporters

of RCTs have written a very influential book (Banerjee and Duflo, 2011) that is, in many ways, a model of consilience. It combines causal relationships observed in multiple RCTs with non-causal observations from a range of studies to support broad hypotheses about the core challenges to improving poor peoples' lives in developing countries. More research that follows this model would help make the most of the resources being allocated to policy RCTs by social scientists.

REFERENCES

- Aapro, Matti S., Claus-Henning Köhne, Harvey Jay Cohen and Martine Extermann. 2005. "Never Too Old? Age Should Not Be a Barrier to Enrollment in Cancer Clinical Trials." *The Oncologist* 10(3):198–204.
URL: <http://theoncologist.alphamedpress.org/content/10/3/198.abstract>
- Acemoglu, Daron and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York: Crown Publishers.
- Allcott, Hunt and Sendhil Mullainathan. 2012. External Validity and Partner Selection Bias. Working paper 18373 National Bureau of Economic Research.
- Banerjee, Abhijit V. and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. Perseus Book Group.
- Bloom, Nicholas, Effert, Mahajan, McKenzie and Roberts. 2011. Does Management Matter? Evidence from India. Working paper National Bureau of Economic Research.
- Bombardier, Claire, Loren Laine, Alise Reicin, Deborah Shapiro, Ruben Burgos-Vargas, Barry Davis, Richard Day, Marcos Bosi Ferraz, Christopher J. Hawkey, Marc C. Hochberg, Tore K. Kvien and Thomas J. Schnitzer. 2000. "Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis." *New England Journal of Medicine* 343(21):1520–1528.
URL: <http://www.nejm.org/doi/full/10.1056/NEJM200011233432103>
- Briceño, Bertha, Laura Cuesta and Orazio Attanasio. 2011. "Behind the Scenes: Managing and Conducting Large-Scale Impact Evaluations in Colombia." *Journal of Development Effectiveness* 3(4):470–501.
- Callen, Michael and James D. Long. 2012. Institutional Corruption and Election Fraud: Evidence from a Field Experiment in Afghanistan. Working paper UC San Diego.
- Callen, Michael, Mohammad Isaqzadeh, James D. Long and Charles Sprenger. Forthcoming. "Violent Trauma and Risk Preference: Experimental Evidence from Afghanistan." *American Economic Review* .
- Cartwright, Nancy. 2010. Will this policy work for you? Predicting effectiveness better: How philosophy helps. Presidential address psa 2010 LSE and UCSD.
- Cartwright, Nancy and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press.
- Das, Jishnu and Quy-Toan Do. 2012. U.S. and Them: The Geography of Academic Research. Working paper World Bank.
- Deaton, Angus. 2009. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Working paper Center for Health and Wellbeing, Princeton University.
- Fearon, James D., Macartan Humphreys and Jeremy M. Weinstein. 2009. "Can Development Aid Contribute to Social Cohesion after Civil War? Evidence from a Field Experiment in Post-conflict Liberia." *American Economic Review* 99(2):287–91.

- Fortin, Martin, Jonathan Dionne, Geneviève Pinho, Julie Gignac, José Almirall and Lise Lapointe. 2006. "Randomized Controlled Trials: Do They Have External Validity for Patients With Multiple Comorbidities?" *The Annals of Family Medicine* 4(2):104–108.
URL: <http://www.annfammed.org/content/4/2/104.abstract>
- Group, Development Research. 2012. *World Bank Development Indicators*. World Bank.
- Heckman, James J. 1991. Randomization and Social Policy Evaluation. Working paper no. 107 National Bureau of Economic Research.
- Heston, Alan, Robert Summers and Bettina Aten. 2011. *Penn World Table Version 7.0*. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.
- Humphreys, Macartan and Jeremy M. Weinstein. 2009. "Field Experiments and the Political Economy of Development." *Annual Review of Political Science* 12:367–78.
- Humphreys, Macartan, Raul Sanchez de la Sierra and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21(1):1–20.
- Imbens, Guido. 2009. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). Working paper Harvard University.
- Iyengar, Radha. 2010. Does Arrest Deter Violence? Comparing Experimental and Nonexperimental Evidence on Mandatory Arrest Laws. In *The Economics of Crime: Lessons for and from Latin America*, ed. Rafael Di Tella, Sebastian Edwards and Ernesto Schargrotsky. University of Chicago Press pp. 421–452.
- Juni, Peter, Linda Nartey, Stephan Reichenbach, Rebekka Sterchi, Paul A Dieppe and Matthias Egger. 2004. "Risk of cardiovascular events and rofecoxib: cumulative meta-analysis." *The Lancet* 364(9450):2021–2029.
- Laitin, David D. 2013. "Fisheries Management." *Political Analysis* 21(1):42–47.
- LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4):604–620.
- Leamer, Edward E. 2010. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives* 24(2):31–46.
- Malani, Anup. 2008. "Patient enrollment in medical trials: Selection bias in a randomized experiment." *Journal of Econometrics* 144(2):341–351.
- Marshall, Monty G., Keith Jagers and Ted Robert Gurr. 2010. Polity IV Project: Political Regime Characteristics and Transitions, 1800-2010. Technical report University of Maryland.
- Masoudi, FA, EP Havranek, P Wolfe, CP Gross, SS Rathore, JF Steiner, DL Ordin and HM Krumholz. 2003. "Most hospitalized older persons do not meet the enrollment criteria for clinical trials in heart failure." *American Heart Journal* 146(2):250–7.
- Monogan III, James E. 2013. "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections." *Political Analysis* 21(1):21–37.

- Mukherjee, D, SE Nissen and EJ Topol. 2001. "Risk of cardiovascular events associated with selective COX-2 inhibitors." *Journal of the American Medical Association* 286(8):954–9.
- Murphy, L, J Bolen, CG Helmick and TJ Brady. 2009. Comorbidities Are Very Common Among People With Arthritis. Poster 43 National Conference on Chronic Disease Prevention and Control, CDC.
- Robinson, Delbert, Margaret G. Woerner, Simcha Pollack and Gail Lerner. 1996. "Subject Selection Biases in Clinical Trials: Data From a Multicenter Schizophrenia Treatment Study." *Journal of Clinical Psychopharmacology* 16(2).
- Van Spall, Harriette G C, Andrew Toren, Alex Kiss and Robert A. Fowler. 2007. "Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review." *JAMA: The Journal of the American Medical Association* 297(11):1233–1240.
URL: <http://dx.doi.org/10.1001/jama.297.11.1233>
- White, Howard. 2009. "Theory-based impact evaluation: principles and practice." *Journal of Development Effectiveness* 1(3):271–284.

SUPPORTING MATERIALS

.1. Additional Plots

This section first provides plots presented in the main text in which the summaries of the key covariates for non-OECD countries are re-weighted, first unweighted, then by the total amount of overseas development assistance averaged over the same period, and finally by proportion of the world's population making under \$2 per day. We then provide additional measures of government performance and show how the RCT and non-RCT study samples differ from that population. Finally, we replicate Table 1 from the main text in graphic form.

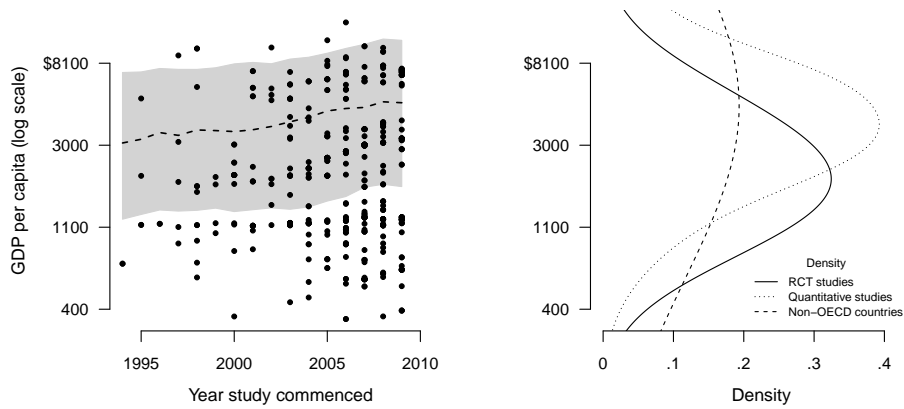


Figure S1: GDP weighted by population average, 1994 - 2009

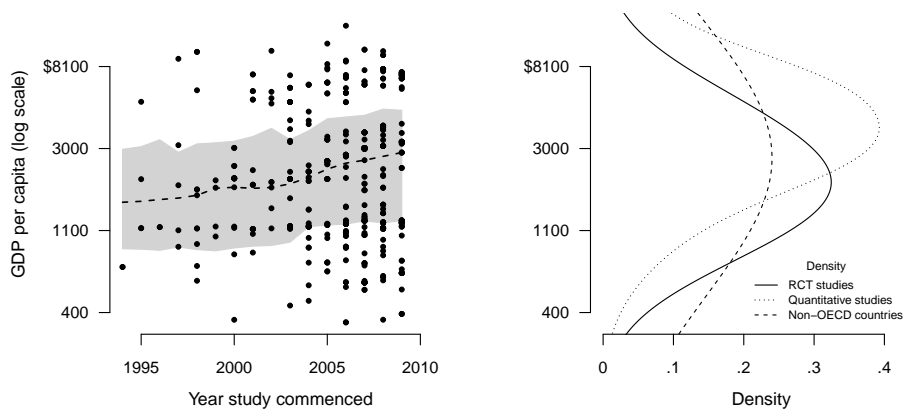


Figure S2: GDP weighted by total ODA average, 1994 - 2009

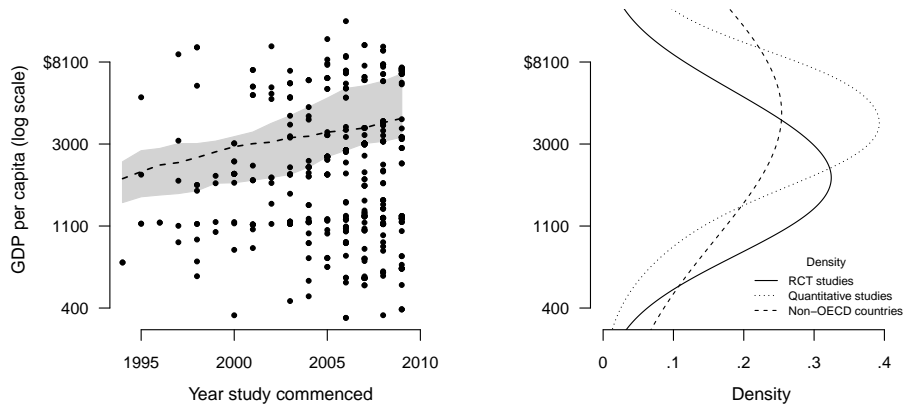


Figure S3: GDP weighted by proportion of population earning ≤ 2 /day, 1994 - 2009

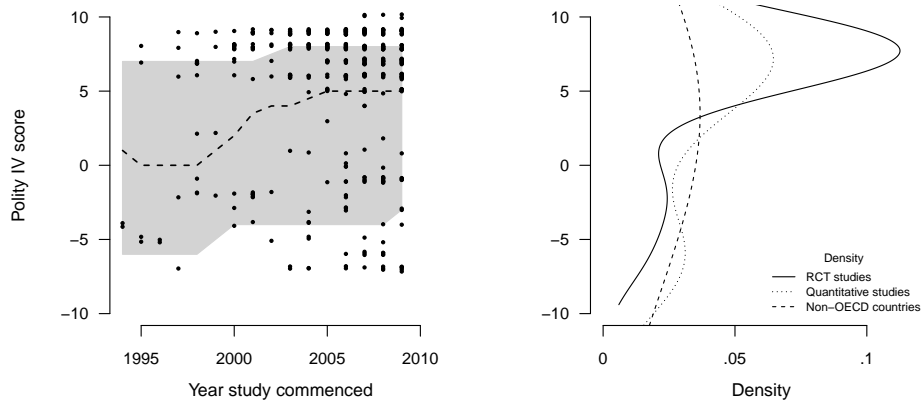


Figure S4: Polity weighted by population average, 1994 - 2009

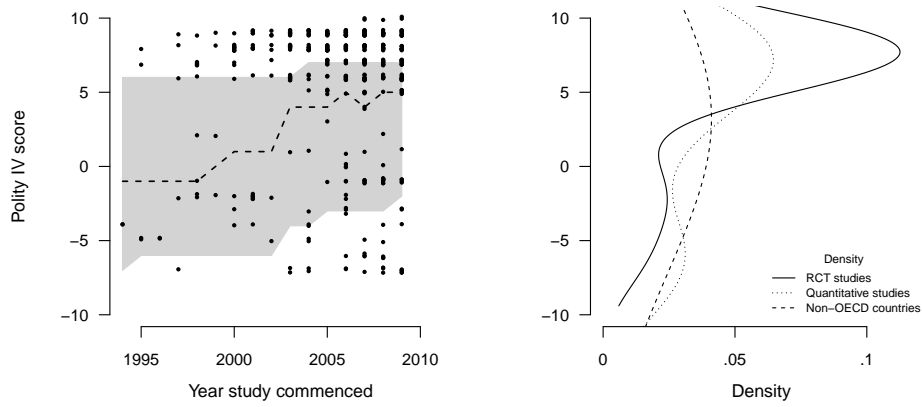


Figure S5: Polity weighted by total ODA average, 1994 - 2009

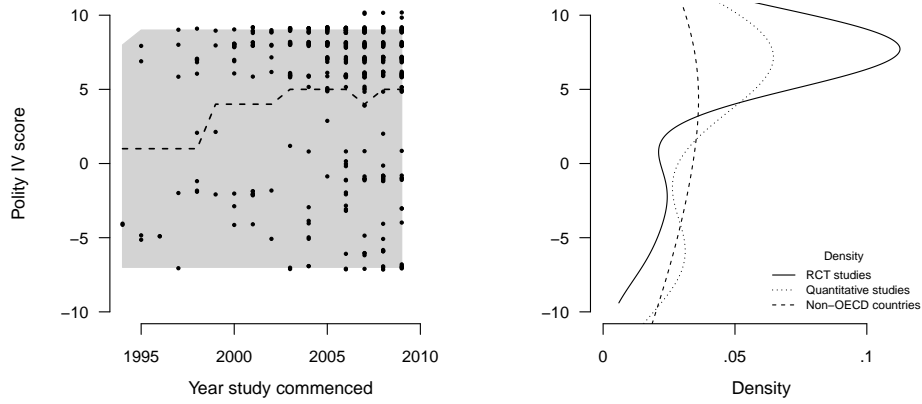


Figure S6: Polity weighted by proportion of population earning <\$2/day, 1994 - 2009

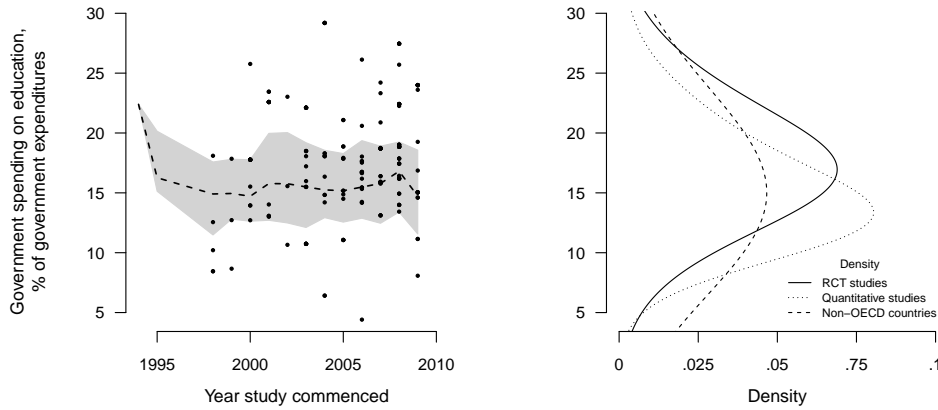


Figure S7: Education expenditures weighted by population average, 1994 - 2009

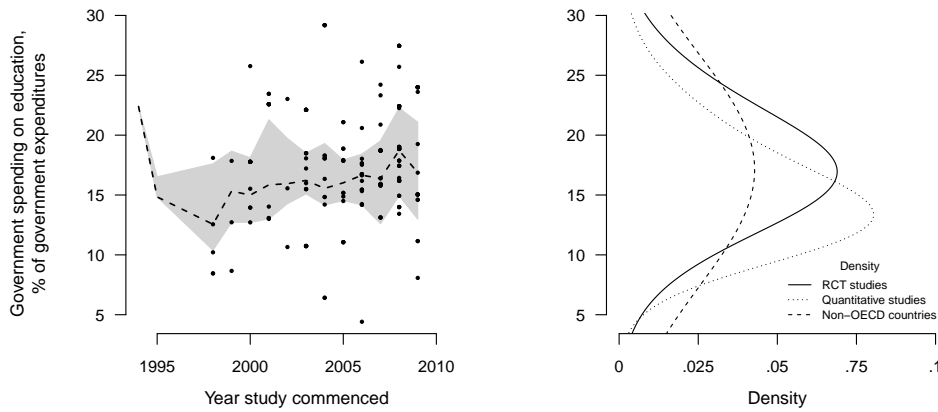


Figure S8: Education expenditures weighted by total ODA average, 1994 - 2009

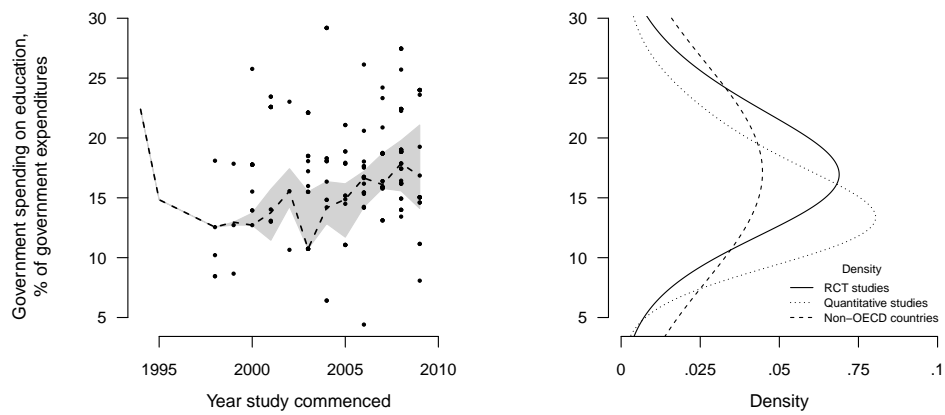


Figure S9: Education expenditures weighted by proportion of population earning <\$2/day, 1994 - 2009

This section provides additional figures showing the distribution of RCTs along other dimensions.

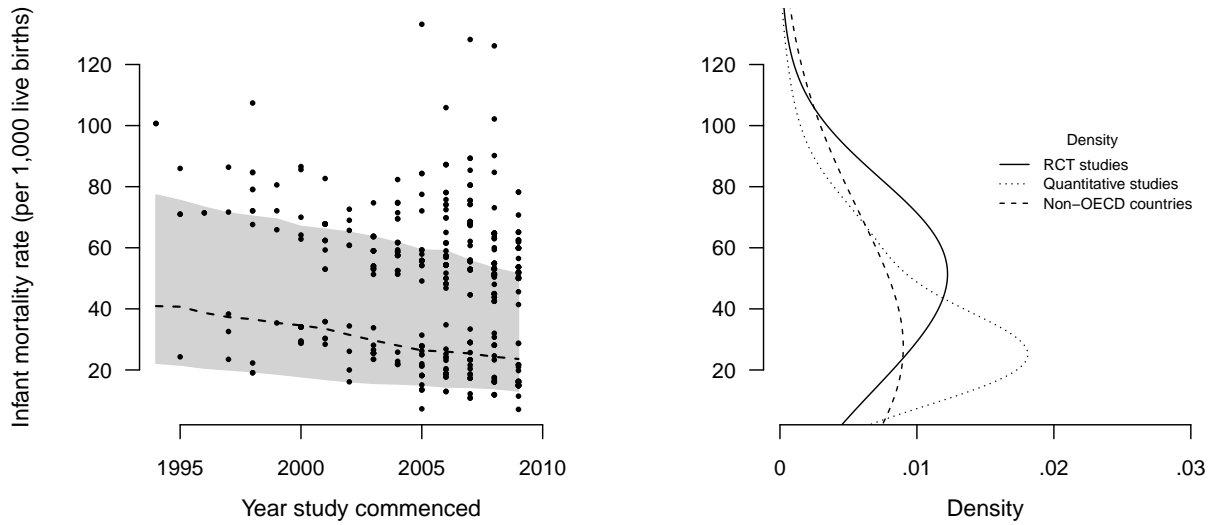


Figure S10: Left panel shows infant mortality rates by RCT each year with the grey band showing the interquartile range of all non-OECD countries. Right panel shows the distribution of infant mortality rates in 2009 for all non-OECD countries vs. the distribution in the RCT sample and that in the non-RCT study sample.

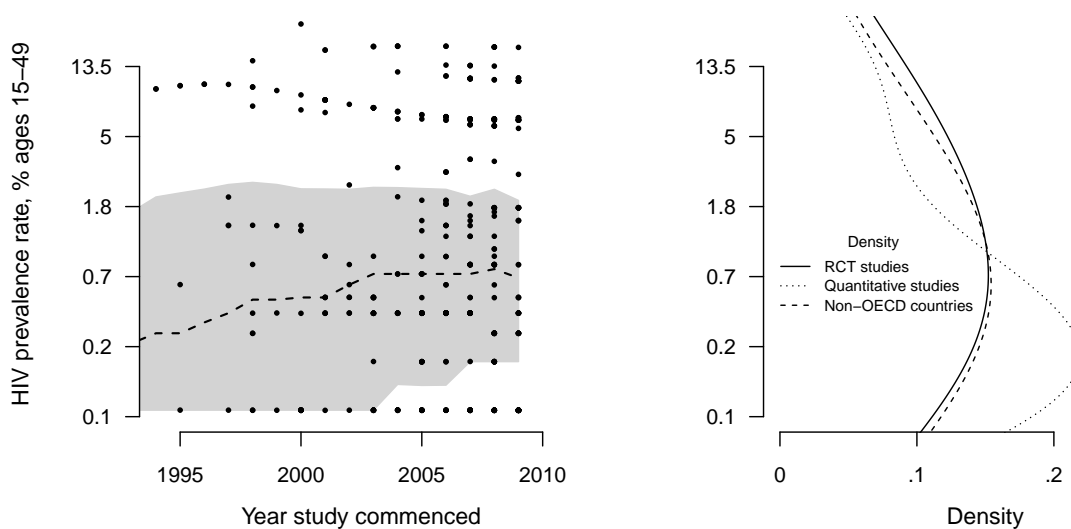


Figure S11: Left panel shows HIV infection rates by RCT each year with the grey band showing the interquartile range of all non-OECD countries. Right panel shows the distribution of HIV infection rates in 2009 for all non-OECD countries vs. the distribution in the RCT sample and that in the non-RCT study sample.

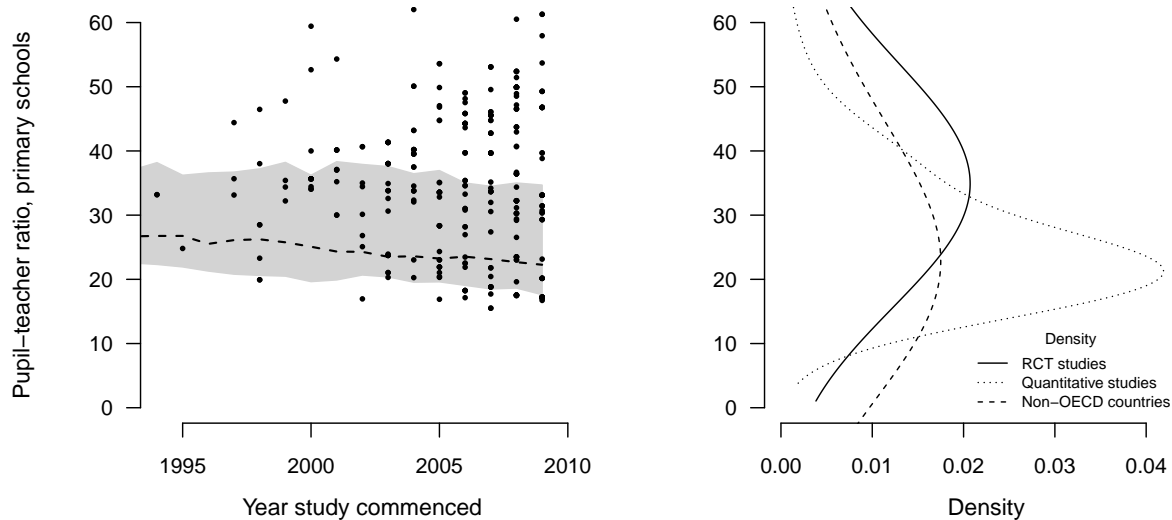


Figure S12: Left panel shows student-to-teacher ratio by RCT each year with the grey band showing the interquartile range of all non-OECD countries. Right panel shows the distribution of student-to-teacher ratios in 2009 for all non-OECD countries vs. the distribution in the RCT sample and that in the non-RCT study sample.

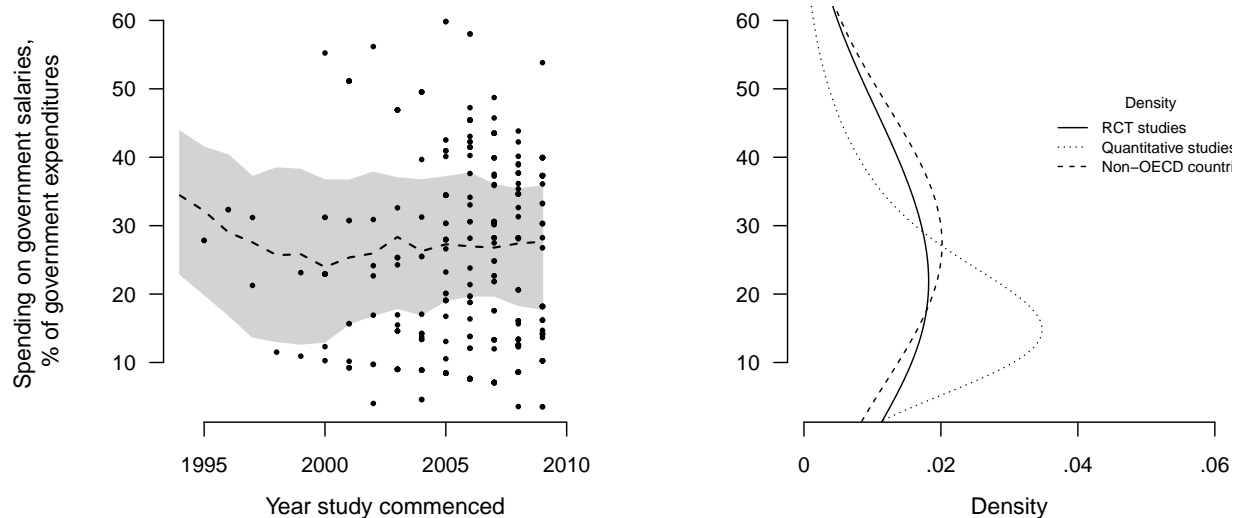


Figure S13: Left panel shows spending on government salaries as a proportion of government expenditures by RCT each year with the grey band showing the interquartile range of all non-OECD countries. Right panel shows the distribution of salary expenditures in 2009 for all non-OECD countries vs. the distribution in the RCT sample and that in the non-RCT study sample.

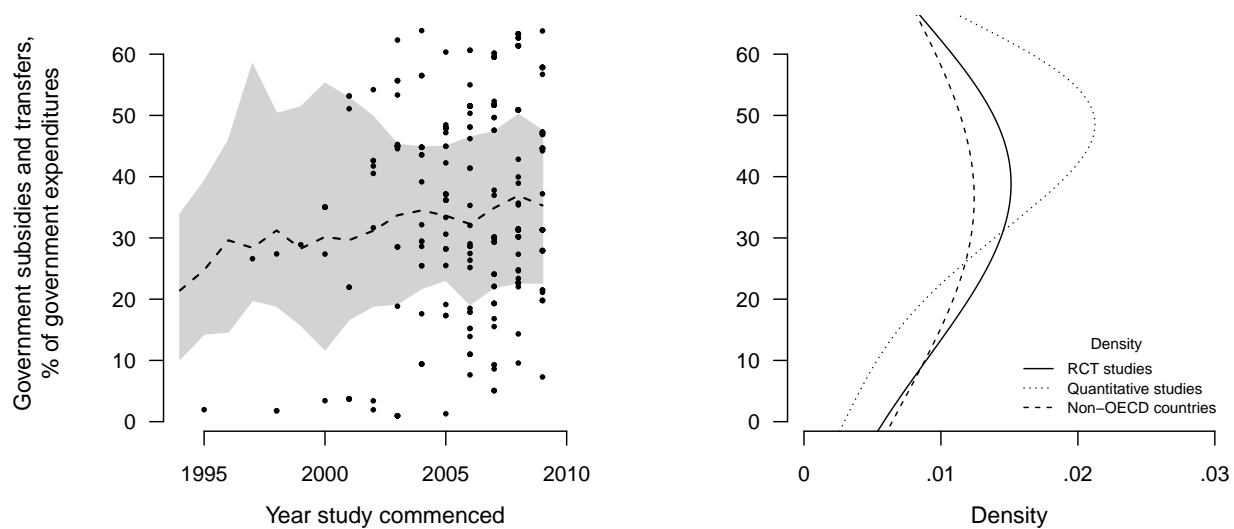


Figure S14: Left panel shows welfare spending as a proportion of government expenditures by RCT each year with the grey band showing the interquartile range of all non-OECD countries. Right panel shows the distribution of welfare expenditures in 2009 for all non-OECD countries vs. the distribution in the RCT sample and that in the non-RCT study sample.

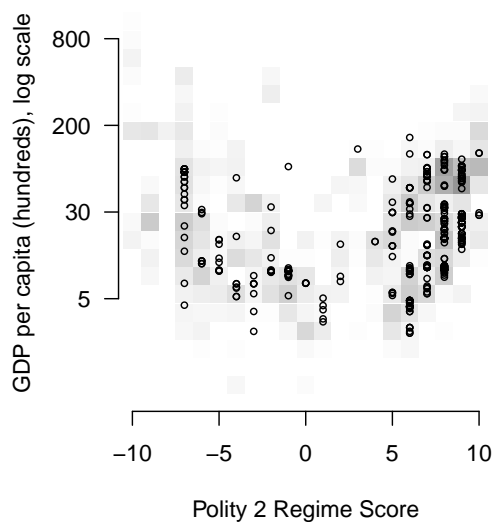


Figure S15: The x-axis shows GDP and the y-axis shows Polity IV scores for 2009 for all non-OECD countries. Each cell is colored by the density of the population of states in that GDP \times Polity cell. Each circle is an RCT, with the RCTs jittered to show the density of RCTs.

.2. *A Short Note on the VIOXX Trial*

The voluntary recall and subsequent release of information on the selective COX-II inhibitor, VIOXX, provides a striking example of what can go wrong due to differential attrition. Initial data from clinical trials of the drug released by Merck indicated only minor differences in adverse events (such as serious cardiovascular thrombotic events, e.g. heart attacks) between treated and untreated patients, leading to FDA approval in 1999. Over the next four years, more than 90 million prescriptions were filled for the drug. Tragically, it turns out that when individuals initially in the treatment group but later excluded due to comorbidity factors were also considered, VIOXX usage led to a large increase in side-effects (Mukherjee, Nissen and Topol, 2001). For the average patient actually taking VIOXX, the risks of heart attack was, in fact, four times higher than for a patient taking the most common alternative (Bombardier et al., 2000).

Why did this happen? It turns out that there were, in fact, significantly more withdrawals in the VIOXX group for hypertension, edema and other pathological circulatory concerns than in the control group (Juni et al., 2004). This differential attrition resulted in the appearance of no greater risk for heart attacks when used in the general population. Why were so many excluded? Because over 40 percent of those suffering from arthritis (the primary users of VIOXX) also suffer from hypertension (Murphy et al., 2009). In the context of social science, because RCTs are more likely to fail in places with weak governance, potentially leading to similar errors.