Experimentation at Scale *

Karthik Muralidharan† Paul Niehaus‡
UC San Diego UC San Diego

July 31, 2017

Abstract

This paper makes the case for greater use of randomized experiments "at scale". We review various critiques of experimental program evaluation in developing countries, and discuss how experimenting at scale along three specific dimensions – the size of the sampling frame, the number of units treated, and the size of the unit of randomization – can help alleviate them. We find that program evaluation randomized controlled trials published over the last 15 years have typically been "small" in these senses, but also identify a number of examples – including from our own work – demonstrating that experimentation at much larger scales is both feasible and valuable.

*We thank David Lagakos, Abhijeet Singh, and Gordon Hanson (the editor) for many helpful discussions and suggestions; special thanks to Sandip Sukhtankar, our co-author on the AP Smartcards evaluation which has shaped many of our ideas on this topic. Sam Krumholz provided excellent research assistance.

†Department of Economics, University of California, San Diego, NBER, and J-PAL; kamurali @ucsd.edu.

‡Department of Economics, University of California, San Diego, NBER, and J-PAL; pniehaus@ucsd.edu.

The growing use of randomized field experiments to evaluate public policies has been one of the most prominent trends in development economics in the past fifteen years. These experiments have advanced our understanding in a broad range of topics including education, health, governance, finance (credit, savings, insurance), and social protection programs, as summarized in a recent handbook (Duflo and Banerjee, 2017). In this paper we argue that experimental evaluations could have a greater impact on policy if more of them were (literally) bigger. We believe this for two reasons.

First, governments (regrettably) often do not follow a process of testing prototypes and scaling up those that work. On the contrary, they often roll out new programs representing millions (or billions!) of dollars of expenditure with little evidence to say whether they will work. Randomizing these rollouts can generate direct evidence on policy questions that are inarguably of interestafter all, they are already heavily funded. Working with governments to evaluate these programs as they are being deployed, and before political constituencies have calcified around them, thus represents a tremendous research opportunity with immediate policy applications.

Second, scale can help to improve "external validity, or the accuracy with which the estimates of impact from a randomized controlled trial (RCT) predict the effects of some subsequent policy decision. Critiques of the experimental movement have highlighted three substantial limits to external validity: (1) study samples may not be representative of the population that policy makers want to generalize their results to; (2) program effects may differ when implemented at smaller scale (say, by a highly motivated non-profit organization) and when implemented at a larger scale (typically by governments); and (3) the experiment may not capture important spillover effects, such as general equilibrium effects (for an overall discussion, see Deaton and Cartwright (2016) and the symposium in the Spring 2010 issue of this journal). Our goal here is not to re-litigate these well-known issues, but instead to highlight one way in which the field experimental literature can (and to some extent already *is*) making progress in addressing them through the use of larger-scale experiments.[1]

When we refer to "scale, our focus is on three specific dimensions in which experiments could be bigger, corresponding to the three threats to validity described above. First, experiments can be conducted in samples that – while not necessarily large themselves – are representative of large populations, addressing concerns about non-representative sampling. Second, experiments can evaluate the impacts of interventions that are implemented at a large scale, which addresses the concern that results will be different (likely worse) when the scale of the operation increases. Third, experiments can be randomized in large units

[1]In a similar vein, Fryer (2017) discusses several limitations of RCTs, and notes that several of these "can be sidestepped by running more, larger, and better-designed RCTs.

such as villages or regions. This enables researchers to test directly for spillovers such as to market prices and quantities, which might otherwise undermine external validity.

We begin this paper by documenting the scale of recent program evaluation experiments run in developing countries and published in top general interest journals over the last 15 years. We find they have typically been small in each of the three senses mentioned earlier: the median evaluation was representative of a population of 10,885 units, studied a treatment delivered to 5,340 units, and was randomized in clusters of 26 units per cluster. We then discuss some of the prominent exceptions, beginning as early as the landmark evaluations of the Progresa program rollout in Mexico (Gertler and Boyce, 2003; Schultz, 2004). We argue using these examples – and drawing on our own experiences over the past decade – that it is both feasible and valuable to conduct experimental evaluations at larger scales than has been the norm.

Of course, not all experiments should be big. Big experiments are expensive, time-consuming, and risky. Many experiments should stay small and present results with a clear discussion of where along the dimensions above the lack of scale does or does not limit the generalizability of their findings. In many cases, a sequence from small to large experiments, such as that proposed in this same symposium by Banerjee et al. (2016), will be best. In our closing section we discuss these tradeoffs, including some of the main organizational and financial considerations in enabling experimentation at scale, and how these constraints might be loosened in ways that could increase the possibilities for large-scale experimentation.

1 How big are recent experiments?

To ground the discussion in a set of basic facts, we collected measures of scale for all randomized controlled trials conducted in developing countries and published in five top general-interest journals (the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Review of Economic Studies*, and *Quarterly Journal of Economics*) from January 2001 to July 2016. We restricted our focus to experiments framed as program evaluations — that is, estimates of the impact of interventions that are candidates for large-scale implementation more or less as is — and excluded experiments framed as tests of theoretical mechanisms. Our substantive conclusions are not sensitive to how we categorize borderline cases. Appendix A1 describes the protocol for the exercise in more detail and provides a full list of studies included and excluded. We identified 29 experimental program evaluations to include in the exercise, with annual counts varying from zero to two each year from 2001-2007 and then from two to five each year from 2008-2016. These figure illustrate the upward trend in publication of experimental program evaluations, but also show that they remain a relatively

small share of total publications in top general interest journals.

1.1 The scale of the population represented

The frame from which an experimental sample is drawn may not be usefully representative of any broader population. This is obviously the case if the frame is not chosen at random, but instead reflects factors such as the availability of a willing implementing partner, researcher preferences, local demand for the intervention, and so on. Such factors can lead to biased estimates of treatment effects when seeking to extrapolate experimental treatment effects to the larger population of interest. For example, Allcott (2015) finds that the first evaluations of a US energy conservation initiative were conducted in sites with substantially higher average treatment effects than the overall average. But more broadly, even if the sampling frame is itself selected in a random or near-random fashion from some larger population, it may yield noisy measures of population parameters if it is itself small. Choosing one district at random from a country within which to test an intervention, for example, produces an estimate of mean impacts that is unbiased for the country-wide average treatment effect, but also very imprecise. As is well known, it is thus valuable to draw experimental samples from large frames (Heckman and Smith, 1995).

To measure the scale of experiments on this dimension, we code two metrics. First, we code an indicator for whether the study was conducted in a sample drawn randomly from *any* larger frame. For example, a study conducted in 10 villages selected at random from the list of villages in the district would be coded as a one, but a study conducted in 10 villages that are not randomly chosen would be scored as a zero. Second, we identify the size of the sampling frame whenever available. For studies that do not report drawing their analysis sample from a larger frame, the sampling frame is the same as the sample size; for those that do report a larger frame, we measure or estimate the size of this frame wherever possible. Overall we were able to estimate the size of the frame for 26 of 29 studies. The first two rows of Table 1 show summary statistics for these two measures, and Figure 1 plots the distribution of the (log of the) latter. Note that we measure size here and throughout by the number of primary units of analysis included in a set, where we define the primary unit is the unit at which the outcome(s) we believe are most important for the studys thesis are measured.[2]

[2]In many cases this measure is unambiguous: for example, in a study that measures the impacts of deworming drugs on individual people, we treat the individual as the base unit of analysis. In others the choice is less clear. For example, a study of incentives for teachers might measure both teacher outcomes and student outcomes, and we must then make a judgment call whether to count teachers, students, etc. as the primary unit of analysis. In these cases we use the tie-breaking rule described, selecting as the primary unit of analysis the unit from which the most important outcomes are collected (which in the example above

Generally speaking, the samples in the studies we reviewed are representative of small populations. Only 31 percent of the studies report sampling respondents from any population larger than the sample itself. Among the 26 studies that report the size of their sampling frame, the median frame contains 10,815 units, while the 75th percentile frame contains 46,418 units. These figures are obviously modest compared to tens or hundreds of millions of impoverished people in the countries in which the studies are run. There are notable exceptions to this rule, however, which we discuss further below. For instance, Alatas et al. (2012a) perform an experiment on poverty targeting in Indonesia on a representative sample of three large provinces in Indonesia; their study results are representative of a population of over 50,000,000 people.

1.2 The scale of implementation

The scale at which an intervention is implemented can matter if the quality of implementation, and thus the effect of treatment, varies with scale.[3] For example, implementing at larger scale spreads managerial oversight more thinly within a given organization, and may require a shift to entirely different organizations (e.g. governments) than the ones that initially developed and tested an intervention (e.g. NGOs). Deaton (2010) similarly worries that "the scientists who run the experiments are likely to do so more carefully and conscientiously than would the bureaucrats in charge of a full scale operation.

Indeed, recent research has documented large variation in organizational effectiveness. For example, Bold et al. (2013) discuss a teacher recruitment intervention that was highly cost-effective when a nongovernment organization ran a pilot study, and also when scaled up to the remaining sites managed by that nongovernment organization, but had no impact when scaled up further and run by the Kenyan government. In a non-experimental meta-analysis of experimental estimates, Vivalt (2015) finds that evaluations of an intervention tend to yield larger estimated effect sizes when the intervention is implemented by an nongovernmental organization as opposed to a government body. More broadly, the productivity literature finds wide dispersion in the productivity of firms (e.g. Hsieh and Klenow (2009)) and plants (e.g. Bloom et al. (2013)) producing relatively standardized products. Given these data, we see no prima facie case to focus solely on what intervention to deliver, and ignore the scale and scalability of the organization delivering it.

would typically be students).

[3]Medical researchers draw a similar distinction between efficacy, or impact under ideal conditions, and effectiveness, or impact under a set of real-world conditions. For example, the antibiotic regimens recommended for treating common strains of tuberculosis are known to be efficacious if closely adhered to, but can also be ineffective if not; adherence may depend on the patient, how the physician explains treatment to the patient, what monitoring protocols are put in place, etc.

To measure the scale of implementation, we record for each study the total number of units treated as part of the experiment. Importantly, this includes all units treated, not just those from whom outcome data were collected. Row 3 of Table 1 shows summary statistics for this measure, and Row 2 of Figure 1 plots its full distribution.

We find that the median study evaluated an intervention delivered to roughly 5,000 units. In the 75th percentile study, roughly 29,000 units were treated. As with frame size there are some substantial outliers. For example, Tarozzi et al. (2014) performed information interventions that had the potential to reach more than 40,000 households, although their primary treatment was more concentrated. But overall, it seems fair to say that most program evaluations have studied implementation at a scale which was modest compared to the scale on which the policies evaluated were (presumably) ultimately intended to be run.

1.3 The scale of units randomized

The size of the units randomized may matter because of spillovers, mechanisms through which an individuals outcomes depend not only on their own treatment status, but also on that of surrounding individuals (or households, firms, and so on). If spillovers are important, comparing outcomes for (randomly) treated and untreated neighbors will yield a doubly biased estimate of the average impacts of treating both, since it nets out spillovers from the treated to the untreated and also fails to capture the effects of spillovers from the untreated to the treated (as highlighted for example by Miguel and Kremer (2004).

Spillovers can arise for various reasons. There may be general equilibrium effects, where relative prices shift in response to treatment intensity (Deaton and Cartwright, 2016). For example, Cunha et al. (2015) find that transferring food to a large proportion of the residents of rural Mexican villages reduced the local price of food. As we discuss below, we find in our own work that improving a government employment scheme in Andhra Pradesh had effects on market prices and earnings much larger than the direct effects. There may also be political economy effects, where the behavior of rent-seeking groups changes in response to treatment intensity. For example, Bold et al. (2013) conjecture that one reason government implementation failed in their scaled-up evaluation of contract teachers in Kenya is that the teachers union mobilized to thwart the reform. In such cases it is difficult to extrapolate from the results of experiments conducted with small units of randomization to predict the results of full-scale implementation (Acemoglu, 2010).

When (as is often the case) spillovers decay with distance, experimenting with larger units can alleviate this concern. Suppose that the effects of a de-worming intervention spill over onto untreated households in the same village as treated ones, but not across villages. In this

example, randomizing the intervention within villages will produce estimates that are biased for the at-scale impact, but randomization across villages will produce unbiased estimates. More generally, if spillovers operate over some bounded distance, then randomizing in larger geographical units will reduce the proportion of each control unit that is affected by spillovers (as more units will be in the unaffected "interior of each control area), and analogously increase the proportion of treated units that are affected by spillovers from all their neighbors. This will in turn will shift the (expected) mean difference in outcomes between treated and control areas closer to the "total treatment effect a policy-maker would obtain from treating all units.

To measure the scale of randomization in our sample of studies, we code two metrics. The first measure is equal to 1 if the study randomizes at a level of aggregation greater than the primary unit of analysis and 0 otherwise. As above, we define the primary unit of analysis as the unit at which (in our judgment) the papers most important outcomes are measured. The second measure is the size of the average cluster randomly assigned to treatment or control, in number of primary analysis units. (While the geographic size of the average cluster is arguably a more useful metric than the number of units it contains, geographic size is not commonly reported.) Rows 4 and 5 of Table 1 report summary statistics for these two variables, and Row 3 of Figure 1 plots the full distribution of the (log of the) latter.

We find that randomization is commonly "clustered: 66 percent of the studies we reviewed randomized at a level of aggregation higher than the primary unit of measurement. At the same time, the units in which randomization is clustered are typically quite small: the median design featured 26 units per cluster. In fact, the largest mean unit of randomization we identified contained just 2,500 households (in Björkman and Svensson (2009)). Of course, the right cluster size – conceptualized as one which controls potential biases due to spillovers to an adequate level – is likely to be highly context and intervention-dependent. That said, the bulk of program evaluations have been conducted at scales of randomization at which general equilibrium effects, political economy effects, or other forms of spillovers are – if present – seem unlikely to be fully captured.

Overall, impact evaluation has for the most part been conducted at small scales – that is, in samples representative of small populations, with implementation for small groups, and with small units of randomization. We also examined whether this pattern has evolved over time by regressing each of the metrics above on calendar year, but found no evidence of a shift in either direction: none of the relationships we estimated were either statistically significant or economically meaningful.

2 Experimenting at scale: some examples

While impact evaluations have typically been small, there are a number of exceptions which demonstrate that it can be feasible and valuable to experiment at much larger scales. We develop this argument below, highlighting a number of experimental studies that evaluate programs at large scale in one or more senses of that term to illustrate the broad range of settings where this has been possible. For illustration we draw on lessons from our work over the past decade and in particular on work (joint with Sandip Sukhtankar) evaluating the introduction of a biometric payment system ("Smartcards) into two large anti-poverty programs in rural Andhra Pradesh (Muralidharan et al., Forthcoming, 2017). We were fortunate for this project to obtain government agreement to an experimental design that was "large relative to the distributions above in all three senses of the word – randomizing treatment across a population of 20M people, for example, and in clusters of 62,000 people.

2.1 Experiments in (near) representative samples of large populations

Conceptually, the benefits of conducting experiments in representative samples are well-understood. In practice, however, the data above suggest that few even among the best-published studies make a claim to be representative of larger populations. This leaves open the possibility of site-selection bias in the location of the experiment, or (even in the absence of bias) of imprecision due to the small number of sites.

To illustrate the potential importance of these issues, we conduct a simple exercise using data from the Smartcards evaluation, which was carried out across eight districts of Andhra Pradesh. One of our main findings was that Smartcards significantly reduced *average* levels of leakage the difference between government outlays and funds actually received by beneficiariesacross these eight districts. In Figure 2, we plot the mean treatment effect of Smartcards on leakage for each district *separately*, ordered by the magnitude of the effect. Notice that these district-specific effects vary widely. A study that evaluated Smartcards within any one district chosen at random would thus run a meaningful risk of producing unrepresentative results. Worse, a study that evaluated Smartcards in a district where (say) the government felt more confident in the prospects for a smooth implementation would very likely be biased.[4]

[4]In the online Appendix, Figure A1 and Table A4 offer a further illustration of this point by looking at the distribution of treatment effects that would be estimated if our study had only one randomly sampled district. Specifically, we simulate 500 experimental samples drawn from any one study district with the same number of sub-districts and sampled villages/households (sampled with replacement) and plot the distribution of treatment effects that would be obtained from such a study sample. As both Figure A1 (Panel B) and Table

While running experiments in samples that are representative of large populations may seem logistically challenging, such a protocol has been successfully implemented in multiple studies in Africa, South Asia, and Southeast Asia. For example, Muralidharan and Sundararaman (2010, 2011, 2013) first select a representative study sample of 600 primary schools across five districts of Andhra Pradesh (with a population over 10 million), and then randomly assign these to various treatments and a control group. Alatas et al. (2012a) first randomly sample 640 villages from three Indonesian provinces (population 50 million) and then randomly assign them to various treatments and a control group. Muralidharan and Sundararaman (2015) first sample a representative universe of villages with a private school (in the study districts), and then randomly assign villages into treatment and control status for studying a school choice program. De Ree et al. (2015) first construct a near-representative sample of 360 schools across 20 districts and all geographic regions of Indonesia and then randomly assign schools to receive accelerated access to a teacher certification program that led to a doubling of pay of eligible teachers. Mbiti et al. (2016) first construct a representative sample of 350 schools across 10 districts in Tanzania before randomly assigning them to various treatments and a control group.[5]

In most of these cases, the incremental cost of first constructing a representative sample and then randomizing the study sample into treatment and control groups was not much higher than using an alternative non-representative sample of the same size – these largely took the form of higher travel costs for survey teams. In addition, many of these studies above feature implementation by government, or by large non-government organizations with the ability to implement programs in wider jurisdictions. In such cases, the implementing partners typically welcomed the wide geographic spread of the study, because they intuitively grasped the importance of testing ideas across a more representative set of study sites, and also because it was politically easier to support pilots across a broader geographical area. Our interactions with government officials also suggest a considerable appetite for large over small experiments in the public sector – as exemplified by a quote from a senior government

A4 (row 2) show, the resulting estimates would be much less precise and 90% confidence interval around the estimates would be over twice as wide as in the case with the larger, more representative sample (a similar point is made by **?**. One procedure to potentially improve external validity would be to reweight the estimates by the inverse of the probability of a household being sampled in order to match to the distribution of observed covariates in the non-study districts. This method has been recommended in a recent discussion of randomized trials by Deaton and Cartwright (2016). The distribution of estimates from such a procedure is shown in Figure 4 (Panel C) and Table 2 (row 3) and the 90 percent confidence interval around the estimates is still nearly twice as wide as in the case with the larger more representative sample.

[5]Large-population representativeness is of course made much easier by the availability of high-quality administrative data, as for example in Kleven et al. (2011) who study tax compliance in a representative sample of taxpayers in Denmark. But as the examples above illustrate it has proven possible even where such data are lacking.

official in India who told one of us that it was "not worth his time to run an experiment in only 100 schools. Thus, neither logistics nor cost appear to be binding constraints to carrying out experiments at scales that are representative of larger populations than have been typical to date.

Of course, even results that are representative for a given large population may need to be extrapolated to others, and this must be done with care. If we seek to extrapolate the Smartcards results from Andhra Pradesh to, say, Indonesia, or Tanzania, we need to take into account the fact that Andhra Pradesh was not randomly selected from the universe of possible states or countries.[6] But we would be better positioned to make such a forecast having run an experiment across all of Andhra Pradesh than had we run it in a single district (say). External validity is after all a continuous and not a binary concept, and all else equal a sample representative of 10 million people does more for external validity than one that is representative of 10 thousand.

2.2 Experiments implemented at scale (or by governments)

Governments often roll out new programs at enormous scales despite little or no existent evidence on their effectiveness. These rollouts create exceptionally high-value opportunities for experimentation at scale, which researchers have already begun to exploit. We provide three examples below.

The first and arguably best-known example is Progresa-Oportunidades in Mexico. This was one of the original "conditional cash transfer programs, which aimed to provide income support to poor households while also promoting human capital accumulation of the next generation (Levy, 2006). It was introduced to randomly-selected communities and households during the program roll-out, which was unique at the time, and enabled high-quality experimental evaluation on program impacts (Schultz, 2004; Gertler and Boyce, 2003; Rivera et al., 2004). Further, because program implementation during this initial roll-out was done by the government, the estimates would reflect at least some of the implementation challenges that would be relevant when further scaling up.

A second example is the Smartcards evaluation we described above, in which the intervention was implemented by the government of Andhra Pradesh at full scale and thus reflected all the administrative, logistical, and political economy factors that typically affect the large-scale implementation of a major program. Moreover, because implementation protocols had

[6]One approach to this challenge is to conduct multi-site experiments where the same/similar program is experimentally evaluated in multiple locations. Such an approach is exemplified by Banerjee et al. (2015) who report results on the impact of a graduation program in reducing poverty across six different countries. Note however that the paper does not report the representativeness of the study populations within each country.

been refined and stabilized in the earliest districts to implement the scheme, they were more likely to reflect the steady-state approach to implementation. As a result the evaluation was able to produce highly policy-relevant point estimates.

A third example is De Ree et al. (2015) who study the effect of doubling teacher pay as part of the rollout of a teacher certification program in Indonesia. The program was implemented nationwide by the government, and the experiment followed exactly the same implementation protocol that was followed across Indonesia, simply accelerating its rollout in randomly-selected schools. Thus, while the experiment was not designed to test the extensive margin impacts of raising teacher salaries (since the announcement of a policy change happened nationally), it was able to study the intensive-margin impacts under government implementation at scale.

In addition to feasibility, the examples above illustrate the potential policy impact of evaluating government roll-outs. Progresa might well have been discontinued after the election of a new government, which was not enthusiastic about a program originated by its predecessor. However, the existence of high-quality evidence of impact likely played an important role in the continuation of the program, albeit with a name change Levy (2006). The evidence of impact from a government-implemented program (combined with its political popularity) is also thought to have played an important role in the rapid spread of CCTs to other countries in Latin America.

Smartcards were similarly found to be highly effective, improving almost every aspect of the affected programs: they reduced leakage, reduced payment delays, reduced time to collect payments, and increased access to work. However, opponents of the program (including lower-level officials whose rents were being squeezed) tended to convey negative anecdotes about Smartcards (such as cases in which genuine beneficiaries were excluded from receiving benefits for lack of a Smartcard), which created doubts among political leaders. This negative feedback was serious enough that the government nearly scrapped the program in 2013. The program survived in part because of the evaluation results and data showing that most beneficiaries strongly favored it.

The study in Indonesia, on the other hand, may have come a little too late. The study itself found that, while doubling teacher salaries increased teacher income and satisfaction with their income, and reduced financial stress and the likelihood of holding a second job it had zero impact on either the effort of incumbent teachers or on the learning outcomes of their students. Thus, a very expensive policy intervention (that cost over $5 billion every year) had no impact on the main stated goal of the government of Indonesia, which was to improve learning outcomes. In principle, such results are crucial for policy in a public sector setting, where there is no market test and where ineffective spending can often continue

indefinitely. A former Finance Minister of Indonesia wistfully expressed to one of us in a meeting that such results would have extremely useful in 2005 when the policy change was being debated. He also expressed optimism that the results would help in a renewed debate on the most effective ways of spending scarce public resources to improve human capital accumulation.

We hope that the three examples here – and other projects currently in progress – may be useful for researchers to highlight in conversations with potential government counterparts to demonstrate both the feasibility and the value of testing major policy reforms at scale.

2.3 Experiments with large units of randomization

A large-scale unit of randomization can potentially enable researchers to test for the *existence* of spillovers between treated and control units, and also to estimate *aggregate treatment effects* inclusive of such spillovers. We illustrate each type of study below, highlighting examples in which the ability to test for / measure spillovers was crucial to accurately estimating policy parameters.

A first prominent example is Miguel and Kremer (2004), who conduct a school-level randomization in Kenya to study the effects of deworming of primary school students on school attendance and test scores. They show using within- and between-school control groups that there are significant spillovers from treated to untreated students because treatment reduces the probability not just of having a worm infection but of transmitting one. As a result they obtained results quite different from earlier studies, which had randomized treatment at the individual level and thus likely under-estimated its impact. Randomizing at the larger unit was thus essential to obtaining unbiased results of the total treatment effect of a policy of universal deworming.[7]

A second example is provided by Muralidharan and Sundararaman (2015), who study the impact of school choice in the Indian state of Andhra Pradesh. A number of studies in the school choice literature have examined the relative effectiveness of private and public schools at improving test scores using student-level experiments that provide some students with vouchers to attend a private school. But these studies raise the question of whether there are spillovers on students left behind in public schools, perhaps due to the departure of their more motivated peers to better schools, or on students in private schools which receive an influx of potentially weaker peers. The study employs a two-stage design that first randomizes entire villages into treatment and control groups (where the treatment villages are eligible to receive the voucher program), and then further randomizes students in treatment villages into those

[7]Note that the results in Miguel and Kremer (2004) do not adjust for the downward bias from between-school spillovers and are hence still likely to be a lower bound on the true effects in their setting.

who receive vouchers and those who do not. Because the choice of primary school attended is highly sensitive to distance, the village-level randomization created an experiment at the level of a plausibly closed economy that enabled the authors to both estimate the spillovers from a school choice program and to estimate the aggregate effects of the program. As it turned out, spillovers were not meaningful in this setting – but this was itself an important finding, as the possibility of spillovers had been widely conjectured in the earlier school choice literature.

A final example is the Smartcards evaluation, which randomly assigned sub-districts of Andhra Pradesh to treatment and control categories. Since a sub-district contained an average population of 62,000 spread out across 20 to 25 large villages, this design allowed the authors to study impacts on rural labor markets more broadly. These effects are found to be quantitatively meaningful. Specifically, nearly 90 percent of the total increase in beneficiary income from the Smartcard program came from increases in private labor market earnings, while only 10 percent came from direct increases in earnings from the public employment program. They also find a significant increase in both stated reservation wages and realized market wages for beneficiaries in treatment areas. Finally, they find strong evidence that these effects "spill over across geographic sub-district boundaries, and estimate that correcting for these spillovers yields estimates of total treatment effects that are typically double or more in magnitude relative to the naive unadjusted estimates. Both sets of results underscore the potential importance of general equilibrium effects for program evaluation. In this sense the study is related to Cunha et al. (2015) who find using a village-level randomization design that transfers of food led to a decrease in food prices in remote villages, which is another example of successful randomization at a level that allowed the authors to estimate market spillover effects of policies.

These examples illustrate both the importance of randomizing at larger units in cases where spillovers may be salient, and the feasibility of doing so. Of course, designing such experiments will never be easy when the researcher does not know whether spillovers exist and/or the distances over which they are likely to be salient. The appropriate size of the unit of randomization will depend on the nature of spillovers, and so there is no uniform sense in which units can be considered "large. Thus, experimental designs need to rely on both theory and prior evidence to help in making the trade-off between larger units of randomization (that mitigate concerns of spillovers) on one hand and cost/feasibility on the other (for a discussion of the optimal unit of randomization in education experiments, for example, see Muralidharan (2017)).

3 Some practical considerations

Their merits aside, running large-scale experiments can be risky and hard. We have personally invested months of effort raising funds, negotiating, and designing studies, only to see them unwind because of political changes or administrative mishaps. How should researchers strike the right balance between experimentation, large and small? And what changes to the organization and financing of field research would be needed to successfully execute on more large-scale evaluations?

3.1 When to go big, and how to do small well

Certainly balance is needed; not all experiments should be "big. The lowest-hanging fruit may be to make samples more representative of the populations about which we wish to learn. From the data above and from personal experience, we think it safe to say that researchers have devoted more effort to persuading their partners to randomize (for internal validity) than to be representative (for external validity). We could often push harder to draw samples from frames that are larger and more representative – if less conveniently located near an NGOs headquarters or a research units office.

Implementation at scale, and randomization across large units, must be paid for in different coin. Opportunities for scale on these dimensions will most often arise when a government (say) has committed to rolling out some intervention. The choice will then be whether to evaluate that "status quo intervention at scale, or whether to instead evaluate some *other*, "challenger intervention – one that does not yet have political or budgetary support – at a smaller scale.[8] In terms of immediate policy impact, evaluating the status quo has a higher expected value the more resources it is receiving and the *lower* are the researchers priors that it works, as an evaluation will change decision-making only if it returns negative results. Evaluating the challenger, on the other hand, has higher expected value the *higher* are the researchers priors.

Where large-scale evaluations are not feasible, there is still scope to make smaller pilots as informative as possible about effects at scale. To address concerns about representativeness, smaller-scale experimental studies would do well to discuss their sampling procedure in more detail (which is often not done) and show tables comparing the study sample and the

[8]This smaller-scale evaluation might itself be the first step in an optimal sequence of experimentation, as discussed by Banerjee et al. (2016) in this volume. In another example, one of us has been evaluating a series of lump-sum cash transfers conducted by the nongovernment organization GiveDirectly (which one of us co-founded). The first evaluation, which was randomized at both household and village levels, did not find significant effects on prices, but this may reflect the limited number (126) of villages included. The next, larger evaluation (currently in progress) is randomized solely at the village level across 653 villages, and is designed with an explicit emphasis on estimating the dynamics of price and factor responses.

universe of interest on key observable characteristics (similar to tables showing balance on observable characteristics across treatment and control units). Plotting the distributions of key population characteristics in the universe and study samples (even if only in an appendix) will make it easier for readers to assess the extent to which results may apply to a broader population (a point also made by Deaton and Cartwright (2016)). More generally, tests of external validity and representativeness of the study sample should be as standard and taken as seriously as tests of internal balance between treatment and control group.

To address concerns around the scale of implementation, it is helpful at a minimum to describe implementation in sufficient detail to let others assess its scalability. For example, researchers can do more to scrutinize claims about fixed and marginal costs made by implementing partners that is currently the norm. Another useful approach is to pilot new programs at small scale but with implementation done by an organization capable of then scaling much further (e.g a government). Examples of experimental papers that successfully follow this approach include: (a) Olken (2007) who studies the impacts of increased audits on reducing corruption in Indonesia by using government auditors to conduct the (randomly-assigned) audits, (b) Muralidharan and Sundararaman (2013) who study the impact of an extra contract teacher on learning outcomes in India by having the government follow the standard implementation protocol for hiring an extra contract teacher (in randomly-selected villages), (c) Dal Bó et al. (2013) who study the impact of varying the salary offered on the quality of public employees recruited in Mexico, and (d) Khan et al. (2016) who study the impact of varying incentives for tax collectors on tax receipts and taxpayer experiences in Pakistan. The scale of implementation in these studies was often smaller in scope or duration than would be seen under a universal scale up. However, the experiment in each of these cases was implemented by government officials in ways that would plausibly mimic a scaled up implementation protocol.

Finally, researchers can to some extent anticipate potential general equilibrium effects even in small-scale studies by measuring impacts on behaviors which would be likely to affect prices in general equilibrium, and then forecasting the likely impacts. For example, if an intervention is found to affect household level labor supply, one could combine these data with estimates of the wage elasticity of labor demand to forecast the likely impact on wages at larger scale.

Another potential alternative for addressing external validity concerns is to embed small experiments within structural models in order to credibly estimate model parameters which then enable out of sample predictions (for discussion see Deaton and Cartwright (2016) or Low and Meghir (2017) in the Spring 2017 issue of this journal. We see potential value in this toolkit, but also limitations – for instance, it is unclear how well model-based extrapolation

can account for the implementation challenges that arise when small programs are scaled up, or account for the multiple margins on which programmatic interventions (which are often bundles of distinct components) change beliefs, preferences, and constraints of the agents whose optimizing behavior the model is trying to solve for. We therefore see large-scale experiments as the most direct way to estimate policy parameters of interest, and the structural approach as a sensible, complementary way to formalize and discipline extrapolation assumptions when they are required.

Finally, large experiments can be useful for testing and estimating deeper relationships in addition to policy parameters. For example, estimates of the effects of fiscal stimulus needed for macroeconomic calibrations could be obtained from large-scale experiments in redistribution such as the one ongoing at the NGO GiveDirectly, which studies the effects of capital inflows equivalent to $\tilde{\%}15$ of GDP in treated communities (pre-registered at `https://www.socialscienceregistry.org/trials/505`). Experimentation at such scales could help to bridge the gap between micro- and macro-development economics.

3.2 Organizing large-scale evaluations

Running large-scale experiments often requires a different set of skills and a different division of labor than smaller projects. Our partnership with the government of Andhra Pradesh, for example, was possible only because one of us had made a sustained investment over the years in building credibility and strong relationships with a number of senior decision-makers in government, who then lent their support when the opportunity for an evaluation arose. Building this sort of relationship-specific capital requires interpersonal skills which typically are neither taught nor screened for in graduate programs.

Once the project in Andhra Pradesh was approved, we faced the challenge of building a 150-person organization to collect data across the state in the course of a few months. This task requires strong people and process management skills – comparable perhaps to the work of building a state-level presidential campaign operation, a task which is generally assigned to veteran political organizers. Again, these organizational skills are not directly taught or screened for in most PhD programs (as our exceptionally hard-working research assistants from Andhra Pradesh can perhaps attest).

These specialized skills, along with a more productive division of labor, could be added to the research enterprise in several ways. Graduate programs could begin teaching them. Research organizations like J-PAL and Innovations for Poverty Action (IPA) could continue to add more specialized functions: for example, the J-PAL South Asia team has created a policy team focused on building and maintaining relationships with government. Researchers

could hire with greater emphasis on continuity, keeping teams together for longer periods of time and across multiple projects so that greater specialization can arise – something we are currently doing as part of a long-term initiative on direct benefit transfers in India. The training of young scholars could include a post-doctoral phase where these specialized skills are taught and learnt both explicitly and tacitly (a common model in the natural sciences, and one that we are increasingly supporting in our own work). And – though this issue can be a delicate one for economists – principal investigators could not only adopt more specialized roles but also signal these to the research community, as for example in the natural sciences where the distinct contributions of different authors are often acknowledged. These changes would involve tradeoffs, of course, but we believe some combination of them will be necessary to support large-scale experimentation.[9]

In terms of finance, large-scale experiments often require different models than small-scale ones. To be clear, grant size is often not the main issue here. After all, project costs are typically driven by the size of samples and the duration of measurement, which are largely independent of the dimensions of scale we highlighted above. But large projects – and especially collaborations with government – do often require greater *flexibility* in the timing of funding than smaller ones. In Andhra Pradesh, for example, the government agreed to randomize the Smartcards rollout and gave us weeks to arrange financing and commit to the project. Had our funder (the Omidyar Network) not evaluated our proposal far more quickly than the typical research grant cycle, the project would never have run.

Large-scale projects also benefit enormously from funding before they begin. Building the team necessary to execute well on a large-scale experiment requires a significant up-front investment in identifying talent, onboarding and training staff, developing good internal processes and culture, and so on. It would be more effective to organize and finance such work around a sustained program of work rather than to build and then dismantle such teams on a project-by-project basis. We therefore see increased value to financing research through broader and longer-term initiatives. Funding mechanisms such as the Agricultural Technology Adoption Initiative or the J-PAL Post-Primary Education, and Governance Initiatives represent a step in this direction, as they can be relatively flexible about purposing and repurposing funds, but they still fund on a project basis.

Experimenting at larger scales may also alter the optimal design of experiments themselves. For example, a large-scale impact evaluation with a government exposes a researcher to

[9]These issues are not restricted to field-experimental research. Similar changes may be needed to support work in teams working with administrative data sets from different settings or for teams of economists working with experts from other fields. More generally, as economics as a discipline shifts from an 'artisan to a 'team model of knowledge production (Jones, 2009), similar organizational innovations are likely to be required for the production of new economics knowledge.

significant risk, as it can be difficult to hold the government to an any agreed-upon rollout plan and timeline. In such scenarios the appropriate balance of risk and return might be to eschew the traditional baseline survey done before an experiment and conserve resources in order to run a larger endline survey (or multiple endline surveys), so that the bulk of research spending is incurred only after adherence to the study protocol is observed. For example, in a recent study one of us worked on the initial randomization was conducted using administrative data on schools while field data collection was conducted only after successful implementation of the intervention in treatment areas.

Funders could then take a similar approach to risk management, providing initial seed capital to enable research teams to negotiate experimental designs and then making the disbursal of funds for measurement contingent on proof of adherence to the experimental protocol. We are increasingly seeing funding committees on which we serve take exactly this approach, and encourage young researchers to frame proposals this way to increase their chances of receiving funding (in incremental tranches contingent on demonstrating success in prior phases). Innovations like these are important to keep the barriers to entry into impact evaluation low, so that resources do not become excessively concentrated in the hands of more established researchers.

One promising way of managing these issues is to create formal institutional frameworks for collaboration between researchers and government implementing partners, with dedicated funding. For instance, J-PAL South Asia has signed a Memorandum of Understanding with the government of the Indian state of Tamilnadu to undertake a series of experimental evaluations (typically with government implementation and funding) with a view to generating evidence that will help the state government to allocate financial and organizational resources when scaling up successful interventions. Another recent example is the MineduLab set up in Peru by J-PAL Latin America in partnership with the Ministry of Education in Peru to conduct a series of experimental evaluations. A third example is the partnership between J-PAL Southeast Asia and the Government of Indonesia to evaluate the design and delivery of social protection programs in Indonesia that has yielded several high-quality papers that have influenced both research and policy (Alatas et al., 2012b, 2016a; Banerjee et al., forthcoming). All these partnerships are broad based and allow for several researchers to work with the government counterpart, and are therefore likely to yield a stream of high-quality policy-relevant evidence.

Working hand-in-glove with implementing partners, whether large or small, will always create some risk of "researcher capture. A researcher who depends on maintaining a good relationship with an NGO or a government to publish well has weakened incentives for objectivity. While this is hardly a new issue, we wish to highlight safeguards which we have

found important in practice. First, researchers should use Memorandums of Understanding (MoUs) and pre-analysis plans judiciously as a means of protecting themselves against ex post pressure to shade or spin their analysis. Second, researchers should seek funding from independent sources to ensure they have allies who will support their objectivity, regardless of the results. Third, researchers should invest in a reputation for objectivity among local policy figures in the countries, as this helps to avoid entanglement with partners who expect a "rubber stamp. Finally, researchers can position themselves strategically in relation to the various factions within a government. For example, in settings where politicians routinely give bureaucrats new schemes to implement, the bureaucrats may be quite happy to have help in weeding out the ones that do not work. Alternatively, while line ministries may be overly enthusiastic about their latest schemes, finance ministries are typically more keen on identifying (and de-funding) the ones that do not work. We have often found that counterparts in ministries of finance and planning are more open to learning about negative results (as seen by the quote from the former Indonesian Finance Minister).[10]

In conclusion, the past fifteen years have seen an explosion in the number of randomized controlled trials in development economics across topics and geographic regions. This trend has been accompanied by extensive debate in the economics profession regarding the strengths and limitations of RCTs for policy evaluation. Our goal in this paper has been to demonstrate one practical way forward to combine the credibility and transparency of RCTs with greater policy relevance, which is to run experiments at a larger scale.

We believe that this is a fruitful approach to pursue both because large-scale RCTs are likely to be directly decision-relevant (as by their nature they will often evaluate expensive new programs being rolled out), and also because they can overcome some of the limitations of smaller experiments with respect to external validity. Specifically, we have characterized the scale of existing studies on three dimensions (representativeness of populations studied, scale of implementation, and spillovers to non-treated participants), discussed the extent to which the external validity of individual studies can be improved by conducting more of them at a larger scale, and illustrated with several examples the feasibility of doing so. We have also aimed to provide a brief discussion on factors that can facilitate experimentation at scale, and hope that this paper helps to encourage more such work going forward.

[10]See (Gueron, 2017) for an insightful historical review of the economics and politics of the increased use of RCTs for evaluating welfare programs in the US. The chapter provides a US-focused discussion of several of the themes in this section.

References

Acemoglu, Daron, "Theory, general equilibrium, and political economy in development economics," *The Journal of Economic Perspectives*, 2010, *24* (3), 17–32.

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias, "Targeting the poor: Evidence from a field experiment in Indonesia," *The American Economic Review*, 2012, *102* (4), 1206–1240.

_ , _ , _ , Benjamin A. Olken, and Julia Tobias, "Targeting the Poor: Evidence from a Field Experiment in Indonesia," *American Economic Review*, June 2012, *102* (4), 1206–40.

_ , _ , _ , _ , Ririn Purnamasari, and Matthew Wai-Poi, "Self-Targeting: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, 2016, *124* (2), 371–427.

_ , Ririn Purnamasari, Matthew Wai-Poi, Abhijit Banerjee, Benjamin A Olken, and Rema Hanna, "Self-Targeting: Evidence from a field experiment in Indonesia," *Journal of Political Economy*, 2016, *124* (2), 371–427.

Allcott, Hunt, "Site selection bias in program evaluation," *The Quarterly Journal of Economics*, 2015, *130* (3), 1117–1165.

Angrist, Joshua D, "American education research changes tack," *Oxford review of economic policy*, 2004, *20* (2), 198–212.

_ and Jörn-Steffen Pischke, "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *The Journal of Economic Perspectives*, 2010, *24* (2), 3–30.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer, "Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment," *The American Economic Review*, 2002, *92* (5), 1535–1558.

Ashraf, Nava, Dean Karlan, and Wesley Yin, "Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines," *The Quarterly Journal of Economics*, 2006, (2), 635–672.

_ , Erica Field, and Jean Lee, "Household bargaining and excess fertility: An experimental study in Zambia," *The American Economic Review*, 2014, *104* (7), 2210–2237.

_ , **James Berry, and Jesse M Shapiro**, "Can higher prices stimulate product use? Evidence from a field experiment in Zambia," *The American Economic Review*, 2010, *100* (5), 2383–2413.

Attanasio, Orazio P, Costas Meghir, and Ana Santiago, "Education choices in Mexico: Using a structural model and a randomized experiment to evaluate Progresa," *The Review of Economic Studies*, 2012, *79* (1), 37–66.

Baird, Sarah, Craig McIntosh, and Berk Özler, "Cash or condition? Evidence from a cash transfer experiment," *The Quarterly Journal of Economics*, 2011, (4), 1709–1753.

Banerjee, Abhijit and Esther Duflo, "Handbook of Field Experiments," 2016.

_ , _ , **Clement Imbert, Santhosh Mathew, and Rohini Pande**, "Can e-governance reduce capture of public programs? Experimental evidence from a financial reform of Indias employment guarantee," Technical Report 2015.

_ , _ , **Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry**, "A multifaceted program causes lasting progress for the very poor: Evidence from six countries," *Science*, 2015, *348* (6236), 772–788.

_ , **Rema Hanna, Jordan Kyle, Benjamin A. Olken, and Sudarno Sumarto**, "Tangible Information and Citizen Empowerment: Identification Cards and Food Subsidy Programs in Indonesia," *Journal of Political Economy*, forthcoming.

_ , **Rukmini Banerji, Jim Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton**, "From proof of concept to scalable policies: Challenges and solutions, with an application," *Journal of Economic Perspectives*, 2016, (4).

Banerjee, Abhijit V, Esther Duflo, and Rachel Glennerster, "Putting a Band-Aid on a corpse: Incentives for nurses in the Indian public health care system," *Journal of the European Economic Association*, 2008, *6* (2-3), 487–500.

_ , **Shawn Cole, Esther Duflo, and Leigh Linden**, "Remedying education: Evidence from two randomized experiments in India," *The Quarterly Journal of Economics*, 2007, (4), 1235–1264.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova, "Powerful Women: Does Exposure Reduce Bias?," *The Quarterly Journal of Economics*, 2009, *124* (4), 1497–1540.

Beath, Andrew, Fotini Christia, Georgy Egorov, and Ruben Enikolopov, "Electoral rules and political selection: Theory and evidence from a field experiment in Afghanistan," *The Review of Economic Studies*, 2016, (3), 932–968.

Behrman, Jere R, Susan W Parker, Petra E Todd, and Kenneth I Wolpin, "Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools," *Journal of Political Economy*, 2015, *123* (2), 325–364.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman, "What's advertising content worth? Evidence from a consumer credit marketing field experiment," *The Quarterly Journal of Economics*, 2010, *125* (1), 263–306.

_ , Simeon Djankov, Rema Hanna, and Sendhil Mullainathan, "Obtaining a driver's license in India: An experimental approach to studying corruption," *The Quarterly Journal of Economics*, 2007, (4), 1639–1676.

Björkman, Martina and Jakob Svensson, "Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda," *The Quarterly Journal of Economics*, 2009, *124* (2), 735–769.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez, "Generating skilled self-employment in developing countries: Experimental evidence from Uganda," *Quarterly Journal of Economics*, 2014, (2), 697–752.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts, "Does management matter? Evidence from India," *The Quarterly Journal of Economics*, 2013, *128* (1), 1–51.

_ , James Liang, John Roberts, and Zhichun Jenny Ying, "Does working from home work: Evidence from a Chinese experiment," *The Quarterly Journal of Economics*, 2015, *130* (1), 165–218.

Bó, Ernesto Dal, Frederico Finan, and Martín A Rossi, "Strengthening state capabilities: The role of financial incentives in the call to public service," *The Quarterly Journal of Economics*, 2013, *128* (3), 1169–1218.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, "Scaling up what works: Experimental evidence on external validity in Kenyan education," *Center for Global Development Working Paper*, 2013, (321).

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak, "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh," *Econometrica*, 2014, *82* (5), 1671–1748.

Bursztyn, Leonardo and Lucas C Coffman, "The schooling decision: Family preferences, intergenerational conflict, and moral hazard in the Brazilian favelas," *Journal of Political Economy*, 2012, *120* (3), 359–397.

_ , Florian Ederer, Bruno Ferman, and Noam Yuchtman, "Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions," *Econometrica*, 2014, *82* (4), 1273–1301.

Callen, Michael and James D Long, "Institutional corruption and election fraud: Evidence from a field experiment in Afghanistan," *The American Economic Review*, 2014, *105* (1), 354–381.

Cartwright, Nancy, "Are RCTs the gold standard?," *BioSocieties*, 2007, *2* (1), 11–20.

Casey, Katherine, Rachel Glennerster, and Edward Miguel, "Reshaping institutions: Evidence on aid impacts Using a preanalysis plan," *Quarterly Journal of Economics*, 2012, *127* (4), 1755–1812.

Chattopadhyay, Raghabendra and Esther Duflo, "Women as policy makers: Evidence from a randomized policy experiment in India," *Econometrica*, 2004, *72* (5), 1409–1443.

Cohen, Jessica and Pascaline Dupas, "Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment," *The Quarterly Journal of Economics*, 2010, *125* (1), 1–45.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment," *Quarterly Journal of Economics*, 2013, *128* (2), 531–580.

Cunha, Jesse M, Giacomo De Giorgi, and Seema Jayachandran, "The price effects of cash versus in-kind transfers," 2011.

_ , _ , and _ , "The Price Effects of Cash versus In-Kind Transfers," 2015.

Deaton, Angus, "Instruments, randomization, and learning about development," *Journal of Economic Literature*, 2010, *48* (2), 424–455.

_ **and Nancy Cartwright**, "Understanding and misunderstanding randomized controlled trials," 2016.

Duflo, Esther and Abhijit Banerjee, *Handbook of Field Experiments*, North Holland, 2017.

_ **, Michael Greenstone, Rohini Pande, and Nicholas Ryan**, "Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India," *The Quarterly Journal of Economics*, 2013, *128* (4), 1499–1545.

_ **, Michael Kremer, and Jonathan Robinson**, "Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya," *The American Economic Review*, 2011, *101* (6), 2350–2390.

_ **, Pascaline Dupas, and Michael Kremer**, "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *The American Economic Review*, 2011, *101* (5), 1739–1774.

_ **,** _ **, and** _ , "Education, HIV, and early fertility: Experimental evidence from Kenya," *The American Economic Review*, 2015, *105* (9), 2757–2797.

_ **, Rema Hanna, and Stephen P Ryan**, "Incentives work: Getting teachers to come to school," *The American Economic Review*, 2012, *102* (4), 1241–1278.

Dupas, Pascaline and Jonathan Robinsona, "Why don't the poor save more? Evidence from health savings experiments," *The American Economic Review*, 2013, *103* (4), 1138–1171.

Feigenberg, Benjamin, Erica Field, and Rohini Pande, "The economic returns to social interaction: Experimental evidence from microfinance," *The Review of Economic Studies*, 2013, (4), 1459–1483.

Fryer, Roland, "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments," in "Handbook of Field Experiments," Vol. 2, North Holland, 2017, chapter 2, pp. 95–322.

Gertler, Paul J and Simone Boyce, "An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico," Royal Economic Society Annual Conference 2003 85, Royal Economic Society June 2003.

Glewwe, Paul and Karthik Muralidharan, "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications," *Handbook of the Economics of Education*, 2016, *5*, 653–743.

_ , Michael Kremer, Sylvie Moulin, and Eric Zitzewitz, "Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya," *Journal of Development Economics*, 2004, *74* (1), 251–268.

Gueron, J.M., "The Politics and Practice of Social Experiments: Seeds of a Revolution," in "Handbook of Field Experiments," Vol. 1, North Holland, 2017, chapter 2, pp. 27–69.

Gueron, Judith M, "The Politics and Practice of Social Experiments," 2016.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning through noticing: Theory and evidence from a field experiment," *The Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Heckman, James J and Jeffrey A Smith, "Assessing the case for social experiments," *The Journal of Economic Perspectives*, 1995, *9* (2), 85–110.

Hsieh, Chang-Tai and Peter J Klenow, "Misallocation and manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 2009, *124* (4), 1403–1448.

Jensen, Robert, "The (perceived) returns to education and the demand for schooling," *Quarterly Journal of Economics*, 2010, *125* (2), 515–548.

_ , "Do labor market opportunities affect young women's work and family decisions? Experimental evidence from India," *The Quarterly Journal of Economics*, 2012, (2), 753–792.

Jones, Benjamin F., "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?," *Review of Economic Studies*, 2009, *76* (1), 283–317.

Kaboski, Joseph P. and Robert M. Townsend, "A Structural Evaluation of a LargeScale QuasiExperimental Microfinance Initiative," *Econometrica*, 09 2011, *79* (5), 1357–1406.

Karlan, Dean and Jonathan Zinman, "Observing unobservables: Identifying information asymmetries with a consumer credit field experiment," *Econometrica*, 2009, *77* (6), 1993–2008.

_ , **Robert Osei, Isaac Osei-Akoto, and Christopher Udry**, "Agricultural decisions after relaxing credit and risk constraints.," *The Quarterly Journal of Economics*, 2014, *129* (2), 597–652.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan, "Self-control at work," *Journal of Political Economy*, 2015, *123* (6), 1227–1277.

Khan, Adnan Q, Asim I Khwaja, and Benjamin A Olken, "Tax farming redux: Experimental evidence on performance pay for tax collectors," *The Quarterly Journal of Economics*, 2016, *131* (1), 219–271.

Kleven, Henrik Jacobsen, Martin B Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez, "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark," *Econometrica*, 2011, *79* (3), 651–692.

Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Peterson Zwane, "Spring cleaning: A randomized evaluation of source water quality improvement," *Quarterly Journal of Economics*, 2011, *4* (2), 9–30.

LaLonde, Robert J, "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, 1986, pp. 604–620.

Leamer, Edward E, "Let's take the con out of econometrics," *The American Economic Review*, 1983, *73* (1), 31–43.

Levy, Santiago, *Progress Against Poverty: Sustaining Mexico's Progresa-Oportunidades Program*, Brookings Institution Press, 2006.

Low, Hamish and Costas Meghir, "The Use of Structural Models in Econometrics," *Journal of Economic Perspectives*, May 2017, *31* (2), 33–58.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani, "Inputs, Incentives and Complementarities in Primary Education: Experimental Evidence from Tanzania," 2016.

Mel, Suresh De, David McKenzie, and Christopher Woodruff, "Returns to capital in microenterprises: Evidence from a field experiment," *The Quarterly Journal of Economics*, 2008, (4), 1329–1372.

Miguel, Edward and Michael Kremer, "Worms: Identifying impacts on education and health in the presence of treatment externalities," *Econometrica*, 2004, *72* (1), 159–217.

Moffitt, Robert A, "The Role of Randomized Field Trials in Social Science Research A Perspective from Evaluations of Reforms of Social Welfare Programs," *American Behavioral Scientist*, 2004, *47* (5), 506–540.

Muralidharan, Karthik, "Field experiments in education in developing countries," *Handbook of Economic Field Experiments*, 2017, *2*, 323–385.

_ **and Venkatesh Sundararaman**, "The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India," *The Economic Journal*, 2010, *120* (546), F187–F203.

_ **and** _ , "Teacher performance pay: Experimental evidence from India," *The Journal of Political Economy*, 2011, *119* (1), 39–77.

_ **and** _ , "Contract teachers: Experimental evidence from India," 2013.

_ **and** _ , "The aggregate effect of school choice: Evidence from a two-stage experiment in India," *The Quarterly Journal of Economics*, 2015, *130* (3), 1011–1066.

_ **, Paul Niehaus, and Sandip Sukhtankar**, "General Equilibrium Effects of (Improving) Public Employment Programs," Technical Report, UC San Diego August 2017.

_ **,** _ **, and** _ , "Building state capacity: Evidence from biometric smartcards in India," *The American Economic Review*, Forthcoming.

Olken, Benjamin A, "Monitoring corruption: Evidence from a field experiment in Indonesia," *Journal of Political Economy*, 2007, *115* (2), 200–249.

Ravallion, Martin, "Should the Randomistas Rule?," *The Economists' Voice*, 2009, *6* (2).

Ree, Joppe De, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for nothing? Experimental evidence on the impact of an unconditional teacher salary increase on student performance in Indonesia," 2015.

Rivera, Juan A, Daniela Sotres-Alvarez, Jean-Pierre Habicht, Teresa Shamah, and Salvador Villalpando, "Impact of the Mexican program for education, health, and nutrition (Progresa) on rates of growth and anemia in infants and young children: a randomized effectiveness study," *Jama*, 2004, *291* (21), 2563–2570.

Schultz, Paul T., "School subsidies for the poor: evaluating the Mexican Progresa poverty program," *Journal of Development Economics*, June 2004, *74* (1), 199–250.

Tarozzi, Alessandro, Aprajit Mahajan, Brian Blackburn, Dan Kopf, Lakshmi Krishnan, and Joanne Yoong, "Micro-loans, insecticide-treated bednets, and malaria: Evidence from a randomized controlled trial in Orissa, India," *The American Economic Review*, 2014, *104* (7), 1909–1941.

Todd, Petra E and Kenneth I Wolpin, "Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility," *The American Economic Review*, 2006, *96* (5), 1384–1417.

Vivalt, Eva, "Heterogeneous treatment effects in impact evaluation," *American Economic Review*, 2015, *105* (5), 467–70.

Wolpin, Kenneth, *The limits of inference without theory*, MIT Press, 2013.

Table 1: Summary statistics: program evaluation RCTs in top journals, 2001-2016

Variable	25th %	Median	75th %	Mean	SD	N
Sample represents larger population?	0	0	1	0.31	0.47	29
Size of sampling frame	529	10,885	46,418	681,955	2,715,907	26
Units treated	401	5,340	29,325	21,233	49,857	29
Clustered randomization?	0	1	1	0.66	0.48	29
Mean size of randomization unit	1	26	99	167	477	28

This table reports summary statistics for measures of experimental scale for randomized controlled trials published in Econometrica, American Economic Review, the Quarterly Journal of Economics, Review of Economics and Statistics and the Journal of Political Economy between January 2001 and July 2016 which we categorized as primarily "program evaluations" (as opposed to mechanism experiments). Counting metrics are defined in "primary units of analysis," which we define as the level at which the studies' primary outcomes are measured (e.g. the household). "Sample represents larger population?" is an indicator equal to one if the paper reports systematically drawing its evaluation sample from any larger population of interest. "Size of sampling frame" is the size of the frame sampled (equal to size of the evaluation sample itself if no larger frame is indicated). "Units treated" is the number of units treated by the organization implementing the intervention being studied. "Clustered randomization?" is an indicator equal to one if randomization was assigned in geographic groupings larger than the primary analysis unit, and "mean size of randomization unit" is the average number of primary analysis units per cluster (equal to 1 for unclustered designs).

Figure 1: Distributions of (log) measures of experimental scale

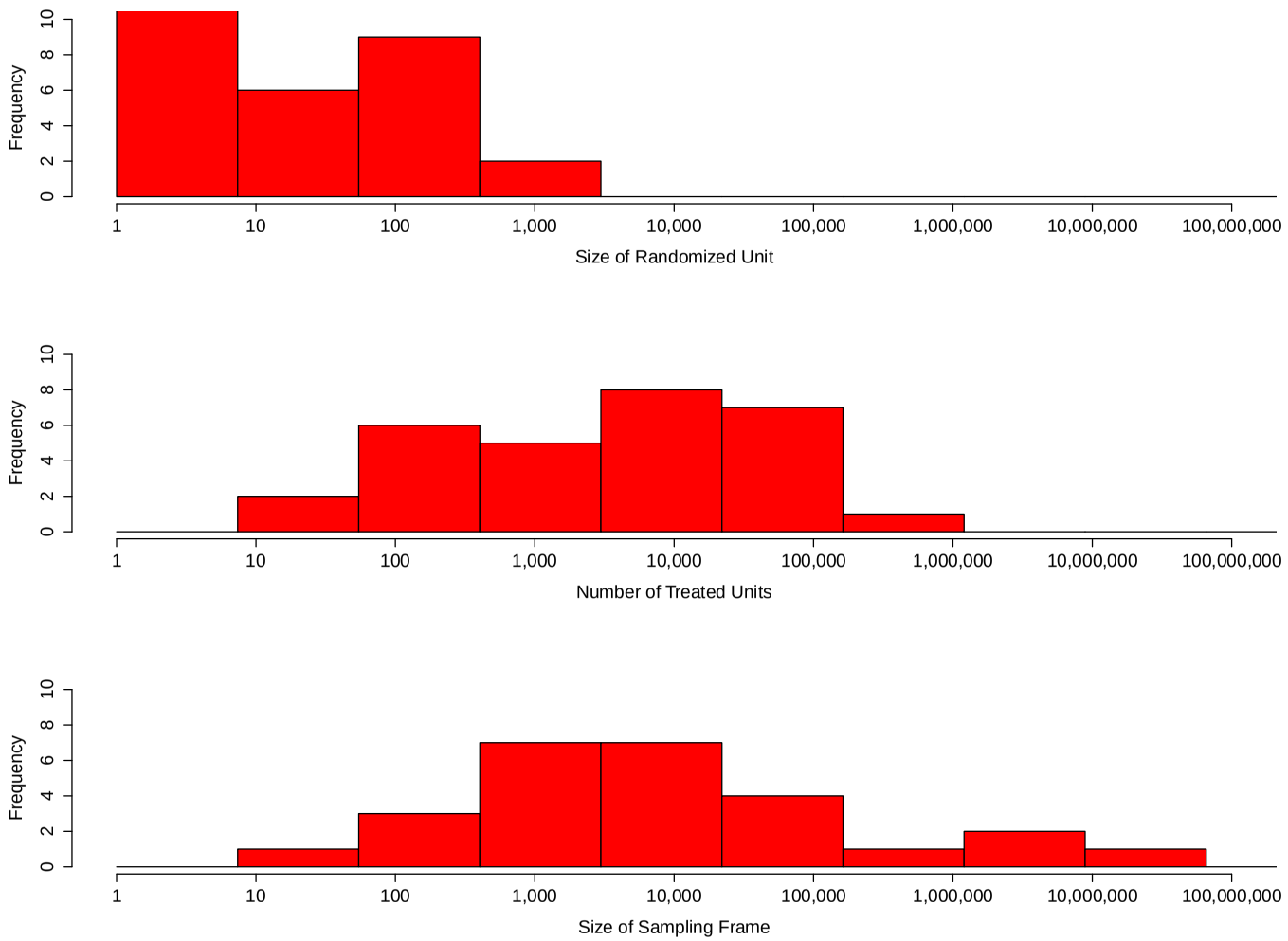
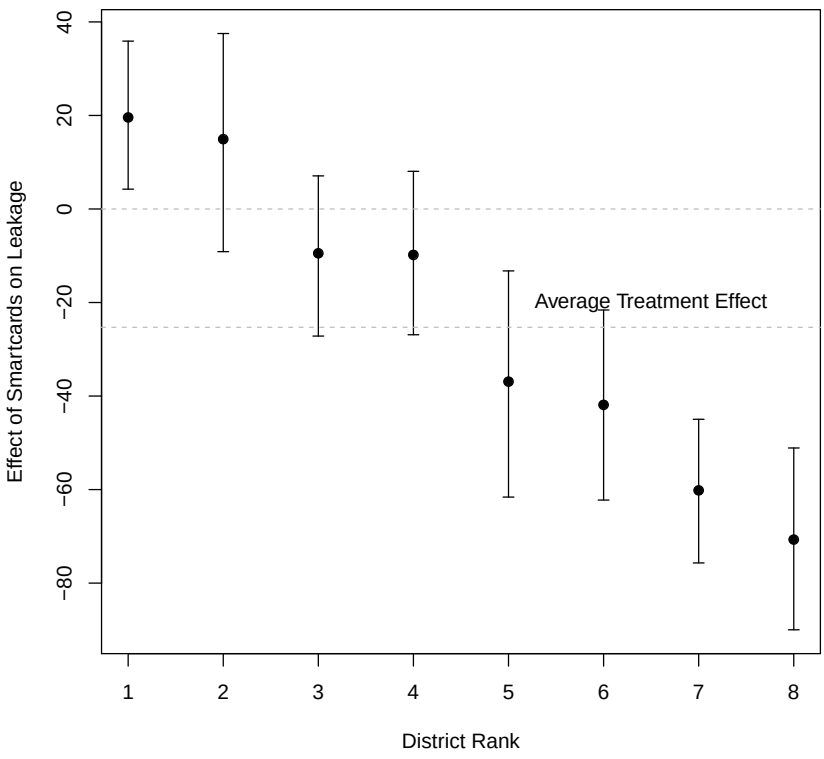


Figure shows the distribution of key attributes of "program evaluation" RCTs published in selected economics journals between January 2001 and July 2016. For more detail on sample and variable construction, refer to notes for Table 1.

Figure 2: Mean effects of Smartcards on leakage, by district



Note: This figure shows average treatment effect of Smartcards on program leakage for each of the 8 districts in the experimental sample of Muralidharan, Niehaus and Sukhtankar (2016). Error bars show the 90% confidence interval generated through a block bootstrap.

A Dataset Creation

A.1 Sample Construction

We constructed our sample of program evaluation randomized-control trials as follows.

First, we examined the abstracts of all papers published in Econometrica, the Quarterly Journal of Economics, the American Economic Review (excluding Papers and Proceedings), the Review of Economics and Statistics and the Journal of Political Economy between January 2001 and July 2016 to identify studies involving randomized controlled trials or policy lotteries. We excluded all studies that either took place in North America (except Mexico), Europe, Japan or Australia/New Zealand or were re-analyses of previously published experiments. This yielded an initial list of 45 studies. These studies are listed in Appendix Table A1.

From this sample, we identified studies that evaluated an intervention which could plausibly (or was already) scaled up as is. This criterion was meant to differentiate "mechanism experiments from experiments evaluating (potential) policies. For example, Muralidharan and Sundararaman (2011) considers a teacher performance pay program that could plausibly be scaled up province-wide and thus was coded as a "program evaluation. By contrast, one of the major treatments in Dupas and Robinsona (2013) analysis of saving technologies is a lockbox maintained by experiments program officer. Scaling up such a treatment would require significant changes and therefore we coded Dupas and Robinsona (2013) as a mechanism experiment. Column 5 of Appendix Table A1 indicates how we coded each study we considered. .

This categorization is of course inevitably somewhat subjective. For instance, we excluded a study by Jensen (2012) which estimated the effects of the experimenters randomly sending call-center recruiters to Indian villages on young womens fertility and labor market outcomes because it is framed as a test of a specific theoretical mechanism: does an exogenous shock to the perceived labor market value of women lead to changes in fertility behavior? One could also argue, however, that this intervention should be considered as a broader policy. (Cohen and Dupas, 2010) is another borderline case: we coded it as a program evaluation because it examined the impacts of an existing policy (subsidizing bed nets) even though the framing of the paper is focused on distinguishing between potential mechanisms for the interaction between subsidies and consumer uptake.

That said, our final results turn out to be insensitive to changing these borderline classifications. Appendix Table A2 shows mean and median values from Table 1 for our primary sample, our primary sample plus all borderline cases, and our primary sample minus all borderline cases. Evidence, inclusion or exclusion of borderline cases has little effect on our substantive conclusions.

A.2 Coding experimental size

We coded five metrics of scale: representativeness of sample, size of sampling frame, number of units treated, whether or not the experiment randomized at the cluster-level and size of unit of randomization. This section describes how each statistic was obtained.

A.2.1 Units of Analysis

Because several of our metrics are counts of number of units, a necessary first step is to define the relevant unit of analysis for each study – is it the individual, the household, the class, etc. To do this we first defined a primary outcome for each study – that is, the outcome on which the papers central claims most directly rest – and defined the unit at which this outcome was measured as the primary unit of analysis. For example, the primary unit of analysis in Cohen and Dupas (2010) is the household because the authors study the effect of subsidization of bed-nets on household bed-net purchases and utilization. The primary unit of analysis in Olken (2007) is road projects because the studys primary outcome are road project missing expenditures.

In cases where major outcomes were recorded for more than one level of analysis (e.g. teachers and students as in Duflo et al. (2012) or villages and households as in Alatas et al. (2012a)), we broke ties by choosing the lower level of aggregation. Appendix Table A3 shows the outcomes and level of outcomes chosen for each paper in our primary analysis.

While we believe this is the conceptually most defensible way to define scale, one might worry that it leads us to understate the size of experiments by focusing on units of aggregation larger than the individual person. We therefore also re-created our metrics using the individual as the primary unit of analysis throughout. Appendix Table A3 replicates Table 1 using this variable definition. As expected, experiment sizes are larger using this metric, but our substantive conclusions do not change at all – experiments are typically small relative to the size of the population of interest.

A.2.2 Sample Representativeness

We coded a study as representative of a larger population if the study sample was randomly drawn from some larger population of interest. (Note that this statistic is invariant to our choice of primary unit of analysis.)

A.2.3 Size of Sampling Frame

The size of the subject sampling frame was constructed in one of two ways. In the cases where the experiment was not drawn from a larger population, the size of the sampling frame is equal to the number of primary units in the study. In the cases where the experiment is a representative sample of a larger population, the number of primary units in this larger population size was used. In many cases, this population size was not stated explicitly, but could be reasonably estimated from outside sources.

Importantly, we restricted our estimate of the sampling frame to only those individuals potentially affected by an intervention. For instance, Baird et al. (2011) is a conditional cash transfer experiment focused on education and fertility outcomes for young women. Thus, the sampling frame from this study was the total number of *young women* in the population from which the sample was drawn, not the overall population.

A.2.4 Number of Units Treated

The number of units treated was constructed in the following manner. We defined a unit of randomization as treated if they received any intervention from the experimenters. In most cases, this accords exactly with how treatment and control is defined by a studys authors. However, in some cases, all units in a study received some intervention by the experimenters. For example, in Tarozzi et al. (2014) both treatment and control villages (as defined by authors) received an information intervention. Thus, even though the authors defined information intervention-only villages as the control,.for the purposes of our statistic all villages in Tarozzi et al. (2014) were considered treated. Our final metric is equal to the total number of treated primary units in the study.

A.2.5 Cluster Randomized

We coded a study as cluster randomized if its unit of randomization contained more than one of its primary unit of analysis.

A.2.6 Size of Unit of Randomization

We defined the size of the unit of randomization as the total number of primary units within a unit of randomization. For instance, Callen and Long (2014) uses polling centers in Afghanistan as a unit of randomization. Although each polling center encompasses hundreds of voters, the primary outcome in Callen and Long (2014) is aggregation fraud at the polling-center level. Thus, in our primary classification we define the size of the unit of randomization for Callen and Long (2014) as 1. When we instead measure size by number of individuals, we define the size of the unit of randomization as 269, the average number of voters per polling center.

Table A1: Full list of development RCTs in top journals

Author	Title	Journal	Year	PE?	Close?	Primary unit	Randomized unit
Joshua Angrist, Eric Bettinger, Erik Bloom, Elizabeth King, Michael Kremer	Vouchers For Private Schooling In Colombia: Evidence From A Naturalized Field Experiment	AER	2002	1	0	students	individual
Edward Miguel, Michael Kremer	Worms: Identifying Impacts On Education And Health In The Presence Of Treatment Externalities	EMA	2004	1	0	students	schools
Raghabendra Chattopadhyay, Esther Duflo	Women As Policymakers: Evidence From A Policy Experiment In India	EMA	2004	1	0	GP council members	gram prachyat
Nava Ashraf , Dean Karlan, Wesley Yin	Tying Odysseus To The Mast: Evidence From A Commitment Savings Product In The Phillipines	QJE	2006	0	1	individuals	individuals
Benjamin Olken	Monitoring Corruption: Evidence From A Field Experiment In Indonesia	JPE	2007	1	1	projects	village
Abhijit Banerjee, Shawn Cole, Esther Duflo, Leigh Linden	Remedying Education: Evidence From Two Randomized Experiments In India	QJE	2007	1	0	students	schools
Marianne Bertrand, Simeon Djankov, Rema Hanna, Sendhil Mullainathan	Obtaining A DriverS License In India: An Experimental Approach To Studying Corruption	QJE	2007	0	0	individual	individuals
Suresh Del Mel, David McKenzie, Christopher Woodruff	Returns To Capital In Microenterprises: Evidence From A Field Experiment	QJE	2008	0	0	firm	firms
Dean Karlan, Jonathan Zinman	Observing Unobservables: Identifying Information Asymmetries With A Consumer Credit Field Experiment	EMA	2009	0	0	individual	individuals
Martina Bjorkman, Jacob Svensson	Power To The People: Evidence From A Randomized Field Experiment On Community-Based Monitoring In Uganda	QJE	2009	1	0	households	catchment area of clinic
Nava Ashraf , James Berry, Jesse Shaprio	Can Higher Prices Stimulate Product Use? Evidence From A Field Experiment In Zambia	AER	2010	0	0	individual	individual
Marianne Bertrand, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, Jonathan Zinman	WhatS Advertising Content Worth? Evidence From A Consumer Credit Marketing Field Experiment	QJE	2010	1	1	individuals	individual
Jessica Cohen, Pascaline Dupas	Free Distribution Or Cost-Sharing? Evidence From A Randomized Malaria Prevention Experiment	QJE	2010	1	1	individuals	community health clinic
Robert Jensen	The (Perceived) Returns To Education And The Demand For Schooling	QJE	2010	0	0	individual	school
Esther Duflo, Pascaline Dupas, Michael Kremer	Peer Effects, Teacher Incentives, And The Impact Of Tracking: Evidence From A Randomized Evaluation In Kenya	AER	2011	1	1	students	school (1st grade)

Author	Title	Journal	Year	PE?	Close?	Primary unit	Randomized unit
Esther Duflo, Michael Kremer, Jonathan Robinson	Nudging Farmers To Use Fertilizer: Theory And Experimental Evidence From Kenya	AER	2011	0	0	households	household
Karthik Muralidharan, Venkatesh Sundararaman	Teacher Performance Pay: Experimental Evidence From India	JPE	2011	1	0	students	school
Michael Kremer, Jessica Leino, Edward Miguel, Alex Zwayne	Spring Cleaning: Rural Water Impacts, Valuation, And Property Rights Institutions	QJE	2011	1	0	households	spring (water)
Sarah Baird, Craig McIntosh, Berk Ozler	Cash Or Condition? Evidence From A Cash Transfer Experiment	QJE	2011	1	0	individuals	enumeration area
Esther Duflo, Rema Hanna, Stephen Ryan	Incentives Work: Getting Teachers To Come To School	AER	2012	1	0	students	school
Vivi Alatas, Abhijit Bannerjee, Rema Hanna, Ben Olken, Julia Tobias	Targeting The Poor: Evidence From A Field Experiment In Indonesia	AER	2012	1	0	households	subvillage
Leonardo Bursztyn, Lucas Coffman	The Schooling Decision: Family Preferences, Intergenerational Conflict, And Moral Hazard In The Brazilian Favelas	JPE	2012	0	0	household	household
Robert Jensen	Do Labor Market Opportunities Affect Young WomenS Work And Family Decisions? Experimental Evidence From India	QJE	2012	0	1	individuals	village
Katherine Casey, Rachel Gelnnerster, Edward Miguel	Reshaping Institutions: Evidence On Aid Impacts Using A Preanalysis Plan	QJE	2012	1	0	household	village
Pascaline Dupas, Jonathan Robinson	Why Don'T The Poor Save More	AER	2013	0	0	individual	ROSCA
Nicholas Bloom, Benn Eifert, Aprajit Mahajan, David McKenzie, John Roberts	Does Management Matter? Evidence From India	QJE	2013	1	0	firm	firm
Ernesto Dal Bo, Frederico Finan, Martin Rossi	Strengthening State Capabilities: The Role Of Financial Incentives In The Call To Public Service	QJE	2013	1	1	individuals	locality
Esther Duflo, Michael Greenstone, Rohini Pande, Nicholas Ryan	Truth-Telling By Third-Party Auditors And The Response Of Polluting Firms: Experimental Evidence From India	QJE	2013	1	0	plant	plant
Benjamin Feigenberg, Erica Field, Rohini Pande	The Economic Returns To Social Interaction: Experimental Evidence From Microfinance	ReStud	2013	1	1	individuals	microfinance group
Alessandro Tarrozi, Aprajit Mahajan, Brian Blackburn, Dan Kopf, Lakshmi Krishhnan, Joanne Yoong	Micro-Loans, Insecticide-Treated Bednets, And Malaria: Evidence From A Randomized Controlled Trial In Orissa, India	AER	2014	1	0	households	village

Author	Title	Journal	Year	PE?	Close?	Primary unit	Randomized unit
Nava Ashraf , Erica Field, Jean Lee	Household Bargaining And Excess Fertility: An Experimental Study In Zambia	AER	2014	0	0	individual	individual
Leonardo Bursztyn, Florian Ederer, Bruno Ferman, Noam Yuchtman	Understanding Mechanisms Underlying Peer Effects: Evidence From A Field Experiment On Financial Decisions	EMA	2014	0	0	individual	client pairs
Gharad Bryan, Shyamal Chowdhury, Ahmed Mobarak	Underinvestment In A Profitable Technology: The Case Of Seasonal Migration In Bangladesh	EMA	2014	0	1	households	village
Dean Karlan, Robert Osei, Isaac Osei-Akoto, Christopher Udry	Agricultural Decisions After Relaxing Credit And Risk Constraints	QJE	2014	0	0	household	household
Christopher Blattman, Nathan Fiala, Sebastian Martinez	Generating Skilled Self-Employment In Developing Countries: Experimental Evidence From Uganda	QJE	2014	1	0	individuals	proposal groups
Rema Hanna, Sendhil Mullainathan, Joshua Schwartzstein	Learning Through Noticing: Theory And Evidence From A Field Experiment	QJE	2014	0	0	individual	individual
Michael Callen, James Long	Institutional Corruption And Election Fraud: Evidence From A Field Experiment In Afghanistan	AER	2015	1	1	polling center	polling centers
Esther Duflo, Pascaline Dupas, Michael Kremer	Education, Hiv, And Early Fertility: Experimental Evidence From Kenya	AER	2015	1	0	individuals	school (grade 6)
Jere Behrman, Susan Parker, Petra Todd, Kenneth Wolpin	Aligning Learning Incentives Of Students And Teachers: Results From A Social Experiment In Mexican High Schools	JPE	2015	1	0	individuals	school
Supreet Kaur, Michael Kremer, Sendhil Mullainathan	Self-Control At Work	JPE	2015	0	0	individual	individual
Nicholas Bloom, James Liang, John Roberts, Zhichun Yang	Does Working From Home Work? Evidence From A Chinese Experiment	QJE	2015	1	0	individuals	individual
Karthik Muralidharan, Venkatesh Sundararaman	The Aggregate Effect Of School Choice: Evidence From A Two-Stage Experiment In India	QJE	2015	1	0	individuals	village
Vivi Alatas, Abhijit Bannerjee, Rema Hanna, Ben Olken, Ririn Purnamasari, Matthew Wai-Poi	Self-Targeting: Evidence From A Field Experiment In Indonesia	JPE	2016	1	0	individual	village
Adnan Khan, Asim Khwaja, Benjamin Olken	Tax Farming Redux: Experimental Evidence On Performance Pay For Tax Collectors	QJE	2016	1	0	tax circle	property tax "circles"
Andrew Beath, Fotini Christia, Georgy Egorov, Ruben Enikolopov	Electoral Rules And Political Selection: Theory And Evidence From A Field Experiment In Afghanistan	ReStud	2016	1	0	politician	village

Table A2: Summary statistics: program evaluation RCTs in top journals, 2001-2016 – sensitivity to sample

	Median			Mean		
Variable	Default	Inclusive	Exclusive	Default	Inclusive	Exclusive
Sample represents larger population?	0	0	0	0.31	0.34	0.41
Size of sampling frame	10,885	10,885	18,356	681,918	636,142	883,224
Units treated	5,340	4,696	5,674	13,564	13,572	15,140
Clustered randomization?	1	1	1	0.62	0.62	0.68
Mean size of randomization unit	26	31	50	166.62	173	207

This table reports summary statistics for measures of experimental scale for randomized controlled trials published in Econometrica, American Economic Review, the Quarterly Journal of Economics, Review of Economics and Statistics and the Journal of Political Economy between January 2001 and July 2016 which we categorized as primarily "program evaluations" (as opposed to mechanism experiments). The table shows mean and median values for key variables using three different sample definitions of "program evaluations:" the preferred sample, the preferred sample augmented to include borderline cases, and the preferred sample reduced to exclude borderline cases. Counting metrics are defined in "primary units of analysis," which we define as the level at which the studies' primary outcomes are measured (e.g. the household). "Sample represents larger population?" is an indicator equal to one if the paper reports systematically drawing its evaluation sample from any larger population of interest. "Size of sampling frame" is the size of the frame sampled (equal to size of the evaluation sample itself if no larger frame is indicated). "Units treated" is the number of units treated by the organization implementing the intervention being studied. "Clustered randomization?" is an indicator equal to one if randomization was assigned in geographic groupings larger than the primary analysis unit, and "mean size of randomization unit" is the average number of primary analysis units per cluster (equal to 1 for unclustered designs).

Table A3: Summary statistics: program evaluation RCTs in top journals, 2001-2016 (measured with individual units)

Variable	25th %	Median	75th %	Mean	SD	N
Sample represents larger population?	0	0	1	0.31	0.47	29
Size of sampling frame	706	14,596	52,924	2,321,303	10,813,222	26
Units treated	473	5,468	30,000	61,210	216,984	29
Cluster randomized?	1	1	1	0.83	0.38	29
Mean size of randomization unit	15	70	324	1,030	2,745	28

This table reports summary statistics for measures of experimental scale for randomized controlled trials published in Econometrica, American Economic Review, the Quarterly Journal of Economics, Review of Economics and Statistics and the Journal of Political Economy between January 2001 and July 2016 which we categorized as primarily "program evaluations." Counting metrics are defined in "individual units of analysis," which we define as the total number of individuals within the primary unit of analysis. "Sample represents larger population?" is an indicator equal to one if the paper reports systematically drawing its evaluation sample from any larger population of interest. "Size of sampling frame" is the size of the frame sampled (equal to size of the evaluation sample itself if no larger frame is indicated). "Units treated" is the number of units treated by the organization implementing the intervention being studied. "Clustered randomization?" is an indicator equal to one if randomization was assigned in geographic groupings larger than the primary analysis unit, and "mean size of randomization unit" is the average number of primary analysis units per cluster (equal to 1 for unclustered designs).

Table A4: Treatment effect distributions from simulated sub-sampling

Sample Selection	Mean	SD	5th %	95th %
All districts	-25.23	13.51	-47.99	-3.39
Single district (unweighted)	-23.2	31.04	-72.9	26.41
Single district (reweighted)	-15.06	25.02	-61.59	16.34

This table records summary statistics of the estimated average treatment effect of an intervention (Smartcards) on a primary outcome (leakage) using data from Muralidharan et al. (Forthcoming). Each row summarizes treatment effects estimated from 500 simulated sub-samples of the original data. In the "all districts" exercise we sampled 157 mandals (the unit of randomization) with replacement from the full set of 8 study districts, and then used these data to estimate the treatment effect. In the "single district (unweighted)" exercise we first randomly chose a single district (with probability equal to the proportion of surveyed households in that district), sampled 157 mandals with replacement from that district, and then used these data to estimate a treatment effect. In the "single district (reweighted)" exercise we sampled mandals as in the "unweighted" version but then estimated a treatment effect using a weighted regression. We calculated these weights by estimating the probability that a given mandal was in the bootstrap sample as a function of all available demographic information, and then using the (inverse of) these propensities to weight the treatment effect estimation.

Figure A1: Distribution of Treatment Effects From Different Sample Restrictions, 500 Simulations

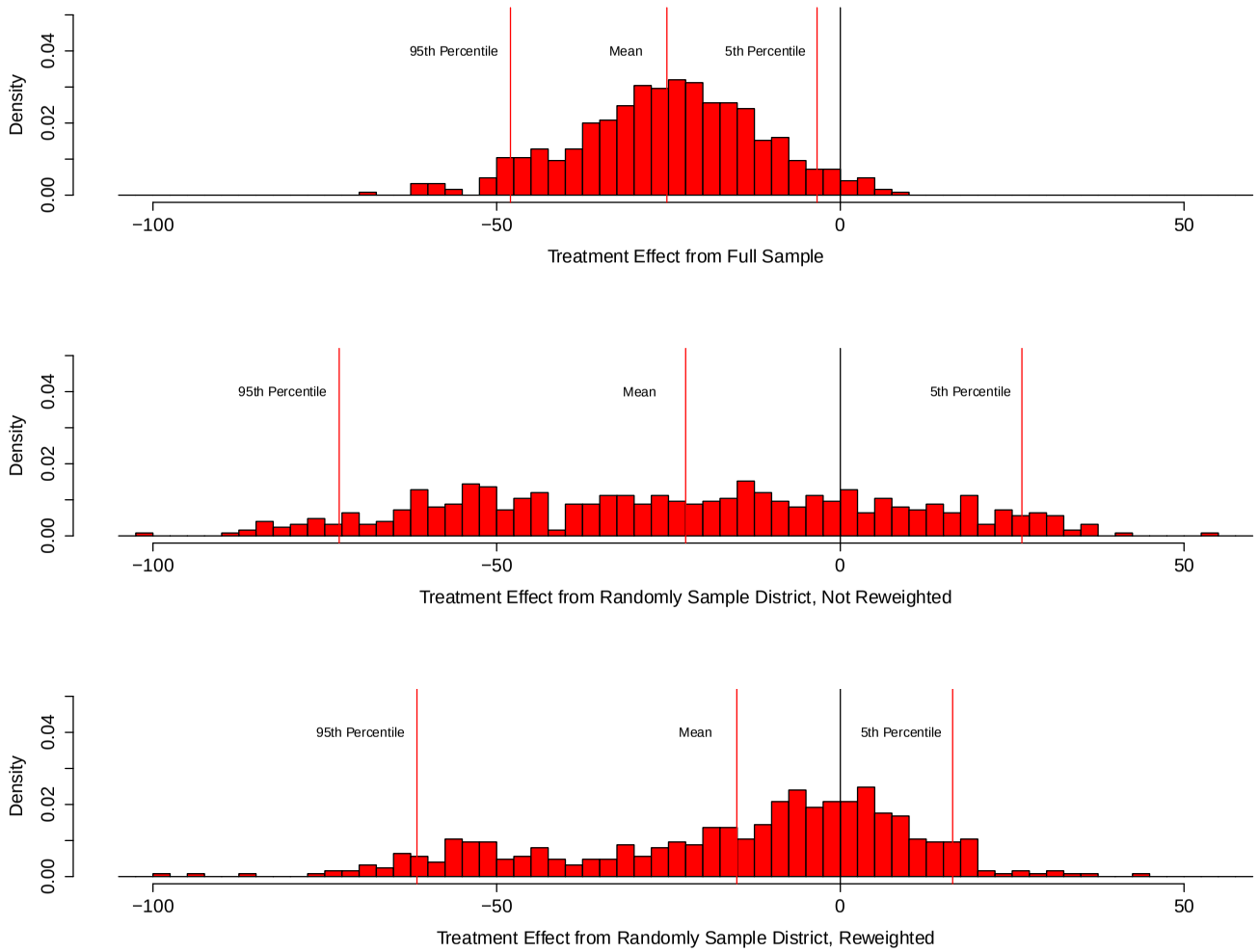


Figure shows the distribution of treatment effects from 500 simulations of data from Muralidharan et al. (Forthcoming) using different sample restrictions. Details on construction of samples are provided in the notes for Table A4.