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. (2012)), 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 significantly – experiments are relatively small with regard 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 Economic Studies 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 Rep. of Larger Pop. (Y/N)	0	0	1	0.31	0.47	29
Size of Sampling Frame	10,442	52,655	1,016,446	2,622,857	10,775,056	26
Number of Treated Units	5,018	30,662	116,250	210,019	436,830	28
Randomized at Cluster Level (Y/N)	1	1	1	0.83	0.38	29
Size of Unit of Randomization	21	131	493	1,608	3,491	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 Economic Studies 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 %
Full Sample	-25.23	13.51	-47.99	-3.39
Randomly Selected District	-23.2	31.04	-72.9	26.41
Re-Weighted Randomly Selected District	-13.58	25.83	-66.26	18.9

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

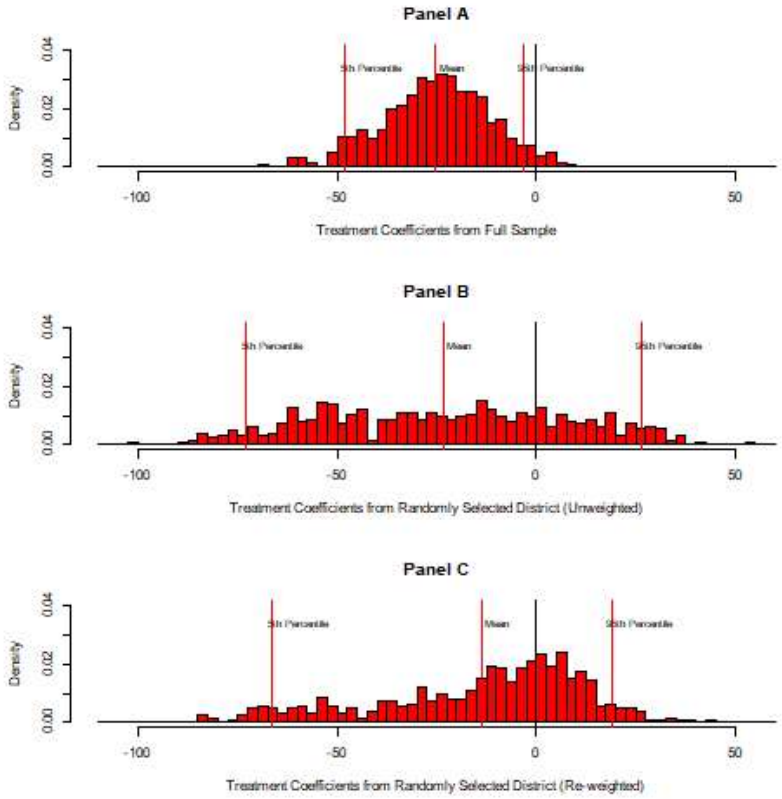

Figure shows the distribution of treatment effects from 500 simulations of data from Muralidharan et al. (Forthcoming) using different sample restrictions. Details on construction of samples are provided in the notes for Table A4.