The Journal of Economic Perspectives

A journal of the
American Economic Association

Fall 2017

The Journal of
Economic Perspectives

A journal of the American Economic Association

Editorial offices:
Journal of Economic Perspectives
American Economic Association Publications
2403 Sidney St., #260
Pittsburgh, PA 15203
email: jep@jepjournal.org

The *Journal of Economic Perspectives* gratefully acknowledges the support of Macalester College.
Registered in the US Patent and Trademark Office (®).

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA.
Printed by LSC Communications, Owensville, Missouri, 65066, USA.
No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.
THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Fall 2017, Vol. 31, No. 4. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 31 • Number 4 • Fall 2017

Symposia

Health Insurance and Choice

From Experiments to Economic Policy

Articles

Features

Statement of Purpose

The *Journal of Economic Perspectives* attempts to fill a gap between the general interest press and most other academic economics journals. The journal aims to publish articles that will serve several goals: to synthesize and integrate lessons learned from active lines of economic research; to provide economic analysis of public policy issues; to encourage cross-fertilization of ideas among the fields of economics; to offer readers an accessible source for state-of-the-art economic thinking; to suggest directions for future research; to provide insights and readings for classroom use; and to address issues relating to the economics profession. Articles appearing in the journal are normally solicited by the editors and associate editors. Proposals for topics and authors should be directed to the journal office, at the address inside the front cover.

Policy on Data Availability

It is the policy of the *Journal of Economic Perspectives* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Details of the computations sufficient to permit replication must be provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the above requirements cannot be met.

Policy on Disclosure

Authors of articles appearing in the *Journal of Economic Perspectives* are expected to disclose any potential conflicts of interest that may arise from their consulting activities, financial interests, or other nonacademic activities.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 3–22

Delivering Public Health Insurance Through Private Plan Choice in the United States

Jonathan Gruber

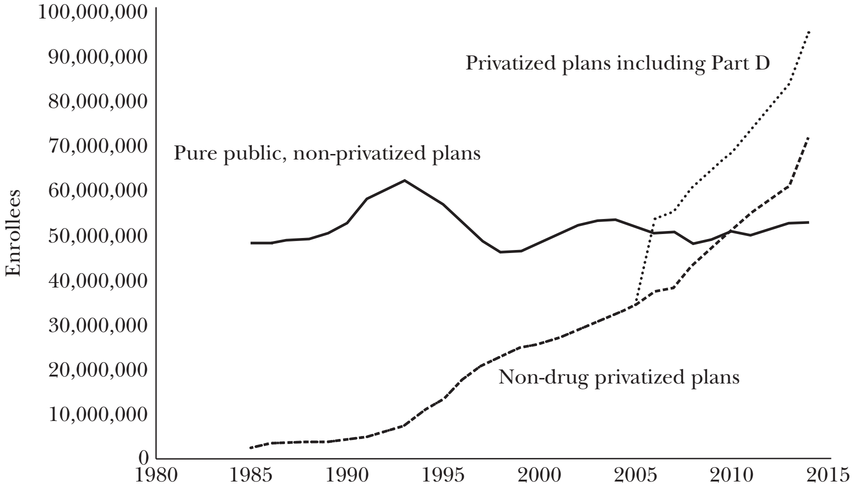
The United States has seen a sea change in the way that publicly financed health insurance coverage is provided to low-income, elderly, and disabled enrollees. When programs such as Medicare and Medicaid were introduced in the 1960s, the government directly reimbursed medical providers for the care that they provided, through a classic "single payer system." Since the mid-1980s, however, there has been an evolution towards a model where the government subsidizes enrollees who choose among privately provided insurance options. Currently, almost one-third of Medicare enrollees are in privately provided insurance plans for all of their medical spending, and another 43 percent of Medicare enrollees have standalone private drug plans through the Medicare Part D program. More than three-quarters of Medicaid enrollees are in private health insurance plans. Those receiving the subsidies made available under the Patient Protection and Affordable Care Act of 2010 do so through privately provided insurance plans that are reimbursed by the government. In most of these cases, individuals have a choice across these private insurance options, and choices are typically from quite large choice sets.

Figure 1 illustrates these trends. The figure has three lines that contrast enrollment in pure public versus privately delivered insurance. The line that starts highest shows enrollment in "single payer only" government insurance since the mid-1980s, excluding enrollment in private plans for Medicare and Medicaid enrollees. The

■ *Jonathan Gruber is Ford Professor of Economics, Massachusetts Institute of Technology, and Research Associate, National Bureau of Economic Research, both in Cambridge, Massachusetts. His email address is gruberj@mit.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.3 doi=10.1257/jep.31.4.3

Figure 1
Medicare and Medicaid Enrollment: Public versus Privatized



Note: This figure contrasts enrollment in pure public versus privately delivered insurance. The line that starts highest shows enrollment in "single payer only" government insurance since the mid-1980s, excluding enrollment in private plans for Medicare and Medicaid enrollees. The lower line shows enrollment in publicly financed private plans through Medicare and Medicaid, while the additional line starting in 2006 includes enrollment for prescription drug coverage only through private Medicare Part D plans.

lower line shows enrollment in publicly financed private plans through Medicare and Medicaid, while the additional line starting in 2006 includes enrollment for prescription drug coverage only through private Medicare Part D plans. If such Part D coverage is included in the privatized bucket, then by 2006 there were more beneficiaries enrolled in some type of private coverage than in government-sponsored single-payer programs. Even absent Part D, starting in 2010 there was more enrollment in privatized plans. Public insurance in the United States is now primarily a privately run endeavor, at least in terms of enrollment. Expenditures on private plans remain below expenditures on directly insured government care, which reflects the fact that, as discussed below, the healthiest enrollees are more likely to join private programs, but even in expenditure terms, the gap is shrinking.

This trend is in contrast to the rest of the developed world. There is a large diversity of health care financing systems around the globe, including direct public health provision in the United Kingdom, publicly provided insurance in Canada, and mandatory nonprofit private insurance in Germany and Switzerland. These countries have had significant changes to their health care systems in recent years, but in each case, the basic context of insurance has remained the same—private has stayed private, while public has stayed public.

This remarkable evolution of the provision of US public health insurance has potentially wide-ranging impacts. Moreover, it is central to current health care policy debates. For example, one of the leading policy disagreements is over moving the Medicare program to a full "premium support" program, in which elders will choose from a wide variety of private (and potentially public) insurance options. Another controversy is over whether to replace the Medicaid program with subsidies to private insurance plans among which enrollees would choose.

Privatized delivery of public health insurance appears to be here to stay, with debates now focused on how much to expand its reach. Yet such privatized delivery raises a variety of thorny issues. Will choice among private insurance options lead to adverse selection and market failures in privatized insurance markets? Can individuals choose appropriately over a wide range of expensive and confusing plan options? Will a privatized approach deliver the promised increases in delivery efficiency claimed by advocates? What policy mechanisms have been used, or might be used, to address these issues?

A growing literature in health economics has begun to make headway on these questions. In this essay, I discuss that literature and the lessons for both economics more generally and health care policymakers more specifically. I begin with a review of the original structure of the major US health insurance programs: Medicare and Medicaid, and discuss the introduction of privatized insurance delivery into Medicare and Medicaid through managed care plans, as well as the broader introduction of exchange-based models through the Medicare Part D program and the Affordable Care Act. I turn to a brief heuristic discussion of the issues raised in shifting from public to private provision. I review the economic literature on what is known about the transition to such a privatized model, and discuss the key lessons and policy issues that must be addressed in evaluating future expansions of privatized public insurance. I highlight in particular two of the key policy issues going forward, risk adjustment and choice consistency, which in turn are the focus of the subsequent two papers in this symposium.

The Changing Nature of Public Health Insurance

The Original Structure of Medicare and Medicaid

Over the long run, the largest single expansion in the US welfare state was the introduction of Medicare and Medicaid in the mid-1960s. These programs started small but grew rapidly. By 2015, these programs accounted for government expenditures of $1.2 trillion, which is 37 percent of total national health expenditures and is larger than total private health insurance expenditure (based on the National Health Expenditure Data from the Centers for Medicare & Medicaid Services). These programs were established as classic single-payer health insurance systems. Individuals chose only whether to enroll in the program, and had no other choices to make with respect to their insurance coverage. The government contracted directly with providers for the cost of providing care to enrollees.

Medicare was originally set up with two different components. Part A is primarily focused on hospital expenditures and is financed by a dedicated payroll tax with no additional premium contribution from individuals; individuals are automatically enrolled at age 65 if they (or their spouses) have sufficient work history. Part B is primarily focused on physician expenditures. Upon turning 65, eligible individuals are able to enroll in Part B by paying a monthly premium set to cover 25 percent of the cost of the program, with the remaining 75 percent coming from general revenue financing.

Medicaid was established as a state program to cover individuals who had low income and assets, but it also covers single mothers, elderly, and disabled. The program is administered by the states but jointly financed by the state and federal governments; the federal government share is an inverse function of state income and averages 57 percent of program costs. Over time, the program has expanded to have higher income limits for certain groups, such as children and pregnant mothers, along with a broader expansion under the Affordable Care Act to all low-income families (in participating states). Although Medicaid is sometimes described as a program for those with low incomes, seniors and the disabled account for 61 percent of program spending (but are only 23 percent of enrollees) (Paradise 2017).

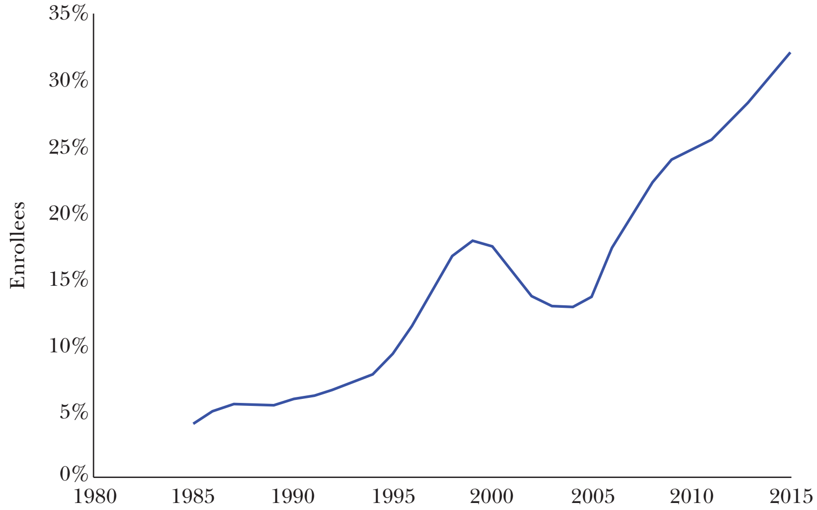
Medicare: The Advent of Choice

Medicare was incredibly successful in expanding insurance coverage among the elderly, and in reducing their exposure to risks of high medical spending (Finkelstein and McKnight 2008). But program costs also exploded in its early years, rising from $64 million in federal expenditures in 1966 to $32.1 billion in 1980 (Gruber 2015). This rapid rise led policymakers to focus on controlling costs through two different channels. The first was regulatory changes in the reimbursement of providers, including the introduction of physician fee schedules and "prospective reimbursement" of hospitals (in which hospitals are paid for services according to pre-established fees). The second cost-saving approach, and the focus of the discussion here, was the introduction of an option for enrollees to join "managed care" plans, starting in 1985.

Managed care plans, traditionally known as health maintenance organizations (HMOs), represented a very different model than the existing model. Under a fee-for-service model, providers billed for the cost of each service and were reimbursed by the government. Under this alternative model, private plans would be paid a fixed amount by the government to cover all of the medical spending of enrollees; the government shifted the entire risk to the managed care entity. This part of Medicare has been called by many names over time but is currently referred to as the Medicare Advantage program.

The government initially reimbursed these plans 95 percent of the county's per enrollee spending on traditional Medicare. Over time, various floors were introduced which raised reimbursement to selected Medicare Advantage plans based on location (for example, urban versus rural plans). The government also "risk adjusts" payments to Medicare Advantage plans to try to reflect the underlying patient health in these plans. In 2006, the government moved to a bidding system whereby plans could submit a bid for the expected costs of providing Medicare-like services to recipients; if the bids were below the county-level benchmarks, three-quarters of

Figure 2
Share Medicare A & B Enrollment in Private Managed Care Organizations (Medicare Advantage)
(out of total Medicare A & B enrollment)



the difference was rebated to consumers in the form of lower premiums or richer benefits, while one-quarter is rebated to the government. The Patient Protection and Affordable Care Act of 2010 kept this system in place, but substantially reduced the county benchmarks, while adding aspects such as bonuses paid to plans based on plan quality.

Figure 2 shows the time series of enrollment in this Medicare managed care option. Enrollment grew rapidly throughout the 1990s, declined significantly due to reimbursement reductions in the early 2000s, and then grew rapidly again after reimbursement increases in the mid-2000s. Despite reimbursement reductions in the 2010 Affordable Care Act, enrollment growth in managed care has not slowed and currently stands at a peak of almost one-third of Medicare enrollees.

From an enrollee perspective, Medicare Advantage plans present a clear tradeoff. Relative to traditional Medicare, these plans significantly limit provider choice and manage care through tools such as utilization review (processes for pre-approval of medical procedures).

On the other hand, traditional Medicare features significant patient cost-sharing, with a large deductible for hospital utilization and an (uncapped) 20 percent coinsurance for physician care, while Medicare Advantage plans typically feature much smaller copayments. Moreover, before drug coverage was added to traditional Medicare through the introduction of Part D in 2006, Medicare Advantage plans covered prescription drugs while traditional Medicare did not. Elders had a variety of options for covering these out-of-pocket costs under Medicare: the poorest elders received coverage of the costs through Medicaid, while others could purchase private "Medigap" insurance that covered the costs. Medicare Advantage was another such option, which was typically much cheaper than Medigap plans. Overall, by signing up for Medicare Advantage, patients were trading off provider choice for a reduction in cost-sharing.

Managed Care Takes Over Medicaid

In its early years, the Medicaid program was delivered exclusively (like Medicare) through fee-for-service reimbursement of providers by states. States set fee schedules and other reimbursement rules for providers who saw Medicaid patients. These fees were typically well below private reimbursement rates, and even below Medicare rates, so that many providers were reluctant to see Medicaid enrollees (Gruber 2015).

In the early 1990s, in an effort to save costs, states began to contract with private Medicaid managed care organization (MMCOs) that were reimbursed a fixed amount by states for absorbing all of the financial risk of covering Medicaid enrollees. Figure 3 shows the share of Medicaid enrollees in MMCOs. This figure grew from under 10 percent of enrollment in 1990 to over 50 percent by 1998, and has grown steadily since, now standing at almost 80 percent of enrollment. The share of Medicaid spending in MMCOs is much lower, however, because the growth in managed care has been much stronger in the low-cost nonelderly/nondisabled population than for the elderly and disabled, who as noted earlier, account for a majority of total Medicaid spending.

The process by which Medicaid enrollees end up in Medicaid managed care organizations has varied over states and across time. In the early years of the program, a number of states started with voluntary systems whereby enrollees could choose whether to enroll. Over time, however, a number of those states have moved to mandatory systems where enrollees are assigned to a MMCO. Currently, there is a mix of the two types of systems.

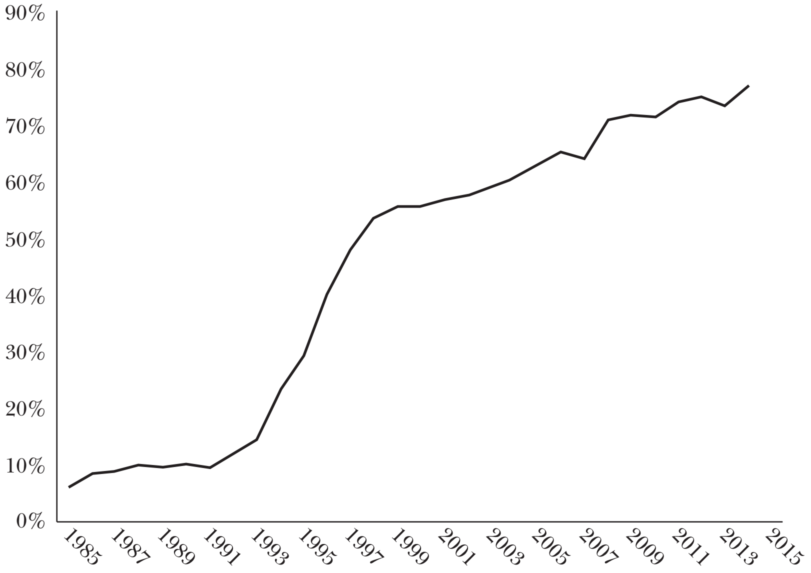
Medicare Part D: The Codification of Choice

A watershed moment in the shift from publicly provided health insurance to publicly financed but privately provided insurance was the introduction of the Medicare Part D program.[1] When Medicare was established in 1965, it covered most medical needs for the elderly and disabled, including hospital and doctor costs, but it excluded coverage for prescription drugs. By the 1990s, the advancement of prescription drug treatments for common illnesses among the elderly drew attention to this gap in Medicare coverage.

The expansion of Medicare to prescription drug coverage could have taken two different paths. In one approach, a drug benefit could have been added to the Medicare program directly, with the government negotiating directly with drug

[1] The following discussion lifts heavily from Gruber (2015). For an excellent overview of Part D, see Duggan, Healy, and Scott Morton (2008).

Figure 3
Share of Medicaid Enrollees in Medicaid Managed Care Organizations (MMCOs)



companies to hold down drug prices. However, after a contentious debate, the policy that was chosen was for government to provide subsidies to private insurers who would then offer prescription drug coverage to the elderly, either through health maintenance organizations or as a stand-alone prescription drug. The Medicare Part D benefit marked the first time that a major government health insurance plan was totally privately provided.

Those eligible for Medicare Parts A and B were after 2006 made eligible for this new Medicare Part D benefit. Individuals chose from a large variety of prescription drug plans that covered at least a basic set of benefits, but typically covered more. Enrollees faced a premium that on average covered 25 percent of the cost of the program. At the same time, Medicare Advantage plans could continue to include drug coverage as part of their coverage packages. This resulted in a large number of options for seniors to choose prescription drug coverage; in 2009, the typical senior could choose from 48 different plans that offered prescription drug coverage (Abaluck and Gruber 2011).

The Affordable Care Act: The Future of Choice?

The most recent major change in choice among private-sector providers for the delivery of government-financed health insurance is the state-based exchanges in the Patient Protection and Affordable Care Act of 2010. Prior to the passage of this law, individuals buying insurance outside of an employer setting faced a fragmented market which featured significant barriers to enrollment, such as highly imperfect information on potential options and active selection against the sickest enrollees (Gruber 2015). The Affordable Care Act put in place a new organizational structure for the health insurance market that sought to address these shortcomings. A new regulatory structure for the market disallowed health-based insurance offerings or pricing, established a minimum set of benefits that must be included in all plans, and organized the market around a set of "metallic tiers" of coverage with similar benefits structures.

Perhaps most significantly, the federal government provided large tax credits to offset the cost of these insurance plans for low-income enrollees. These tax credits were structured so that low-income families pay a capped share of their income (from 2 percent of income at the poverty line to 9.5 percent of income at four times the poverty line), and the government absorbed the rest of the cost for the second-lowest-cost plan in the silver metallic tier (the second lowest tier). The Affordable Care Act also included an individual mandate that penalized individuals who did not sign up for insurance coverage, further incentivizing individuals to enroll through exchanges (unless they had other options available such as employer or government-provided insurance).

Enrollment in the state-based exchanges grew rapidly in their first year of operation, but has stabilized more recently. In a minority of states, the exchanges are state-run entities; in the majority, enrollees enroll through a federal website that shows each state's choices. Initially, 2.6 million enrollees were in state-administered programs and 5.4 million in the federally run exchange. Over time, as some states have stopped providing their own exchanges, enrollment in state-administered exchanges has stagnated at 2.8 million while enrollment in the federally run exchange has risen to 8.7 million. The differences between state- and federally-administered exchanges are not obvious to the enrollee, but Frean, Gruber, and Sommers (2017) find that enrollment grew much faster, and consumers were more responsive to subsidies, in state-administered exchanges (perhaps due to other correlated state outreach programs).

What Are the Tradeoffs in Moving to Choice of Private Plans?

This radical shift from purely public insurance options to a choice of private plans raises a number of important economic issues. In this section, I review these issues. I then turn to the available evidence on these issues from the US experience with privatized insurance delivery.

The issues that arise in moving to a privatized system of insurance choice can be divided into the allocative and production side. On the allocative side, the benefit of choice is that it allows chosen health insurance to reflect preference heterogeneity, allowing allocative efficiency across plans. Individuals differ in their demand for insurance for a number of reasons, ranging from demographic characteristics to tastes for risk. Forcing individuals into a plan that doesn't reflect their preferences imposes an allocative cost. For example, Lucarelli, Prince, and Simon (2008) use aggregate data on plan market shares to conduct a study of how plan features affect demand for prescription drug insurance plans, and they estimate sizeable welfare losses from limiting the option set facing seniors. But the study assumes that seniors are choosing optimally, and therefore, by that definition, restricting their choice set can only be harmful.

On the other hand, having choice across plans invokes two costs. The first is adverse selection: if individuals match their individual tastes, then the plans preferred by the least healthy enrollees may end up with the highest prices. Adverse selection does not necessarily doom an insurance market; as highlighted by Finkelstein and McGarry (2006), it depends on the correlation between tastes for risk and health status. For example, in some insurance markets, those who have lower risk-tolerance and are generally safer tend to buy more insurance, not less. But adverse selection has generally been documented to be a significant issue in health insurance markets, as documented strikingly by Cutler and Reber (1998). A series of articles by Einav and Finkelstein, nicely reviewed in this journal in 2011, show how to measure the consequences of adverse selection. Einav and Finkelstein find that while adverse selection exists across insurance plan choices, it has relatively modest welfare costs.[2]

One reason why the consequences of adverse selection might not be large relates to the second potential cost of choice: choice frictions. These include the standard problems of switching costs, as well as behavioral problems that arise in evaluating a complicated set of health insurance choices. For example, Handel (2013) documents strong evidence for inertia in private health plan choices and argues that it offsets the pressures towards adverse selection; however, Polyakova (2016) finds in the context of Medicare Part D that inertia leads to an increase in adverse selection. So the sign of this effect is not obvious. Likewise, a set of articles reviewed below find "choice inconsistencies" in the context of the Medicare Part D program, whereby individuals are not optimizing their plan choice. A recent paper by Handel, Kolstad, and Sinnewijn (2015) reviews the evidence on how choice frictions impact the demand side of the insurance market and discuss the fact that "more is not always better" when it comes to decisions facing choice frictions.

The other set of issues around privatized choice arises on the producer side. Once again, the standard economics argument is clear: allowing choice across plans will put competitive pressure on those plans to deliver care efficiently, whereas a monopoly public insurer faces no such pressure. Once again, although the basic intuition continues to apply, this issue is much more complicated in the context of insurance markets than in goods markets.

One reason is that it is more difficult to define "efficiency" in an insurance market. Normally economists would define productive efficiency in terms of producing at minimum costs per quality-adjusted unit of output. But quality-adjustment is very difficult in health insurance—particularly when concepts of "quality" may vary across individuals. This insight also further amplifies adverse

[2] Note that these are just the welfare costs of adverse selection across plans, not the welfare costs of adverse selection in terms of accessing insurance markets. The latter may be much larger, as discussed in Hackmann, Kolstad, and Kowalski (2015).

selection concerns, in that private insurers will have an incentive to target outcomes that are most valued by the healthiest potential enrollees.

The tradeoff here is nicely illustrated by two recent articles. On the one hand, Einav, Finkelstein, and Polyakova (2016) document that while public insurance programs typically incorporate uniform cost-sharing across prescription drugs, private prescription drug plans under Medicare Part D distinguish cost-sharing across categories of drugs that are differentially price-elastic, which is more efficient. On the other hand, Geruso, Layton, and Prinz (2016) document that within private plans on the state-level health insurance exchanges established by the Patient Protection and Affordable Care Act of 2010, prescription drug cost sharing is designed in a manner to discourage enrollment among less-healthy enrollees. Similar studies of plan benefits designed to promote virtuous selection are discussed below; in that discussion, where possible, efficiency will be defined relative to observable health outcomes such as mortality.

Two other issues involve the capabilities of public and private insurers. One issue is the ability of public versus private insurers to reduce unit prices for health care, which turns on the dynamics of competitive bidding versus regulatory price setting. The other issue is the ability of private insurers to impose care management restrictions that may be politically difficult to impose with public insurance. These issues are discussed further below in the context of Medicare Part D and Medicare Advantage.

Evidence on the Effects of Privatized Delivery of Health Insurance

Medicare Advantage: More Efficient Care, But Sorting by Risk

There is a rapidly growing literature on the Medicare Advantage program, comparing it to outcomes under traditional Medicare. This literature has drawn three important conclusions.

First, patients who choose Medicare Advantage are much healthier than patients who choose traditional Medicare (as reviewed by Brown, Duggan, Kuziemko, and Woolston 2014). Those who move into Medicare Advantage have costs that are 20–37 percent lower than those who remain in traditional Medicare. Of course, such a gap could theoretically be due either to differences in health or to more efficient provision of care. However, Batata (2004) notes that if those moving into Medicare Advantage were, on average, as healthy as those who remain in traditional Medicare, then when individuals move across programs there should be no change in average costs of traditional Medicare. In fact, as more individuals move from traditional Medicare to Medicare Advantage, the average costs of traditional Medicare do rise. Batata finds that the marginal cost of traditional Medicare disenrollees who move to Medicare Advantage is $1,030 lower, or 20–30 percent cheaper than the average cost of traditional Medicare enrollees.

More recently, Brown et al. (2014) show that substantial health differences exist between enrollees in Medicare Advantage and traditional Medicare. They find that the health care spending of those switching to Medicare Advantage plans from fee-for-service have total annual health care costs that are $2,850 (or 45 percent) lower than those in traditional Medicare, and have risk scores (measures of underlying patient burden) that are 20–30 percent lower than in traditional Medicare. They also find that those who are in good health are more satisfied with Medicare Advantage than with traditional Medicare, and as for the 3 percent of Medicare Advantage enrollees who switch back to fee-for service each year, it is the sickest enrollees who are most likely to switch back.

Second, health care utilization and costs are much lower under Medicare Advantage than under traditional Medicare (for an excellent review, see McGuire, Newhouse, and Sinaiko 2011). Of course, this literature faces the important problem mentioned above—enrollment in Medicare Advantage is not random. To account for this, previous studies have taken a variety of approaches. One subset of research has estimated cross-sectional models that include a rich set of controls for individual's age, health status, and related factors, assuming that there are no remaining unobserved differences between those who choose to enroll in managed care and those who do not (Curto, Einav, Finkelstein, Levin, and Battacharya 2017; Landon et al. 2012). Another branch of studies has used instrumental variable approaches, with their methods assuming that certain factors (for example, the penetration of Medicare Advantage in a local market) influence plan choice but do not affect utilization (Mello, Sterns, and Norton 2002). Yet another strand of this literature has used longitudinal data to follow individuals over time and compare the evolution of Medicare spending or other outcomes of interest among those switching between Medicare Advantage and traditional Medicare and those not switching; Brown et al. (2014) examine cases of voluntary switching, while Parente, Evans, Schoenman, and Finch (2005) examine cases of switching following exit of the plan that people had been using.

Recently, Duggan, Vabson, and I have taken a different approach to the selection problem by exploiting plan exits from counties in New York in the early 2000s (Duggan, Gruber, and Vabson forthcoming). Individuals who were enrolled in Medicare Advantage in those counties were exogenously removed from the program, and we study the subsequent effect on this population's use of health care relative to other counties and relative to the traditional Medicare population that was not affected by exits. Consistent with the previous literature, we find significant increases in inpatient health care utilization and costs when individuals are exogenously moved from Medicare Advantage to traditional Medicare, although in this study we were unable to examine outpatient care.

Third, the literature has generally found no significant impact on patient outcomes, although this evidence on this point is more limited than on the previous two findings. Duggan, Vabson, and I find that the exit from Medicare Advantage when private plans ceased to operate had no impact on patient mortality, hospital readmissions, or reported hospital quality (Duggan, Gruber, and Vabson forthcoming). Taken together with the evidence on utilization and costs, this literature suggests that Medicare Advantage is delivering care more efficiently than traditional Medicare.

Medicaid Managed Care Organizations: Mixed Evidence on Efficiency

The literature on managed care within Medicaid has delivered somewhat less-clear conclusions than studies of managed care within Medicare. Where there is enrollee choice between fee-for-service Medicaid and Medicaid managed care organizations, there is some evidence of selection. For example, Glied, Sisk, Gorman, and Ganz (1997) find that those who voluntarily selected MMCOs in New York City were healthier. As a result, the large literature that cross-sectionally compares MMCO enrollees to those in traditional Medicaid should be taken with some caution (for a review of that literature, see Kaestner, Dubay, and Kenney 2005).

To address this selection, several papers have relied on the introduction of state-level mandates that the enrollees join Medicaid managed care plans, and then applied a difference-in-difference approach. The results using this approach are quite mixed. Duggan (2004) studied such mandates in California and found that they led to increased government spending, while Harman, Hall, Lemak, and Duncan (2014) showed that such a mandate in Florida lead to lower Medicaid spending. Herring and Adams (2011) and Duggan and Hayford (2013) use national samples and find no effect of mandated movements into Medicaid managed care organizations. Duggan and Hayford, do, however find that managed care reduced Medicaid spending in states that had generous baseline fee-for-service provider reimbursement rates in Medicaid. Their results suggest that managed care achieved savings for Medicaid mainly through the government's ability to negotiate lower prices with health plans and that managed care plans had little impact on the actual practice of Medicaid providers. Marton, Yelowitz, and Talbert (2014) confirm this finding using a case study of Kentucky, where they find that MMCOs lowered costs under some contracting arrangements but not under others.

There is more limited evidence on how a shift to managed care affects patient outcomes. Duggan (2004) found that MMCOs improved the quality of health care, while Aizer, Currie, and Moretti (2007), who examined prenatal care and birth outcomes in California, found that managed care decreased the quality of prenatal care and increased the incidence of low-birth-weight, pre-term births, and neonatal mortality.

Medicare Part D: Savings, But Choice Inconsistency

In the decade since the Medicare Part D program was introduced, a substantial literature has emerged to study this new mode of government insurance delivery.

One important conclusion is that the competition between private insurers has led to lower-than-expected expenditures for the program. Expenditures in the first 10 years of Part D were $147 billion below what was initially projected ($740 billion) by the Congressional Budget Office (2013). Duggan and Scott Morton (2010, 2011) show that this result arises from reduced drug prices for customers insured under Medicare Part D relative to what was expected from this new market. This finding is perhaps surprising, because economic intuition suggests that insuring consumers will make their demand less price-elastic, leading to higher prices in the imperfectly competitive prescription drug market. But in this case, insurers used their "formulary design" to negotiate lower prices. That is, insurers could choose the set of drugs for which consumers would face lower or higher prices (or which would be unavailable), and they used this power to negotiate lower prices from manufacturers in return for better formulary placement.

This important finding suggests a sizeable fiscal benefit from the competitive structure of Medicare Part D. Of course, it does not prove that prices would not have been even lower if Medicare had negotiated directly for lower drug prices. In the Medicaid program, the government negotiates prices to be paid for drugs using a "most favored nation" clause that insists that drug prices paid by Medicaid be no higher than those paid by other payers. This provision has significantly lowered drug prices in the Medicaid program, albeit with important external impacts on pricing to other payers (Duggan and Scott Morton 2006). An important ongoing policy debate is the efficacy of government price regulation versus reliance on a competitive mechanism in the context of the imperfectly competitive market for pharmaceuticals.

Another area of significant study examines the ability of consumers to choose across the dozens of Medicare Part D options that they have available. Abaluck and Gruber (2011) and Heiss, Leive, McFadden, and Winter (2012) show that the vast majority of enrollees do not chose the cost-minimizing plan. More specifically, Abaluck and Gruber (2011) argue that there are two major "choice inconsistencies" under Part D: individuals are much more sensitive to premium differentials across plans than to out-of-pocket cost differentials; and consumers consistently overweight "salient" plan characteristics based on their overall impacts, not their impacts on those specific consumers. We estimate that there are welfare losses of 25–30 percent from these choice inconsistencies. Ketcham, Lacarelli, Miravete, and Roebuck (2012) argue that foregone savings from the program are minimized over time through learning; in contrast, Abaluck and Gruber (2016a) find that choice inconsistencies actually grow over time with little individual or cohort learning. For a debate over these findings, see Abaluck and Gruber (2016c) and Ketcham, Kuminoff, and Powers (2016).

There is also evidence of some adverse selection arising in the Medicare Part D market. Polyakova (2016) shows that there is a significantly higher enrollment in the most generous plans by the sickest enrollees. Conversely, Carey (2017) and Lavetti and Simon (2016) demonstrate that formularies are designed strategically by insurers to avoid the sickest enrollees.

On the one hand, these studies suggest that competition in Medicare Part D has restrained price increases over time. On the other hand, this market has experienced choice failures and adverse selection. The existing evidence does not show whether a single-payer option without choice would have done better.

State-Based Exchanges: Increasing Volatility

The latest move towards privatization of public insurance delivery is through the exchanges established with the Patient Protection and Affordable Care Act of 2010. While these exchanges have been in operation for only a few years, a sizeable literature has emerged on their effects. In addition, this literature draws on the experience of the Massachusetts insurance "connector," similar in design to the state-based exchanges, which has been in operation since 2006.

The state-based exchanges established by the Affordable Care Act have been marked by considerable volatility in pricing and plan offerings. For the first two years, the market was relatively stable. In the initial year of 2014, the average premium was 15 percent below what had been projected by the Congressional Budget Office when the law passed (Cohn 2016). Moreover, enrollees in most areas had a wide variety of choices: 74 percent of enrollees had products from at least three insurance companies to choose from, and only 520 of the 3,142 US counties only offered one plan, with most of those 520 counties being very low population (Avery et al. 2015).

But premiums rose and options were reduced for open enrollment in 2016 and even more so for open enrollment in 2017. Premiums rose by 22 percent on average in 2017. Choices fell, with only 57 percent of enrollees having three or more companies to choose from and 21 percent of enrollees having only one company. Despite this large increase in premiums, after taking into account the initial low pricing by insurers, premiums were on average just about where the Congressional Budget Office had projected them to be when the law passed.

This pricing volatility reflects several factors. Predicting insurance costs in a brand new market is challenging. The exchanges were not designed to be the only route to purchase individual coverage (except in Washington, DC, where the exchange had a monopoly), so insurers had to predict which of the existing non-group-insured would migrate to the new exchange products as well as which uninsured would enroll.

This effort was further hampered by important policy developments. First, the Obama administration allowed many more individuals than expected to remain in "grandfathered" insurance plans, which kept some of the best risks out of the new pools. Second, administrators allowed generous use of "special enrollment periods" which allowed flexibility for individuals to enter the market, but also promoted further selection. Third, the Republican Congress refused to appropriate funds for the "risk corridors," the payments to/from unprofitable/profitable insurance plans that were supposed to help buffer the financial risk facing entrants in this new market. These risk corridor payments were supposed to amount to as much as $8.2 billion over 2014 and 2015, but only $362 million in payments were made; these payments could have significantly offset the losses to insurers from early underpricing (Cohn and Young 2017).

A final reason for the volatility is a much higher rate of plan-switching in the state-based exchanges than has been previously experienced in health insurance choice environments. Switching rates in employer-sponsored insurance are quite low and in Medicare Part D are around 10 percent (Abaluck and Gruber 2016a), but in the exchanges the switching rates have been roughly 35 percent (Centers for Medicare and Medicaid Services 2017, p. 7). This likely reflects the fact that subsidies are tied to the second-lowest-cost "silver" plan, so as plans move in the ranking due to relative changes in premiums, enrollees must switch to preserve their subsidy levels.

Structuring Privatized Offerings of Publicly Funded Health Insurance

This ongoing shift to privatized public insurance raises a key set of policy issues involving adverse selection by the private providers of health insurance and problems of the inconsistency of choices of consumers who are confronted with multiple private insurance plans. In this section, I provide an overview of these key policy issues, which are then the focus of the two more detailed papers following in the symposium.

Adverse Selection

Adverse selection is endemic in these types of insurance-market choice environments. There are two approaches to addressing this problem. The first is risk adjustment, which involves using redistributive mechanisms to offset the losses from adverse selection, and in this way to improve market functioning. In theory, such redistributive mechanisms could happen either before care is provided when the government makes payments to insurance companies, or after the care has been provided and the costs are known, or some mixture of the two. For example, the federal government has introduced risk adjustment mechanisms into the reimbursement structure for Medicare Advantage plans that target reimbursement not just to the age and gender of the enrollee, but also to their underlying health care costs (measured before care is provided, based on previous utilization). Evidence on the effectiveness of these government efforts is mixed, with Brown et al. (2014) arguing that selection actually worsened in the early years of this program, and Newhouse, Price, McWilliams, Hsu, and McGuire (2014) arguing that the program was ultimately successful in removing selection from Medicare Advantage. The Geruso and Layton paper in this symposium includes a detailed discussion of risk adjustment approaches.

The second approach to addressing adverse selection is through supply-side restrictions on the structure of the market. As an example, consider how policy should react to low-cost "limited network" plans that restrict the insured to lower-cost providers in return for a lower premium. These plans have proved enormously popular in general, and in particular on the state-based exchanges. Moreover, such plans can represent a useful cost-saving innovation, because the restrictions in provider choice can deliver substantial savings that can be passed on to the consumer. While there have been few studies, there are no documented significant costs to health care outcomes from these network restrictions; Gruber and McKnight (2016) find that the introduction of such limited network plans for public employees in Massachusetts led to significantly lower health expenditures with no adverse impact on care delivery.

However, Sheppard (2016) does find that limited network plans exacerbate adverse selection by excluding the providers who deliver care to the least profitable patients—who also tend to be the patients who place greater value on broader networks of providers with fewer restrictions. This finding raises a critical design tradeoff in setting "minimum standards for network adequacy among insurers": higher standards for network adequacy will reduce the cost savings from narrow networks but will also reduce the risk that sicker enrollees are denied access to necessary providers.

A similar tradeoff arises in the contentious policy arena of setting minimum benefits standards for health insurance plans. Any publicly financed insurance plan must define what constitutes the "insurance" that it being paid for, and there is broad agreement that such insurance should include physician and hospital care. But what about prescription drugs? Mental health? Maternity coverage? These services are vital to some populations, but not others.

In a world without selection, the question of which benefits to include in the minimum package is just a question of distribution. Including a richer set of minimum benefits will redistribute from nonusers of those benefits to users. For example, including maternity coverage in a minimum benefit set will reduce a gap in insurance prices that would otherwise exist between male and female enrollees. But with adverse selection, plans that include more generous benefits must be aware of the risk that they will attract heavier users of health care, which means that such plans may be overpriced or perhaps not be offered at all. A standard approach to this issue is to define a minimum set of benefits so that no one is left without "essential" services, but the definition of "essential" is highly sensitive and open to debate.

Addressing Choice Inconsistencies

As discussed earlier, there is a lively debate over the nature of consumer choice between insurance plans, which has often focused on results from Medicare Part D. To the extent that consumer choices are inconsistent, meaning that a number of consumers have a tendency to choose plans that are not most cost-effective for their personal needs, competition between private health insurance firms will not work effectively. There is as yet little work evaluating the consistency of choices in other health insurance environments such as Medicare Advantage, Medicaid managed care, and the state-based Affordable Care Act exchanges.

If future researchers tend to find that choice inconsistencies are rather widespread and lasting, what policies might be used to address them? Abaluck and Gruber (2016b) discuss this question in detail, using data on health insurance plan choice across roughly 250 school districts in the state of Oregon. They document significant choice inconsistencies in the health insurance decisions made by employees of these school districts and then model three alternative interventions designed to address these inconsistencies. First, government might require enrollees to re-enroll at regular intervals, thus pushing back against inertia and promoting more active choice. However, we find that this approach does nothing to improve the quality of choices. This conclusion is consistent with the finding noted earlier, from Abaluck and Gruber (2016a), that those who switch to a new plan don't choose any more consistently than they originally did. Second, government might provide better decision support to improve choices. However, randomized access to a decision support tool did not significantly improve choices because the recommendation from the decision support tool was often ignored. Third, government might restrict the number of choices, or the number of dimensions across which choices may vary. We find that with improvements in "choice architecture"—and in particular smaller choice sets—individuals make much better choices. This finding of an important role for choice architecture is demonstrated by the work of Ericson and Starc (2016) as well as others. This evidence and the policy implications are reviewed in the paper by Ericson and Sydnor in this symposium.

There is a relatively small but emerging literature raising the issue of whether seniors can choose appropriately across their various Medicare Advantage options. For example, McWilliams, Afendulis, McGuire, and Landon (2011) show that seniors were less likely to choose Medicare Advantage at all when they had more than 15 plans to choose from, and that elders with cognitive limitations were less responsive to benefit generosity across plans. Sinaiko and Zeckhauser (2016) show strong inertia in plan choice between Medicare Advantage and traditional Medicare.

Conclusion

The United States has been experiencing a radical transformation in the delivery of public health insurance that has largely flown under the radar. Policy discussions often treat public health insurance programs as a sort of government monopsony, whereas in fact the majority of enrollees in public insurance programs financed by the federal government are in private insurance plans funded by the government. The pace of this shift from public to private provision has been notable in recent years. In particular, the delivery of expanded health insurance through the use of state private insurance exchanges through the Patient Protection and Affordable Care Act of 2010 has added to the roles of publicly financed, privately provided insurance.

The existing experience with privatized delivery of public insurance has had successes and failures. There is clear evidence that care is delivered more efficiently under Medicare Advantage plans than under traditional Medicare, and the costs of the privatized Part D prescription drug benefit program were much lower than projected due to innovative insurer plan design. On the other hand, there is less clear evidence for efficiency in privatized Medicaid plans, and there is clear evidence of the problems that enrollees face making choices in this complicated environment.

How should policymakers evaluate this tradeoff? One certainty is that unfettered private choice is not optimal, and that, as demonstrated empirically by Ericson and Starc (2016), structuring choice can improve choice quality. Better risk-adjustment should also be incorporated in order to reduce incentives for private insurers to compete over the health risk of their enrollees rather than the delivery of efficient care. In principle, with a sufficiently structured choice environment and proper risk adjustment, the productive benefits of choice can outweigh the allocative difficulties, particularly given the strong evidence for more efficient health care delivery by private Medicare Advantage plans.

Further expansion of the private choice-based model of public insurance is an ongoing subject of policy discussions. For example, one occasionally discussed possibility is that the entire Medicare system could be moved to a "premium support," or defined contribution, model. Under this model, all Medicare enrollees would be entitled to a voucher amount that they could apply to choice from a set of private plans—and perhaps to a public plan as well (the so-called "public option"). Designing a choice-based model for delivery of public health insurance, and evaluating how it compares with a publicly provided alternative, requires addressing issues of adverse selection by private providers, choice inconsistencies by consumers, and the characteristics of competitive forces and government mandates in this setting.

■ *I am grateful to Enrico Moretti and the editors and to Amy Finkelstein for helpful comments, to Jason Abaluck for a series of collaborations that greatly advanced my thinking on these topics, and to Chris Behrer for research assistance.*

References

Abaluck, Jason, and Jonathan Gruber. 2011. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program." *American Economic Review* 101(4): 1180–210.

Abaluck, Jason, and Jonathan Gruber. 2016a. "Evolving Choice Inconsistencies in Choice of Prescription Drug Insurance." *American Economic Review* 106(8): 2145–84.

Abaluck, Jason, and Jonathan Gruber. 2016b. "Improving the Quality of Choices in Health Insurance Markets." NBER Working Paper 22917.

Abaluck, Jason, and Jonathan Gruber. 2016c. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program: Reply." *American Economic Review* 106(12): 3962–87.

Aizer, Anna, Janet Currie, and Enrico Moretti. 2007. "Does Managed Care Hurt Health? Evidence from Medicaid Mothers." *Review of Economics and Statistics* 89(3): 385–99.

Avery, Kelsey, Mathias Gardner, Emily Gee, Elena Marchetti-Bowick, Audrey McDowell, and Aditi Sen. 2015. "Health Plan Choice and Premiums in the 2016 Health Insurance Marketplace." ASPE Research Brief, US Department of Health and Human Services, October 30. https://aspe.hhs.gov/system/files/pdf/135461/2016%20Marketplace%20Premium%20Landscape%20Issue%20Brief%2010-30-15%20FINAL.pdf.

Batata, Amber. 2004. "The Effect of HMOs on Fee-for-Service Health Care Expenditures: Evidence from Medicare Revisited." *Journal of Health Economics* 23(5): 951–63.

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston. 2014. "How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program." *American Economic Review* 104(10): 3335–64.

Carey, Colleen. 2017. "Technological Change and Risk Adjustment: Benefit Design Incentives in Medicare Part D." *American Economic Journal: Economic Policy* 9(1): 38–73.

Centers for Medicare & Medicaid Services. No date. National Health Expenditure Data. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/index.html.

Centers for Medicare & Medicaid Services. 2017. *Health Insurance Marketplaces 2017 Open Enrollment Period: January Enrollment Report for the Period: November 1– December 24, 2016*. Washington, DC: Department of Health and Human Services.

Cohn, Jonathan. 2016. "Obamacare May Be Getting More Expensive, But It's Still Cheaper Than Predicted." *Huffington Post,* August 1. http://www.huffingtonpost.com/entry/obamacare-insurance-premiums_us_579f89e1e4b08a8e8b5ecb84.

Cohn, Jonathan, and Jeffrey Young. 2017. "Republicans Are Crying about Obamacare Problems They Helped Create." *Huffington Post,* March 18. http://www.huffingtonpost.com/entry/republicans-obamacare-problems_us_58cc48e3e4b0be71dcf4d685.

Congressional Budget Office. 2013. *A Premium Support System for Medicare: Analysis of Illustrative Options*. Washington, DC: Congressional Budget Office.

Curto, Vilsa, Liran Einav, Amy Finkelstein, Jonathan D. Levin, and Jay Battacharya. 2017. "Healthcare Spending and Utilization in Public and Private Medicare." NBER Working Paper 23090.

Cutler, David M., and Sarah J. Reber. 1998. "Paying for Health Insurance: The Trade-off between Competition and Adverse Selection." *Quarterly Journal of Economics* 113(2): 433–66.

Duggan, Mark. 2004. "Does Contracting Out Increase the Efficiency of Government Programs? Evidence from Medicaid HMOs." *Journal of Public Economics* 88(12): 2549–72.

Duggan, Mark, Jonathan Gruber, and Boris Vabson. Forthcoming. "The Consequences of Health Care Privatization: Evidence from Medicare Advantage Exits." *American Economic Journal: Economic Policy***.**

Duggan, Mark, and Tamara Hayford. 2013. "Has the Shift to Managed Care Reduced Medicaid Expenditures? Evidence from State and Local-Level Mandates." *Journal of Policy Analysis and Management* 32(3): 505–35.

Duggan, Mark, Patrick Healy, and Fiona Scott Morton. 2008. "Provider Prescription Drug Coverage to the Elderly: America's Experiment with Medicare Part D." *Journal of Economic Perspectives* 22(4): 69–92.

Duggan, Mark, and Fiona M. Scott Morton. 2006. "The Distortionary Effects of Government Procurement: Evidence from Medicaid Prescription Drug Purchasing." *Quarterly Journal of Economics* 121(1): 1–30.

Duggan, Mark, and Fiona Scott Morton. 2010. "The Effect of Medicare Part D on Pharmaceutical Prices and Utilization." *American Economic Review* 100(1): 590–607.

Duggan, Mark G., and Fiona Scott Morton. 2011. "The Medium-Term Impact of Medicare Part D on Pharmaceutical Prices." *American Economic Review* 101(3): 387–92.

Einav, Liran, and Amy Finkelstein. 2011. "Selection in Insurance Markets: Theory and Empirics in Pictures." *Journal of Economic Perspectives* 25(1): 115–38.

Einav, Liran, Amy Finkelstein, and Maria Polyakova. 2016. "Private Provision of Social Insurance: Drug-Specific Price Elasticities and Cost Sharing in Medicare Part D." NBER Working Paper 22277.

Ericson, Keith M. Marzilli, and Amanda Starc. 2016. "How Product Standardization Affects Choice: Evidence from the Massachusetts Health Insurance Exchange." *Journal of Health Economics* 50: 71–85.

Finkelstein, Amy, and Kathleen McGarry. 2006. "Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market." *American Economic Review* 96(4): 938–58.

Finkelstein, Amy, and Robin McKnight. 2008. "What Did Medicare Do? The Initial Impact of Medicare on Mortality and Out of Pocket Medical Spending." *Journal of Public Economics* 92(7): 1644–68.

Frean, Molly, Jonathan Gruber, and Benjamin Sommers. 2017. "Premium Subsidies, the Mandate, and Medicaid Expansion: Coverage Effects of the Affordable Care Act." *Journal of Health Economics* 53: 72–86.

Geruso, Michael, Timothy J. Layton, and Daniel Prinz. 2016. "Screening in Contract Design: Evidence from the ACA Health Insurance Exchanges." NBER Working Paper 22832.

Glied, Sherry, Jane Sisk, Sheila Gorman, and Michael Ganz. 1997. "Selection, Marketing, and Medicaid Managed Care." NBER Working Paper 6164.

Gruber, Jonathan. 2015. *Public Finance and Public Policy*. 5th ed. London: Worth Publishers.

Gruber, Jonathan, and Robin McKnight. 2016. "Controlling Health Care Costs through Limited Network Insurance Plans: Evidence from Massachusetts State Employees." *American Economic Journal: Economic Policy* 8(2): 219–50.

Hackmann, Martin B., Jonathan T. Kolstad, and

Amanda E. Kowalski. 2015. “Adverse Selection and an Individual Mandate: When Theory Meets Practice.” *American Economic Review* 105(3): 1030–66.

Handel, Benjamin R. 2013. “Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts.” *American Economic Review* 103(7): 2643–82.

Handel, Benjamin R., Jonathan T. Kolstad, and Johannes Spinnewijn. 2015. “Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets.” NBER Working Paper 21759.

Harman, Jeffrey S., Allyson G. Hall, Christy H. Lemak, and R. Paul Duncan. 2014. “Do Provider Service Networks Result in Lower Expenditures Compared with HMOs or Primary Care Case Management in Florida’s Medicaid Program?” *Health Services Research* 49(3): 858–77.

Heiss, Florian, Adam Leive, Daniel McFadden, and Joachim Winter. 2012. “Plan Selection in Medicare Part D: Evidence from Administrative Data.” NBER Working Paper 18166.

Herring, Bradley, and E. Kathleen Adams. 2011. “Using HMOs to Serve the Medicaid Population: What Are the Effects on Utilization and Does the Type of HMO Matter?” *Health Economics* 20(4): 446–60.

Kaestner, Robert, Lisa Dubay, and Genevieve Kenney. 2005. “Managed Care and Infant Health: An Evaluation of Medicaid in the US.” *Social Science and Medicine* 60(8): 1815–33.

Ketcham, Jonathan D., Nicolai V. Kuminoff, and Christopher A. Powers. 2016. “Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program: Comment.” *American Economic Review* 106(12): 3932–61.

Ketcham, Jonathan D., Claudio Lucarelli, Eugenio J. Miravete, and M. Christopher Roebuck. 2012. “Sinking, Swimming, or Learning to Swim in Medicare Part D.” *American Economic Review* 102(6): 2639–73.

Landon, Bruce E., Alan M. Zaslavsky, Robert C. Saunders, L. Gregory Pawlson, Joseph P. Newhouse, and John Z. Ayanian. 2012. “Analysis of Medicare Advantage HMOs Compared with Traditional Medicare Shows Lower Use of Many Services during 2003–09.” *Health Affairs* 31(12): 2609–17.

Lavetti, Kurt, and Kosali Simon. 2016. “Strategic Formulary Design in Medicare Part D Plans.” NBER Working Paper 22338.

Lucarelli, Claudio, Jeffrey Prince, and Kosali Simon. 2008. ”Measuring Welfare and the Effects of Regulation in a Government-Created Market: The Case of Medicare Part D Plans.” NBER Working Paper 14296.

Marton, James, Aaron Yelowitz, and Jeffery C. Talbert. 2014. “A Tale of Two Cities? The Heterogeneous Impact of Medicaid Managed Care.” *Journal of Health Economics* 36: 47–68.

McGuire, Thomas G., Joseph P. Newhouse, and Anna D. Sinaiko. 2011. “An Economic History of Medicare Part C.” *Millbank Quarterly* 89(2): 289–323.

McWilliams, J. Michael, Christopher Afendulis, Thomas McGuire, and Bruce Landon. 2011. “Complex Medicare Advantage Choices May Overwhelm Seniors—Especially Those with Impaired Decision Making.” *Health Affairs* 30(9): 1786–94.

Mello, Michelle M., Sally C. Stearns, and Edward C. Norton. 2002. “Do Medicare HMOs Still Reduce Health Services Use after Controlling for Selection Bias?” *Health Economics* 11(4): 323–40.

Newhouse, Joseph P., Mary Price, J. Michael McWilliams, John Hsu, and Thomas G. McGuire. 2014. “How Much Favorable Selection Is Left in Medicare Advantage?” NBER Working Paper 20021.

Paradise, Julia. 2017. *10 Things to Know about Medicaid: Setting the Facts Straight.* Menlo Park, CA: Kaiser Family Foundation.

Parente, Stephen, William Evans, Julie Schoenman, and Michael Finch. 2005. “Health Care Use and Expenditures of Medicare HMO Disenrollees.” *Health Care Finance Review* 26(3): 31–43.

Polyakova, Maria. 2016. “Regulation of Insurance with Adverse Selection and Switching Costs: Evidence from Medicare Part D.” *American Economic Journal: Applied Economics* 8(3): 165–95.

Shepard, Mark. 2016. ”Hospital Network Competition and Adverse Selection: Evidence from the Massachusetts Health Insurance Exchange.” NBER Working Papers 22600.

Sinaiko, Anna, and Richard Zeckhauser. 2016. “Forced to Choose, Again: The Effects of Defaults on Individuals in Terminated Health Plans.” Chap. 23 in *Nudging Health: Health Law and Behavioral Economics,* edited by I. Glenn Cohen, Holly Fernandez Lynch, and Christopher Robertson. Baltimore: Johns Hopkins University.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 23–50

Selection in Health Insurance Markets and Its Policy Remedies

Michael Geruso and Timothy J. Layton

Selection (adverse or advantageous) is *the* central problem that inhibits the smooth, efficient functioning of competitive health insurance markets. Even—and perhaps especially—when consumers are well-informed decision makers and insurance markets are highly competitive and offer choice, such markets may function inefficiently due to risk selection. Selection can cause markets to unravel with skyrocketing premiums and can cause consumers to be under- or overinsured. In its simplest form, adverse selection arises due to the tendency of those who expect to incur high health care costs in the future to be the most motivated purchasers. The costlier enrollees are more likely to become insured rather than to remain uninsured, and conditional on having health insurance, the costlier enrollees sort themselves to the more generous plans in the choice set. These dual problems represent the primary concerns for policymakers designing regulations for health insurance markets.

In practice, identifying selection problems and designing policy responses is not always straightforward. A natural starting point for uncovering selection distortions is a comparison of the chronic health conditions of consumers who elect more- versus less-generous insurance. However, selection can play out in complex ways that extend beyond issues of who remains uninsured and who chooses which plan.

■ *Michael Geruso is Assistant Professor of Economics, University of Texas, Austin, Texas. Timothy J. Layton is Assistant Professor of Health Care Policy, Harvard Medical School, Boston, Massachusetts. Both authors are Faculty Research Fellows, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are mike.geruso@austin.utexas.edu and layton@hcp.med.harvard.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.23 doi=10.1257/jep.31.4.23

Consider a market in which two essentially identical plans compete for enrollees and earn zero profits. All consumers opt to purchase insurance (in one plan or the other) due to a generous government subsidy. Because consumers perceive these plans as indistinguishable, they choose between them seemingly at random. Thus, neither plan differentially attracts sick or healthy consumers, and no one decides to remain uninsured. At first, it might appear that there are no selection problems. But what if we notice that neither plan offers good coverage for cancer treatments? It might be that both plans believe that offering such coverage would attract especially costly patients and would drive the plan to insolvency. Both plans thus attempt to screen out these patients by offering coverage that is unappealing to cancer patients. Because the two plans act identically, neither succeeds in avoiding cancer patients, and both plans get an equal share of such patients. But the result is that cancer patients cannot find good coverage in the market, and currently healthy consumers cannot find a plan to protect them against the possibility of needing cancer care in the future. Despite the fact that that we observe no systematic sorting of sick consumers between the available plans, this too would be a selection-driven distortion: there is a missing market for cancer coverage due to the anticipation of how the sick would sort themselves if a certain kind of coverage were offered.

In this essay, we review the theory and evidence concerning selection in competitive health insurance markets and discuss the common policy tools used to address the problems it creates. We begin in the next section by outlining some important but often misunderstood differences between two types of conceptual frameworks that economists use to think through selection. The first, the *fixed contracts* approach, takes insurance contract provisions as given and views selection as influencing only insurance prices in equilibrium. This is useful for thinking through selection problems on the extensive margin like "death spirals," in which the healthy choose to remain uninsured and prices can spiral upwards as the consumers remaining enrolled are increasingly sick and costly. This framework is also helpful in understanding how government subsidies to purchase insurance can arrest this feedback mechanism. The second broad framework, the *endogenous contracts* approach, treats selection as also influencing the design of the contract itself, including the overall level of coverage and coverage for services that are differentially demanded by sicker consumers. This approach is useful for understanding "cream skimming," in which various contract features are designed to attract or deter certain kinds of enrollees, such as with our cancer patients above. This modeling framework is also helpful in understanding the motivation for policy tools like risk adjustment and requirements that insurance policies offer certain minimum essential health benefits.

After outlining the selection problems, we discuss four commonly employed policy instruments that affect the extent and impact of selection: 1) premium rating regulation, including "community rating"; 2) consumer subsidies or penalties to influence the take-up of insurance; 3) risk adjustment, which is a policy that adjusts payments to private insurance companies based on the expected health care costs of enrollees; and 4) contract regulation, often involving rules for the minimum of what must be covered by the privately provided health insurance contract. We discuss the economics of these policy approaches and present available empirical evidence on their consequences, with some emphasis on the two markets that seem especially likely to be targets of reform in the short and medium term: Medicare Advantage (the private plan option available under Medicare) and the state-level individual insurance markets.

Adverse Selection, Through the Lenses of Fixed and Endogenous Contracts

We describe two conceptual approaches to modeling adverse selection: a *fixed contract approach*, in which the available insurance policies are taken as given, and an *endogenous contracts approach*, in which insurance providers design the elements of contracts in a way that seeks to attract those with relatively low expected use of health care. Neither the fixed nor endogenous contracts framework is superior in all applications; instead, each is useful in characterizing certain types of selection problems and in designing appropriate regulatory responses. As we work through these ideas, we will use the term *selection* primarily to describe actions by consumers as they sort themselves into and out of insurance and across plans. We will use the term *screening* to differentiate the actions of plans as they respond to and anticipate consumer sorting.

Generally, adverse selection arises because consumers have private information that is not accessible to the insurance provider or because the insurer is prohibited from conditioning insurance prices on observable information like age, gender, and medical history, effectively making such information private. Given a fixed set of contracts (not an innocuous assumption), consumers who expect, based on their private information, to have low health care expenses will select themselves into the lower-cost plans, while those who expect to have high health care expenses will select themselves into higher-cost, higher-coverage plans. In principle, selection in insurance markets need not be adverse in this sense, but in health insurance, the clear empirical pattern across a variety of market settings is one of adverse selection.

What we call the *fixed contracts* approach follows this intuition very directly. It models consumers as selecting across a very limited set of insurance contracts on the basis of private health status information. For example, under this framework a researcher might study, or a policymaker might consider, a market with two differentiated plans: a high-coverage contract (such as a generous preferred provider organization plan with low cost sharing) and a low-coverage contract (such as a high-deductible plan). The reason why the specific coverage levels of the high- and low-benefits contracts are chosen is typically unmodeled. In empirical applications, the *fixed contracts* assumption usually amounts to assuming that whatever plans are currently observed in the market are the only plans that could exist, regardless of changes to consumer demand, changes to the value of the outside option such as charity care, or changes to the regulatory structure of the market.

Under the fixed contracts approach, the insurance company does not respond to the pressures of selection by altering the generosity of the high-benefit contract—say, by limiting the size of the provider network, requiring larger copayments for certain services, or using "utilization review" to limit the provision of certain kinds of care. In equilibrium, the forces of selection affect plan prices, and perhaps whether a segment of the market completely unravels with certain plan types exiting altogether, but that is all.

Under this framework, efficiency losses occur when selection distorts contract prices and consequently consumers do not sort efficiently across contracts (or across the choice between insurance and uninsurance). Einav and Finkelstein (2011, in this journal) treat the welfare economics of this case in detail in a series of intuitive diagrams. We refer the reader to that article for a full discussion of this framework. Here, we briefly note that the typical adverse selection result is that more generous coverage sells only at very high prices. Higher prices for the generous contract imply fewer enrollees, who will be costlier on average than those in the less-generous contract. In a competitive equilibrium, these fewer, costlier enrollees imply higher break-even prices, completing the feedback loop between prices, average costs, and enrollment in the generous contract. Low-cost, healthy consumers, who value generous coverage more than the social cost of providing it to them, are not offered it at a price that they are willing to pay. Thus, from the perspective of social efficiency too few consumers enroll in more generous coverage.

The fixed contracts framework has been popular among empirical economists because in addition to allowing for relatively straightforward characterizations of equilibrium prices and enrollment given exogenous variation in insurance prices, it allows for straightforward welfare analysis (for example, Einav, Finkelstein, and Cullen 2010). It also appears to characterize accurately the employer-sponsored insurance setting, the channel through which the majority of Americans receive health insurance. However, a number of observations about private insurance markets like the Marketplaces created by the Patient Protection and Affordable Care Act of 2010 and Medicare Advantage raise questions about whether a framework that assumes only prices (and not other contract features) respond to selection is sufficient for fully characterizing the impact of adverse selection on social welfare in these settings. Two particular observations raising concerns are: the widespread presence of insurance contracts that offer 1) narrow provider networks and 2) restrictive drug formularies.

First, consider narrow provider networks. In the state-level health insurance Marketplaces, Bauman, Bello, Coe, and Lamb (2015) find that around 55 percent of available plans have hospital networks that are deemed "narrow," meaning that they include from 31–70 percent of area hospitals, or "ultra-narrow," including less than 30 percent of area hospitals. With respect to physician networks, Polsky and Weiner (2015) find that 11 percent of plans have networks with fewer than 10 percent of physicians in the area and 65 percent of plans have networks with fewer than 40 percent of physicians in the area—with networks being even narrower for physicians specializing in the treatment of cancer. In Medicare Advantage, Jacobson, Trilling, Neumann, Damico, and Gold (2016) find that the average plan in a given county covers only around 50 percent of hospitals in the county, and 9 of 20 cities studied in their report do not have a single plan with a "broad" hospital network, defined as more than 70 percent of hospitals in the area.

Second, drug formularies, which list consumer cost-sharing amounts for prescription medications, are often much more restrictive in the state-level Marketplaces than in employer-sponsored plans. The state-level Marketplace plans are much more likely than employer plans to place entire therapeutic classes of drugs on high cost-sharing "specialty" tiers, exposing consumers to significantly more out-of-pocket spending, and to place nonprice barriers on drugs like prior authorization or step therapy requirements (Jacobs and Sommers 2015; Geruso, Layton, and Prinz 2016).

While narrow networks and restrictive formularies could be efficient reactions to consumer preferences for lower-cost insurance products, they could also be driven by adverse selection.[1] Consider an insurer designing a drug formulary in a competitive market setting in which plans cannot directly reject applicants and in which regulation prevents price discrimination between applicants. Competition induces such an insurer to increase the generosity of the formulary until the costs of additional generosity exceed the benefit to its enrollees. But as the insurer improves the quality of its formulary, it may attract a different set of customers who are likely to have high health care costs. For example, using claims data discussed below, it is straightforward to observe that consumers who use immunosuppressant drugs to treat conditions like rheumatoid arthritis generate costs in excess of $30,000 annually, while paying a premium that is a small fraction of that amount. In a market setting that outlaws premium discrimination (also known as "medical underwriting"), all insurers may offer symmetrically poor coverage for classes of drugs like immunosuppressants to discourage such patients from joining their plans. Deviations from that strategy could yield the unhappy outcome for the insurer of cornering the market on these unprofitable patients, with limited ability to spread the costs of such patients across the rest of the risk pool. This dynamic could result in an inefficient equilibrium where all of the available insurance contracts provide too little coverage for immunosuppressants. This type of inefficiency would be missed when using a fixed contracts framework that assumes that adverse selection only distorts prices of observed contracts because, in this case, the equilibrium set of contracts available for purchase is itself distorted.

In short, adverse selection (and its policy remedies) may affect not only the prices of contracts but also the design of the contracts themselves. We refer to models that allow for this possibility as *endogenous contract* frameworks. Such models may include a continuum of potential insurance contracts, including contracts

[1] Limited networks and restrictive formularies could in principle be a socially efficient reaction to consumer preferences for lower-cost coverage or the outcome of a bargaining game between insurers and hospitals/drug manufacturers (Ho and Lee 2016; Duggan and Scott Morton 2010). However, as noted in the text, such patterns are also consistent with adverse selection.

not currently observed in the market, rather than just a small number of observed contracts like the high-benefit and low-benefit contract example mentioned earlier. The key feature of these models is that they allow selection to influence the design of the contracts that insurers offer in equilibrium rather than assuming that the observed set of contracts represents the entire contract space.

To gain intuition for the endogenous contracts approach, consider the case where there are two types of consumers, healthy (inexpensive) and sick (costly). The healthy do not wish to subsidize the sick, so they demand plans that screen out the sick by offering less-generous coverage at a lower price, leading to a separating equilibrium in which the sick purchase full coverage at a high price and the healthy inefficiently purchase only partial coverage at a lower price (Rothschild and Stiglitz 1976). It is important to understand that the degree of partial coverage is an equilibrium outcome: Will the low plan cover 80 percent of expenses or just 60 percent? This depends on the extent of the difference between sick and healthy consumers, as well as the presence of any risk-based transfer payments enforced by the regulator.

The use of a contract feature as a screening device is commonly known as "cream skimming." More recent theoretical work has examined screening when there are many types of consumers (Azevedo and Gottlieb 2017) and when markets are imperfectly competitive (Veiga and Weyl 2016). In some cases, this type of adverse selection may lead to some types of consumers being unable to purchase insurance with *any* level of coverage (Hendren 2013).

A further complication, which matters in practice for consumers but is missed by the fixed contracts framework, is the multidimensional nature of coverage in modern health insurance contracts. Consider the possibility that costs of physical and mental health care may be covered differently. Assume that both the inexpensive and costly consumer types have similar probabilities of using physical health services, but let the costly type have higher probability of requiring mental health services. Again, the inexpensive consumers wish to avoid subsidizing the costly ones, but now, instead of demanding plans that screen out the sick by limiting *total* coverage, the healthy demand plans that screen out the sick by limiting coverage for mental health services only, while maintaining full coverage for physical health services. This dynamic can lead to a separating equilibrium where the costly patients purchase a contract providing full coverage for physical and mental health services at a high price, while the inexpensive consumers purchase a contract providing full coverage for physical but only partial coverage for mental health services (Glazer and McGuire 2000). Again, this outcome is inefficient if the inexpensive types value full coverage for both physical and mental health services more than the social cost of providing it to them. In other models, *all* consumers, both healthy and sick, are worse off when they are combined in the same market because all plans offer poor coverage for services that the sick are more likely to require (Frank, Glazer, and McGuire 2000; Veiga and Weyl 2016). The various models nested in the endogenous contracts framework differ in their equilibrium concepts, whether they assume perfect competition, and in their restrictions on the contract space, but all result in the equilibrium set of contracts being different from, and usually less generous on average than, the efficient set of contracts.

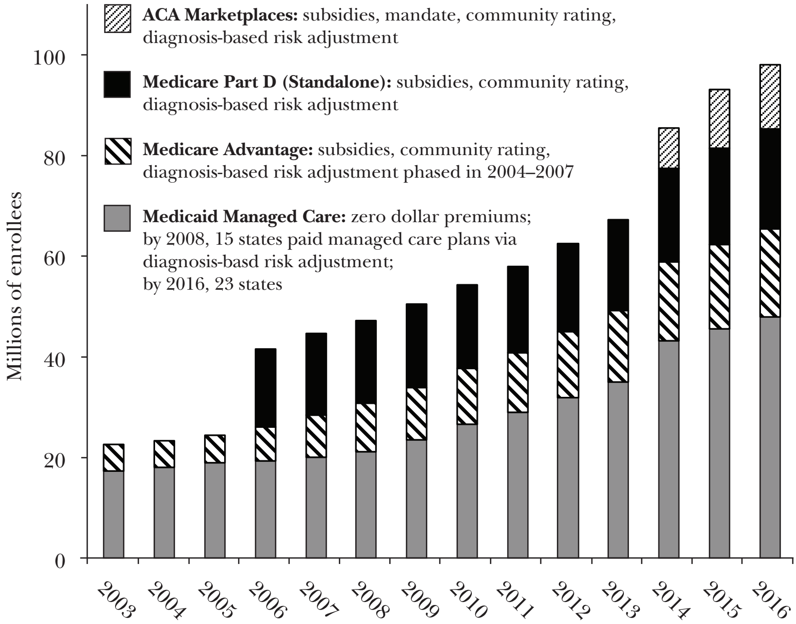
While the endogenous contracts settings are more general than the fixed contracts setting, they are also clearly much more complicated. In part, this is because the contract space is large, with many hard-to-observe and hard-to-measure dimensions—for example, coverage for a specific drug, whether a particular specialist is included in the provider network, or the level of hassle involved with scheduling a visit with an in-network or out-of-network mental health provider. Thus, in contrast to the fixed contracts framework, calculating the welfare consequences of contract distortions has generally involved imposing significant theoretical structure while analyzing calibrated counterfactuals that may extend uncomfortably out-of-sample. The unresolved challenges of estimating welfare in an endogenous contracts framework in a transparent way are a critical avenue for future work, with the potential for important policy applications. In the meantime, however, the endogenous contract problem remains empirically important and motivates much of the regulatory action in Medicare and the Marketplaces—in particular regulations targeting the quality of coverage available rather than just the portion of people who choose to purchase coverage. The lack of a welfare framework equal in elegance to the fixed contracts framework is no reason for economists to ignore these types of market failures.

In the sections that follow, we build on the fixed and endogenous frameworks to discuss the most common selection-related government policies used in regulated private health insurance markets. We begin with premium rating regulations, which are policies primarily aimed at equity concerns but which interact with selection, often making selection problems worse.

Premium Rating Regulations and Community Rating

Outside of large employer settings, health insurance in the United States is increasingly organized around private insurers competing for enrollees in highly regulated and often publicly subsidized markets. As public programs like Medicare and Medicaid turn to private insurers to deliver benefits, selection-related policies have risen in importance. Figure 1 shows the dramatic growth in the use of regulated private health insurance markets to provide public health insurance benefits over the last 15 years. Over 60 percent of Medicaid recipients choose plans in a market-like setting where they face a choice between a public fee-for-service option and a private managed care plan, or more frequently, between multiple private managed care alternatives. In Medicare, 19 million beneficiaries (33 percent) choose to receive their physician and hospital coverage from a private Medicare Advantage plan, and an additional 20 million beneficiaries purchase private prescription drug insurance in the highly subsidized and tightly regulated Medicare Part D program. Finally, for individuals who do not receive health insurance from their employer or from another public program, the Patient Protection and Affordable Care Act

Figure 1
The Rise of Markets, Choice, and Selection Regulation in Public Health Benefits



Note: ACA is the Affordable Care Act.

(ACA) of 2010 introduced state-based Health Insurance Marketplaces ("Marketplaces"), which have provided a new publicly subsidized, privately provided health insurance benefit for millions of lower-income Americans.

The markets/programs listed in Figure 1 share the feature that consumers can choose among competing insurance products and that insurers are not allowed to price discriminate between consumers or to reject applicants. In each of these markets, regulators also require certain minimally acceptable benefits packages, use risk adjustment to compensate insurers for enrolling high-expected-cost patients, and offer subsidies to lower-income enrollees to encourage take-up. We begin here by discussing restrictions on insurers' ability to price discriminate, also known as premium rating restrictions.

Premium rating restrictions govern whether and how prices may vary across consumers for a given insurance product. A complete prohibition against price discrimination within a local rating area is called "community rating." In the case of Medicare Advantage plans, full community rating is used: The small beneficiary contributions to the highly subsidized plan premiums cannot vary across enrollees within the local market, regardless of age, sex, or medical history. The state-level Marketplaces use modified community rating, in which prices can vary within a geographic market only by age and by smoking status in a prescribed way. In most state-level Marketplaces, the premium for a 64 year-old is restricted to be exactly three times the premium for a 21 year-old, with premiums at each intermediate age set using a regulator-specified age-price curve. In the absence of these types of restrictions, one might expect competition to drive an insurer to charge each consumer a premium equal to that buyer's expected cost, thus leading to high premiums for the sick and low premiums for the healthy.

Premium rating restrictions generally exacerbate adverse selection problems, because premium rating *imposes* an information asymmetry between the consumer and insurer. Specifically, insurers are required to ignore signals that would be informative about an individual's expected health care costs when setting prices. Buchmueller and DiNardo (2002) show that the shift of New York's individual and small group health insurance markets to community rating in the 1990s led to consumers shifting from fuller coverage plans to more restrictive health maintenance organizations, consistent with the endogenous contracts literature discussed above. By 2013, prior to the Affordable Care Act taking effect, New York's individual health insurance market had experienced almost a complete "death spiral," with only 17,000 individuals enrolled in the market and 2.1 million individuals uninsured (Rabin and Abelson 2013).

Despite these potential negative consequences, premium rating restrictions are extremely popular among consumers and policymakers and are currently in place in almost all health insurance markets in the United States and in other high-income countries. Why? Fairness motivations are typically cited; indeed, the section of the ACA that establishes premium rating rules is titled "Fair Health Insurance Premiums."

But there is also a clear economic rationale for such rules. It involves long-run risk (Cochrane 1995; Handel, Hendel, and Whinston 2015; Hendren 2017). Risk-averse consumers value not only coverage for fluctuations around their expected annual health spending, such as due to a broken bone; they also value coverage for health state transitions, such as developing diabetes, that may permanently affect their expected health care consumption and thus their health insurance premiums in the absence of premium rating regulations.

Much of the prior literature—from Rothschild and Stiglitz (1976) to Einav and Finkelstein (2011) and some of our own work as well—has focused on the value of insurance in smoothing *one-period* risk, which can be viewed as insurance against the variation in health care spending when fundamental health status is not changing. This focus on one-period risk carries the awkward implication that optimal insurance for an expensive cancer patient may involve a $60,000 premium, because the goal of insurance is to protect that patient from uncertainty over whether treatments cost $50,000 or $70,000 this year. In contrast, restrictions that prohibit plans from setting different premiums based on health status—including those in the federal rules that have governed employer health plans since 1974—take the longer view. In this view, insurance seeks to cover the risk of becoming reclassified as an expensive patient in some future period. In an unregulated market, such reclassification would mean facing significantly higher health insurance premiums.[2]

Empirically, this reclassification risk seems important. More than half of US households contain a member with a pre-existing condition (Kaiser Family Foundation 2016). Calibrations by Handel, Hendel, and Whinston (2015) suggest that the welfare benefits of eliminating reclassification risk may swamp the welfare costs of one-period adverse selection. While we feel obligated to bring attention to this understudied and important issue, we will focus here primarily on the *interaction* between policies like community rating that address this reclassification risk, and selection.[3]

Subsidies and Penalties Related to Taking Up Health Insurance

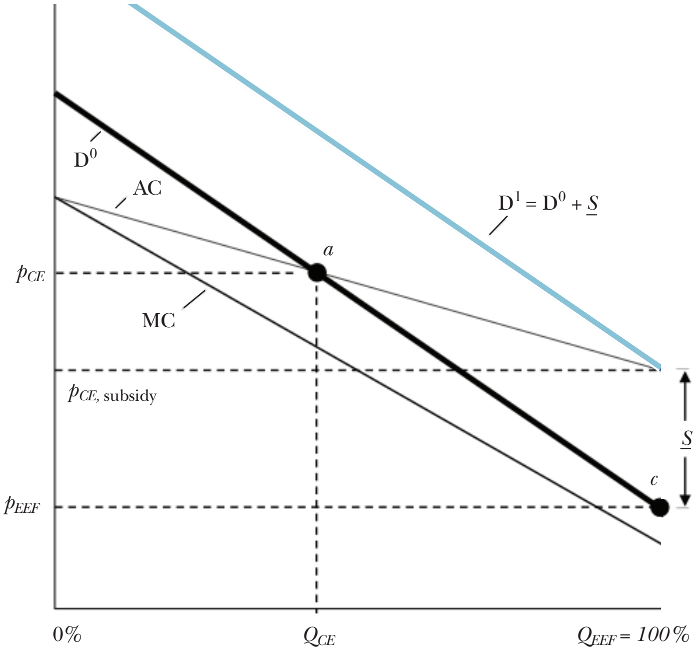
On the extensive margin between purchasing any insurance or none, policies that involve premium subsidies or penalties for not purchasing insurance (also known as coverage "mandates") are often used to combat selection problems, including the information asymmetries introduced by community rating. The rationale from economic theory for subsidies/penalties related to taking up insurance is most clear from the perspective of the fixed contracts framework, in which adverse selection into the risk pool drives prices to become inefficiently high at the market level.

Consider Figure 2, where we follow the basic setup of Einav and Finkelstein (2011), and examine the margin of consumers choosing between taking up insurance and remaining uninsured. The horizontal axis is scaled from 0 to 100 percent enrollment of the population, so that the demand curve D^0 reflects the willingness-to-pay for insurance of the marginal consumer at each level of enrollment. The vertical axis measures prices or costs in dollar terms. The marginal costs (MC) of enrollees slope downward, because adverse selection implies the highest willingness-to-pay consumers are those who generate the highest costs to insure. Demand and costs are more closely linked here than in the typical goods market, where the production technology determines a marginal cost that is independent of the particular consumer who purchases the good. Following the standard model, the competitive equilibrium Q_{CE} is determined by point a, the intersection of average costs (AC) and demand, where insurers earn zero profits. The efficient outcome is at point c, full enrollment, because in this example the demand curve is everywhere

[2] There exist other solutions to the reclassification risk problem, such as long-run insurance contracts, though these solutions face significant barriers to implementation, especially in the presence of the significant choice frictions and behavioral biases described in the other articles in this symposium. See Handel, Hendel, and Whinston (2017) for a detailed treatment of these alternative solutions.

[3] There is a subtle but important additional efficiency cost of premium rating restrictions. If consumers have heterogeneous preferences over insurance plans, a point on the marginal cost curve represents the average cost over a set of heterogeneous consumers who place the same value on insurance, and no uniform price can efficiently sort all consumers (Glazer and McGuire 2011; Bundorf, Levin, and Mahoney 2012; Geruso forthcoming). This feature is unique to selection markets, where the specific consumer who purchases the product determines both its value and its production cost.

Figure 2
Subsidies/Penalties and the Fixed Contracts Price Distortion



Notes: We follow the basic setup of Einav and Finkelstein (2011), and examine the margin of consumers choosing between taking up insurance and remaining uninsured. The horizontal axis is scaled from 0 to 100 percent enrollment. The vertical axis measures prices or costs in dollar terms. The demand curve D^0 reflects the willingness-to-pay for insurance of the marginal consumer at each level of enrollment. The marginal costs of enrollees slope downward, because adverse selection implies the highest willingness-to-pay consumers are those who generate the highest costs to insure. Following the standard model, the competitive equilibrium Q_{CE} is determined by point *a*, the intersection of average costs and demand, where insurers earn zero profits. The efficient outcome is at point *c*, full enrollment, because in this example the demand curve is everywhere above the marginal cost curve. A uniform subsidy, $\underline{S}$, equal to the difference between the rightmost point of the average cost curve and the rightmost point of the demand curve is the minimum uniform subsidy that will induce efficient sorting in this setting. If instead of a subsidy, a penalty were applied to the outside option of remaining uninsured, then $\underline{S}$ would define the minimum uniform penalty.

above the marginal cost curve, implying that willingness-to-pay exceeds individual-specific marginal costs to the plan for every individual.[4] Here, society as a whole (consumers + insurers) would be made better off if all consumers took up insurance.

[4] This diagram is appropriate for considering extensive margin selection from uninsurance to insurance, or for considering the Medicare Advantage/Traditional Medicare choice margin, where selection alters only the price of Medicare Advantage. In markets where the price of both options is endogenous to their risk pools, the equilibrium is more complex (Weyl and Veiga 2016; Layton 2016; Handel, Hendel, and Whinston 2015).

A uniform subsidy equal to the difference between the rightmost point of the average cost curve and the rightmost point of the demand curve is the minimum uniform subsidy that will induce efficient sorting in this setting. Call this minimum subsidy $\underline{S}$. Offering $\underline{S}$ can be viewed as shifting up the effective demand to intersect the average cost curve at exactly Q = 100 percent. Equivalently, offering $\underline{S}$ can be viewed as lowering the effective price perceived by consumers to the efficient price, p_{EFF}. If instead of a subsidy, a penalty were applied to the outside option of remaining uninsured, then $\underline{S}$ would define the minimum uniform penalty.

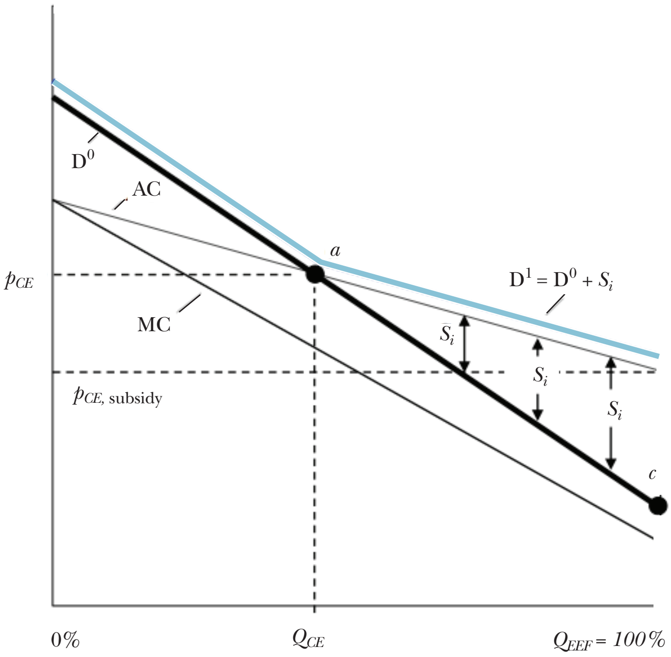
Importantly, $\underline{S}$ here is lower than the difference between p_{CE} (the competitive equilibrium price) and p_{EFF} (the efficient price). In other words, it appears to be the case that we have gotten something for nothing in reducing net prices by more than the subsidy amount. This happens because as more enrollees with lower marginal costs enter the market, they drive down average costs and therefore subsidize the competitive equilibrium price, which is equal to average cost. Thus, in adversely selected markets and under the assumptions given here, subsidies have a greater than one-for-one return in terms of lowering prices.

To determine $\underline{S}$ precisely, the regulator must be able to identify the full demand and cost curves, including any nonlinearities. However, any subsidy greater than the minimum $\underline{S}$ will also efficiently allocate consumers across plans, which allows the regulator significant room for error in setting the subsidy when the goal is full insurance. Of course, it may not be the case that the demand curve is everywhere above the marginal cost curve, implying that there are people whose valuation of insurance is less than the social cost of providing it to them. This could be due to moral hazard or nonnegligible loading costs (for example, costs incurred in marketing and claims administration) combined with low risk aversion. In this case, a subsidy that is too large causes welfare losses by inducing enrollment among consumers who value insurance below its marginal cost.

Yet another practical consideration is the deadweight loss from the taxes funding the subsidy. Even when the social optimum is full insurance, the cost of public funds needs to be taken into account when evaluating any publicly funded subsidy scheme. In light of this, penalties for not purchasing insurance may be preferable to subsidies. Penalties can induce allocative efficiency without requiring government expenditures, other than on enforcement. Incidence also differs: Subsidies fall on all enrollees, whereas penalties are more likely to bite for the population on the margin of making the choice to remain uninsured. However, penalties are politically unpopular, difficult to enforce, and may conflict with additional (and sometimes more prominent) distributional goals related to the notion of affordability. In particular, policymakers may be hesitant to force large penalties on low-income consumers.

Given the difficulties of implementing penalties, one might then ask whether there are ways to improve upon the minimum uniform subsidy, $\underline{S}$. A policy of subsidies targeted to the consumers with lowest willingness-to-pay for insurance may be more efficient than uniform subsidies. In Figure 3, we consider a candidate policy of paying person-specific subsidies S_i for the set of consumers to the right

Figure 3
Variable Subsidies Linked to Willingness-to-Pay



Note: Here we consider a policy of paying person-specific subsidies S_i for the set of consumers to the right of point *a*. For these consumers, the subsidy would be pivotal in their take-up decision. This subsidy schedule would generate the effective demand curve D^1, which adds the variable subsidy to the original demand curve, D^0. This tailored subsidy scheme would achieve the same universal coverage as the uniform subsidy $\underline{S}$ from Figure 2, but cost less.

of point *a*. For these consumers, the subsidy would be pivotal in their take-up decision. This subsidy schedule would generate the effective demand curve D^1, which adds the variable subsidy to the original demand curve, D^0. This subsidy scheme would achieve the same universal coverage as $\underline{S}$, but cost less. Costs would be lower both because fewer consumers would receive the tailored subsidy and because the tailored amounts would be smaller than $\underline{S}$ for all but the lowest willingness-to-pay consumer.

How could such variable subsidies be targeted in practice, for example, in the state-level health insurance Marketplaces? Work in the context of the Massachusetts Exchange that was enacted before the Patient Protection and Affordable Care Act of 2010 has shown that younger consumers are about twice as price sensitive as older consumers (Ericson and Starc 2015). Therefore, targeting subsidies to younger consumers is likely to achieve similar levels of allocative efficiency at a lower cost to the taxpayer than a uniform subsidy (Tebaldi 2017). In contrast, current policy proposals tend to favor subsidizing the *highest* cost enrollees such as via high-risk pool payments, or tying subsidies to age in the opposite pattern—offering larger subsidies to older, more expensive consumers. These are likely to be inefficient ways to address selection problems.[5]

Although we have focused so far on the competitive markets case, an additional complication inherent in designing subsidy schemes is the presence of imperfect competition. The portion of the subsidy that is passed through to consumers rather than extracted by producers (including insurers and health care providers) depends on the level of competition in the market. In the private Medicare Advantage context, on average about half of the dollar value of marginal changes in direct-to-plan subsidies are passed through to consumers in the form of lower premiums or lower cost sharing in a typical market (Curto, Einav, Levin, and Bhattacharya 2014; Song, Landrum, and Chernow 2013), with the largest pass-through rates in the most competitive local markets (Cabral, Geruso, and Mahoney 2014). These results suggest that market structure can have important effects on the consequences of a preset premium subsidy.

In practice, subsidies may also be dynamically linked to local market conditions, including to the prices that insurers set for their plans. This type of subsidy is used in the state-level Marketplaces, where tax credits are benchmarked to the price of the second-lowest price "Silver" plan. Jaffe and Shepard (2017) show that this type of price-linked subsidy distorts insurer prices because insurers that have some probability of having the second-lowest price plan will distort their prices upward to increase the size of the subsidy. On the other hand, this type of subsidy has the potential benefit that it protects subsidized consumers from changes in insurer prices (due to changes in technology, adverse selection, or other features) that are not anticipated by the regulator and thus cannot be incorporated into a fixed subsidy. Such a feature can be important in stabilizing a new market in which there is considerable uncertainty.

Clearly, implementing a mixture of mandates with penalties and subsidies involves a number of practical concerns. But despite these complications, the evidence to date indicates significant welfare gains from their use. For example, Hackmann, Kolstad, and Kowalski (2015) study the implementation of an individual mandate to purchase health insurance in Massachusetts that took the form of tax penalties paid by consumers who chose not to purchase coverage. They study the welfare consequences of the mandate assuming a fixed contracts model like Figure 2 and find an average welfare gain of 4.1 percent per person or a total of $51.1 million annually due to the penalty.

There is still a great deal we do not know about the use of mandates with penalties and subsidies as policy tools. First, there is work to be done to understand optimal subsidy schedules when subsidies can vary across consumers and when competition in the market is imperfect. Second, while economic theory suggests that tax penalties

[5] It is important to note, however, that these implications for efficiency are based on the static, one-period setting, and the efficiency consequences of reinsurance or differentially large subsidies for the healthy may be reversed, or at least weakened, when considering long-run dynamic risk such as the risk of acquiring a chronic disease.

and subsidies for insurance are largely equivalent to the consumer (differing only in their income effect), there is no evidence of which we are aware that suggests consumers react symmetrically to subsidies and penalties in this context. Further, the effect of the combination of subsidies and penalties embedded in the state-level Marketplaces through the Affordable Care Act had heterogeneous effects across states (Kowalski 2014). It is unclear what is driving that heterogeneity. While cross-state differences before the Affordable Care Act in the regulatory environment and in rates of uninsurance are obvious candidates, it is possible that factors like active marketing by states to encourage enrollment or efforts by states to improve the consumer's shopping experience played a role. Such effects may be important, but are inherently difficult to quantify.

Risk Adjustment

While selection along the extensive margin of insurance versus uninsurance is generally addressed via mandates backed by subsidies and penalties, the selection problems that arise on the intensive margin—that is, across plans within a market—are generally addressed by risk adjustment. We argue in this section that understanding why risk adjustment is so widely used requires a focus on this intensive margin and is further helped by examining insurance markets through the lens of endogenous, rather than fixed, contracts. We begin by outlining the mechanics of such a policy.

Mechanics of Risk Adjustment

Although the practical administration of risk adjustment is complicated in ways we will discuss below, the idea is simple: compensate plans for the *expected* costs of their enrollees and thereby remove the incentive to avoid high-expected-cost consumers, such as the cancer patients from our introduction. Thus, when an insurance company considers providing health insurance for a person, expected profits will be the premium received from this person minus the expected costs of providing coverage, plus a risk-adjustment payment.

To illustrate how risk adjustment works, consider Medicare Advantage, the private insurance option for hospital and physician coverage within the Medicare program. Estimation of a risk-adjustment transfer begins with calibrating the relationship between observables and costs in some reference population. In Medicare as in many settings, this is carried out via a simple ordinary least squares regression of annual patient costs on indicators for demographic variables and a small set of chronic disease indicators, derived from diagnosis codes in prior-year insurance claims. In Medicare, the right-hand-side variables also include indicators for Medicaid and disability status. In other settings, prescription drug utilization and other measures of prior health care use may be included. The estimated coefficients are then used to predict expected costs based on individual characteristics. These predictions from the risk adjustment regression are transformed into risk scores that are straightforward to interpret. A Medicare Advantage enrollee with a risk score of 1.5 would have expected costs equal to 150 percent of the costs of the typical enrollee in Traditional Medicare.

Finally, to determine the actual dollar size of the risk adjustment, risk scores are multiplied by some dollar amount, which for Medicare Advantage is roughly the cost of enrolling the typical-health Medicare beneficiary in Traditional Medicare in the local geographic market. In other words, this payment approximates what a person with these observable characteristics would have cost the taxpayer if the person had enrolled in Traditional Medicare. This risk-adjustment mechanism is simultaneously providing a premium subsidy and a selection correction.

In the setting of the Marketplaces established by the Affordable Care Act, there is no equivalent of the Traditional Medicare program with which to benchmark risk scores. The Marketplace scheme uses instead the reference population of large employer plans. Risk scores are normalized against the average risk score in that market, and transfers are sent from plans with low-cost enrollees to plans with high-cost enrollees.

Theoretical Underpinnings of Risk Adjustment

Figure 4 depicts the same baseline demand and risk selection conditions for the insurance/uninsurance setting described in Figure 2. Here we use it to provide intuition for how risk adjustment would only imperfectly address the price distortions that are described by the fixed contracts framework.

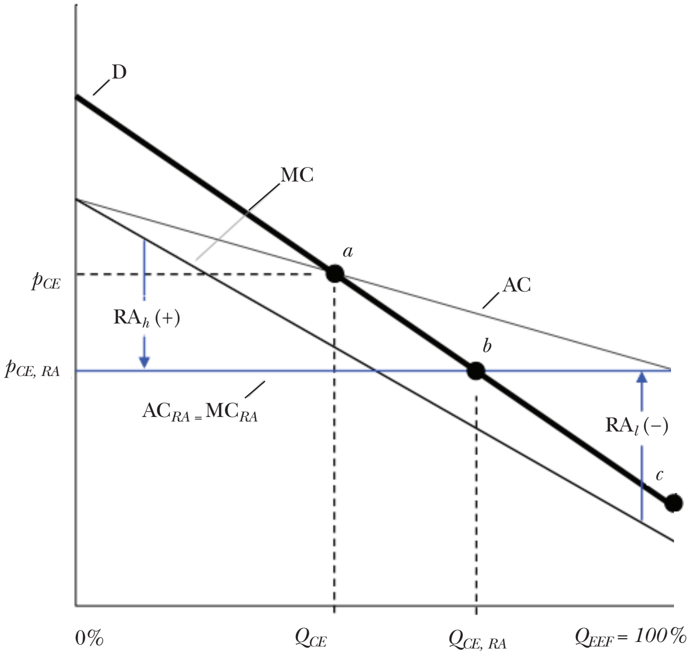
Risk adjustment alters the competitive equilibrium by compensating for the individual-specific difference between marginal costs and the average population cost. This has the effect of rotating the insurer's *perceived* marginal cost curve. Larger positive risk adjustment payments, like RA_h, are made by the regulators for individuals with larger expected costs, and smaller or negative payments like RA_l, for enrollees with lower costs. In the diagram, the solid horizontal line represents the net marginal cost perceived by the insurer in the case in which risk adjustment perfectly compensates for expected costs. This net marginal cost curve is flat at the level of the population average cost, arresting the feedback loop that would otherwise link equilibrium prices to the composition of the enrolled risk pool.

Although the insurer determines pricing according to perceived costs, the social marginal cost curve relevant for welfare analysis remains the original, downward sloping line, which implies that full enrollment remains the efficient outcome. Given the (arbitrary) demand and cost curves drawn in the diagram, the competitive equilibrium under risk adjustment is determined by point *b*. Enrollment with risk adjustment, $Q_{CE,RA}$, is higher than the unregulated case, and closer to the optimum.

However, risk adjustment does not completely resolve the inefficiency by raising the enrollment rate to 100 percent, at least not without an additional subsidy.[6] While risk adjustment subsidizes insurers for enrolling sicker consumers (the positive RA_h), it also taxes them for enrolling healthier consumers (the negative RA_l). This tax on

[6] In the online Appendix available with this paper at http://e-jep.org, we offer an example in which risk adjustment could make the allocation worse in a competitive equilibrium.

Figure 4
Risk Adjustment and the Fixed Contracts Price Distortion



Notes: Larger positive risk adjustment payments, like RA_h, are made by the regulators for individuals with larger expected costs, and smaller or negative payments like RA_l, for enrollees with lower costs. Given the (arbitrary) demand and cost curves drawn in the diagram, the competitive equilibrium under risk adjustment is determined by point *b*. Enrollment with risk adjustment, $Q_{CE,\,RA}$, is higher than the unregulated case, and closer to the optimum. However, risk adjustment does not completely resolve the inefficiency by raising the enrollment rate to 100 percent, at least not without an additional subsidy.

enrolling healthy consumers limits the extent to which risk adjustment can solve the inefficient sorting problem in settings where it is efficient for all consumers to purchase insurance. Mahoney and Weyl (2017) apply the fixed contract framework to show that in both perfectly and imperfectly competitive markets, risk adjustment may improve or worsen the allocation, depending on demand and selection. In our diagram, a subsidy of $\underline{S}$ would need to be employed in addition to risk adjustment to generate efficient sorting. From this perspective, risk adjustment appears to have done little: the same minimum subsidy of $\underline{S}$ from Figure 2 would be needed to achieve the optimum (Q_{EFF} = 100 percent) *with or without risk adjustment.*[7]

[7] Risk adjustment does, nonetheless, break the connection between the enrollee risk pool and the plan's average costs. In this way it stabilizes the market, easing insurer uncertainty about net costs, and reducing the probability of prices evolving uncertainly in a setting like the Marketplaces established by the Affordable Care Act in which the demand and cost curves (determining the competitive equilibrium) were not common knowledge.

If risk adjustment does not solve the inefficiency in this setting, then why is this policy instrument so widely used? For one, budget-neutral risk adjustment *may* improve allocative efficiency to some extent without requiring the regulator to provide non-budget-neutral subsidies. Additionally, it is important to understand that risk adjustment is intended to address *intensive* margin (high- versus low-coverage) selection rather than the *extensive* margin (insurance versus uninsurance) selection problem depicted in Figure 2. Adapting the fixed contracts approach to the intensive margin problem, Layton (forthcoming) and Handel, Hendel, and Whinston (2015) show that conventional risk adjustment can eliminate most of the inefficiency caused by adverse selection across plans in a Marketplace-like setting, assuming consumers cannot opt out of coverage altogether. Handel, Kolstad, and Spinnewijn (2015) use a fixed contracts framework to show that risk adjustment can be complementary to policies that improve consumer choices, limiting the negative consequences of these choice-improving policies for adverse selection (Handel 2013).

But to understand fully the motivation for risk adjustment, one must consider not only intensive-margin selection across differentiated fixed contracts but also the endogenous design of those contracts. The most important objectives of risk adjustment are related to the design of health plan benefits, rather than prices. The Center for Medicare and Medicaid Services, for example, thinks about risk adjustment as a way to counter "cream-skimming" behavior by insurers. The regulatory focus on cream-skimming suggests that regulators and policymakers are worried about the endogenous contracts distortions discussed earlier, rather than the price-feedback mechanism described in Figure 4.

In principle, risk adjustment can address insurer incentives to try to avoid certain patient types because risk adjustment can make all enrollees equally profitable to the insurer on net (Van de Ven and Ellis 2000; Breyer, Bundorf, and Pauly 2011). Intuitively, risk adjustment forces the healthy to subsidize the sick to some extent, no matter what contract they purchase. This limits the possibilities of an inefficient separating equilibrium with higher- and lower-coverage plans and can lead to an efficient pooling equilibrium where all consumers, both healthy and sick, fully insure (Glazer and McGuire 2000).

Risk Adjustment in Practice

In practice, it can be difficult to evaluate whether risk adjustment is functioning well, because risk adjustment is usually introduced to a market alongside other important regulatory changes. But between 2004 and 2007, Medicare Advantage transitioned to a risk adjustment system based on diagnoses for chronic conditions, while holding fixed other important features like community rating. After the implementation of diagnosis-based risk adjustment in 2004, Medicare Advantage plans enrolled beneficiaries who were sicker than their pre-risk adjustment enrollees (Brown, Duggan, Kuziemko, and Woolston 2014; Newhouse and McGuire 2014). This is consistent with risk adjustment successfully removing some of the financial incentive to avoid sicker, costlier patients.

However, the enrollment of additional sicker patients is not a sufficient statistic for judging the success of risk adjustment at combatting contract distortions due to adverse selection. If insurers respond to risk adjustment by switching away from designing contracts to attract low-cost individuals and instead move to designing contracts to attract individuals who are low cost *conditional on their risk scores*, a new class of distortions can arise. Brown, Duggan, Kuziemko, and Woolston (2014) and Newhouse, Price, McWilliams, Hsu, and McGuire (2015) provide evidence that while the set of Medicare beneficiaries switching from Traditional Medicare to Medicare Advantage got sicker after 2004, the costs of these switchers *conditional on their risk scores* actually went down. This result is consistent with insurers cream-skimming by switching their plan design and marketing strategies away from targeting low-cost enrollees to targeting beneficiaries who are low-cost conditional on their risk scores (Aizawa and Kim 2015). But it is also consistent with insurers being willing to attract sicker consumers after the introduction of risk adjustment, and with the lower-cost consumers among the sick simply being more likely to take up Medicare Advantage compared to the higher-cost sick.[8]

A more direct piece of evidence regarding cream-skimming conditional on risk scores comes from Lavetti and Simon (2016). They examine Medicare contracts for pharmaceutical coverage in the post-risk adjustment period. They find that Medicare Advantage drug formularies differ from stand-alone Medicare Part D plan formularies in ways that are consistent with screening-in enrollees who were profitable conditional on risk adjustment. This finding suggests that even if risk adjustment has improved the equilibrium set of contracts in Medicare Advantage, some degree of distortion remains.

In the state-level Marketplaces established by the Affordable Care Act, before-and-after comparisons are less clear. The introduction of risk adjustment in these programs was combined with major contemporaneous policy changes, and sorting out the effects is difficult. But there is at least some prima facie evidence that the Marketplace plans are being designed to attract enrollees who would likely have been highly unprofitable without risk adjustment. For example, Aetna launched Marketplace plans for 2016 that were specifically marketed toward diabetics, with features like differentially low cost-sharing for specialist visits linked to diabetes management (Andrews 2015).

In summary, without risk adjustment, the incentive for an insurer to distort coverage for a particular dimension of the contract is determined only by the cost of the consumers who value that dimension of the contract. With risk adjustment, there is variation in both cost and revenue across consumers. Thus, risk adjustment may change the margin of selection, rather than eliminate it entirely.

[8] The result is complicated by the observation that this pattern appears to have reversed course in 2006, with switcher costs conditional on risk scores returning to their 2001 levels (Newhouse et al. 2015). We note that the introduction of Part D and increases to Medicare Advantage benchmarks that occurred in 2006 represent potential confounders for this time period due to their potential independent effects on the composition of the Medicare Advantage risk pool.

To make these ideas concrete, in Figure 5 we compare consumer costs and risk-adjusted revenues based on the Marketplace risk adjustment scheme. The figure is based on detailed health claims data for about 12 million consumers who are enrolled in plans offered by their large employers.[9] Note that these are not Marketplace claims data. But they are instructive regarding the incentives embedded in the Marketplace payment formulas. The claims data allow direct observation of costs. The claims data also include all of the diagnosis information necessary to calculate risk adjustment payments implied by Marketplace formulas. We use the risk adjustment software from the regulator to generate hypothetical risk adjustment transfers associated with each enrollee, as if the enrollee's claim history had been generated while enrolled in a Marketplace plan.

We focus in Figure 5 on the possibility of cream-skimming via the design of prescription drug benefits. We classify individuals according to whether they have a pharmacy claim for a drug within one of 220 standard therapeutic classes of medications. Each circle in the figure corresponds to a therapeutic class, grouping together all consumers who used a drug in the class. Marker sizes are proportional to the numbers of consumers associated with each class. The horizontal axis measures mean total spending among consumers utilizing a drug in the class, and the vertical axis measures the mean simulated revenue (actuarially fair premiums plus risk adjustment transfers) among those same consumers. Consumers associated with classes below the 45-degree line are profitable to avoid because, for these consumers, insurer costs exceed Marketplace premium plus risk adjustment revenue in expectation.

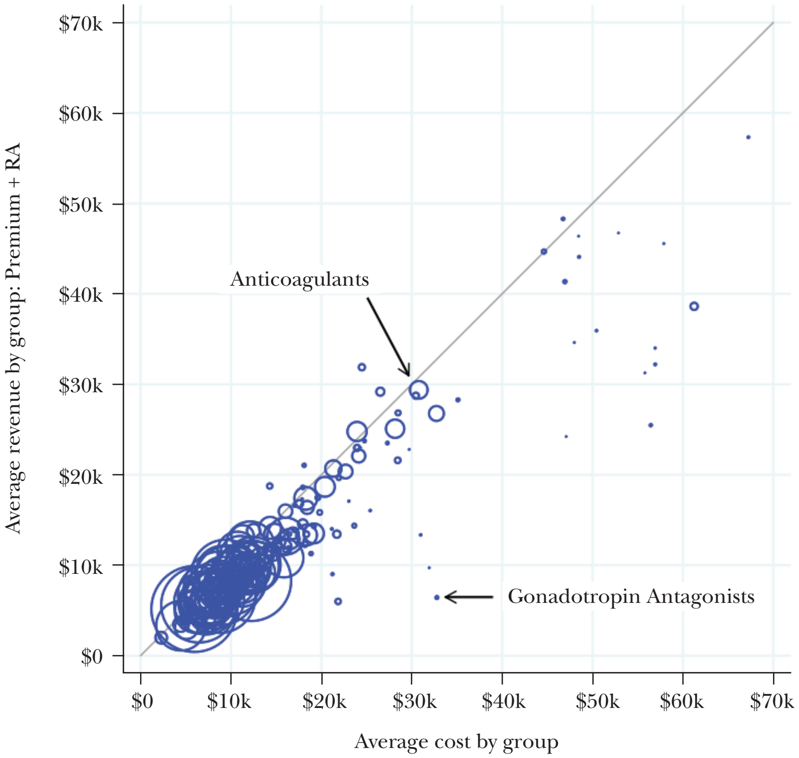
In Figure 5, the majority of drug classes are clustered tightly around the 45-degree line. This pattern implies that the payment system neutralizes the screening incentives for the majority of potential enrollees. For many drug classes that would predict costs several times in excess of premiums, such as anticoagulants (blood thinners), costs do not correlate with unprofitability *net of the risk adjustment payment.* This suggests that the Marketplace risk adjustment is succeeding in protecting consumers whose prescription drug use would otherwise flag them as unprofitable to insure.

However, there are a small number of significant outliers, such as the gonadotropin antagonist class (for infertility in women) far off the diagonal. Geruso, Layton, and Prinz (2016) analyze the universe of state-level Marketplace formularies for 2015 and show that insurers indeed design formularies to be differentially unattractive to the groups that deviate far below the 45-degree line. Within a plan, drug classes used by less-profitable consumers appear higher on the formulary tier structure, implying higher out-of-pocket costs by potentially thousands of dollars per year and/or significant nonprice hurdles, including prior authorization. Even less-expensive and generic drugs that are associated with expensive patients are assigned to high cost-sharing tiers or are left off formularies altogether. Other prior studies have provided similar evidence of insurers responding to imperfect risk adjustment

[9] These large employer claims data are aggregated by Truven Health and cover plan years 2012 and 2013. See the online Data Appendix for full details.

Figure 5

Incentives to Screen May Remain Net of Risk Adjustment



Note: We classify individuals according to whether they have a pharmacy claim for a drug within one of 220 standard therapeutic classes of medications. Each circle in the figure corresponds to a therapeutic class, grouping together all consumers who used a drug in the class. Marker sizes are proportional to the numbers of consumers associated with each class. The horizontal axis measures mean total spending among consumers utilizing a drug in the class, and the vertical axis measures the mean simulated revenue (actuarially fair premiums plus risk adjustment transfers) among those same consumers. Consumers associated with classes below the 45-degree line are profitable to avoid because, for these consumers, insurer costs exceed Marketplace premium plus risk adjustment revenue in expectation. The majority of drug classes are clustered tightly around the 45-degree line, showing that the payment system succeeds in neutralizing selection incentives for the majority of potential enrollees. However there are a number of significant outliers, such as the gonadotropin class of drugs (for infertility in women).

via formulary design in Medicare Part D (Carey 2017) and via hospital network design in the Massachusetts "Connector" marketplace (Shepard 2016), which was set up before the Patient Protection and Affordable Care Act of 2010.

Aside from the tendency of insurers to react to the exploitable errors in any risk adjustment system, risk adjustment faces several challenges due to the need to construct the risk-score based on observable signals of expected costs. For example, risk-adjusted payments to Medicare Advantage plans are ultimately based on diagnoses recorded on health insurance claims. In Traditional Medicare, on the other hand, diagnoses play no role in many payments, such as payments for outpatient physician services. This means physicians face relatively weak incentives to document diagnoses in Traditional Medicare claims, regardless of whether such diagnoses are recorded in the physician's notes and patient's medical records. Therefore, it is perhaps not surprising that if an individual enrolls in Medicare Advantage, the doctors with whom Medicare Advantage plans contract typically record more, and more severe, diagnoses. This leads to patient risk scores that are on average 6–7 percent higher than the score the same patient would generate in Traditional Medicare (Geruso and Layton 2015). The Centers for Medicare and Medicaid Services acknowledges the coding differences, and over time has implemented increasingly large (but likely still too small) deflation factors to risk scores reported by Medicare Advantage plans.

The fact that diagnosis codes (or risk-adjustment variables more generally) are not fixed characteristics of consumers also leads to an efficiency problem in terms of how intensely health care services are provided. In principle, risk adjustment aims at reimbursing plans for who they enroll, rather than what the plans do. This would align the insurer, who is the residual claimant on capitation funds not paid out to providers, with the policymaker's goal of constraining the growth of health care spending. However, Geruso and McGuire (2016) show that risk adjustment in the state-level Marketplaces significantly reimburses plans on the margin for actual care given. Intuitively, this occurs because the recorded diagnoses only arise endogenously via an interaction with a service provider, so risk scores are implicitly tied to utilization, rather than fixed characteristics of consumers. Across major diagnostic categories of services, insurers are reimbursed for services provided between 8 cents on the dollar and 82 cents on the dollar by the Marketplace risk adjustment scheme.

In markets without a public option, such as the state-level Marketplaces, an additional challenge arises: It is not clear what to use as the "baseline" plan when calibrating the relationship between costs and diagnoses. Einav, Finkelstein, Kluender, and Schrimpf (2016) offer evidence that conventional risk-adjustment policies *cannot* perfectly adjust for expected costs in plans with different coverage, implying that at least some cream-skimming incentives will always remain. The Marketplaces include plans with dramatically different cost structures, from low-cost Medicaid-like plans to generous wide-network "Cadillac" plans. However, the current risk adjustment system used in the Marketplaces treats all health insurance plans equally, with all plan risk adjustment transfers based on the average premium in the market, and with only minor modifications for a plan's actuarial value. Layton, Montz, and Shepard (2017) show analytically that this equal treatment has potentially distortionary consequences, with the choice of the benchmark plan determining the extent of the transfer from low-cost plans to high-cost plans. How to deal with this issue remains a key area for future research.

A final complication relates to consumers' outside option. Because risk adjustment forces low-premium advantageously selected plans to transfer money to high-premium adversely selected plans, it likely results in raising the premiums of the lowest-price plans. This results in more people enrolling in the higher-cost, more comprehensive plans, but it may also force marginal enrollees out of the market (Newhouse forthcoming), implying that risk adjustment may need to be accompanied by significant premium subsidies and/or penalties on the insurance/uninsurance margin if it is to be successful in these settings.

The substantial challenges implicit in designing the optimal risk adjustment system suggest important avenues for future theoretical and empirical work. Despite these challenges, conventional risk adjustment is the best tool we have to address selection across plans in competitive health insurance markets, hence its near-universal adoption in individual health insurance markets.

Contract Regulation

Almost all insurance markets feature extensive regulations on the contracts that insurers may offer. In Medicare Advantage, private plans must offer at least the standard set of benefits provided under Traditional Medicare. In the state-level health insurance Marketplaces, plans are required to pay for at least 60 percent of the health care costs of an average patient, to meet network adequacy mandates, and to comply with Essential Health Benefits (EHB) rules, which lay out minimal coverage requirements for services including maternity and newborn care, mental health and substance use disorder services, prescription drugs, and more. Services in these categories must be covered at least as well as they are covered in a "benchmark" plan chosen in each state. The variations in state benchmarks for Essential Health Benefits are in fact reflected in contract design differences across states (Andersen forthcoming).

These types of benefit regulations can be understood as a last line of defense against the endogenous contract distortions. As discussed above, if adverse selection is not adequately counteracted by risk adjustment, then the equilibrium set of contracts could be quite different from the efficient set of contracts, and in such a situation, restraining the equilibrium set of contracts could potentially improve welfare. The potential gains from such provisions can only be understood in an endogenous contracts framework.

However, these types of benefit regulations may also produce unintended consequences. For example, while Andersen (forthcoming) finds that the Essential Health Benefits regulations result in more drugs being covered in the formularies of Marketplace plans, the additional covered drugs are much more likely to be subject to utilization management restrictions, which have the effect of limiting access in practice (Simon, Tennyson, and Hudman 2009). This finding illustrates a key problem with using contract regulations to combat selection problems: it is very difficult for regulators to design rules that limit all possible dimensions of the health care interaction.

Another major tradeoff when using this type of regulatory mechanism is that minimum coverage requirements can lead some consumers who would like to purchase less-generous coverage to go uninsured (Finkelstein 2004). Even if uninsurance can be removed from the choice set with some combination of penalties and mandates, minimum coverage can in principle induce a death spiral for other plans in the market: specifically, as more healthy consumers are required to purchase a medium coverage contract, the price of that medium coverage contract drops, inducing some (relatively healthy) consumers who would have chosen a high coverage contract to inefficiently move to the less-generous minimum coverage (Azevedo and Gottlieb 2017).

A final potential downside to contract regulations is that even if all dimensions of the plan are observable and enforceable, it is difficult for a regulator to know the efficient level of coverage for each particular service. Determining optimal coverage involves a complex optimization problem that incorporates many difficult-to-estimate parameters such as consumer elasticities of demand and insurer and provider market power. Regulations could require insurers to provide too much of some benefits from the standpoint of social welfare. Additionally, the presence of this type of regulation can lead to political economy problems where interest groups lobby the government to require coverage of the services they use or provide, leading to a set of regulations that reflect political influence rather than social efficiency.

Overall, while contract restrictions may play a role in plugging various holes left by imperfectly implemented risk adjustment, such policies have clear limits. Our summary reading of the evidence is that when attempting to limit selection problems in markets, there is no good substitute for a payment system that leverages market forces and addresses insurers' financial incentives with respect to selection, rather than tries to force insurers to act against their own financial interests.

Conclusion

Publicly financed health insurance programs in the United States have in recent years come to rely more heavily on private insurance markets where individuals choose from a variety of plans designed by private sector insurers. This change is especially apparent in the growth of the Medicare Advantage program and the creation of the state-level health insurance Marketplaces by the Patient Protection and Affordable Care Act of 2010. The health insurance contracts actually offered to individuals by private insurers clearly reflect the reality that selection incentives matter. Although the consequences of adverse selection can be limited by risk adjustment, premium rating regulations, mandates/subsidies, and contract regulations, there is still a great deal that we don't know about what optimal plan payment policies look like. Glazer and McGuire (2000, 2002) took early steps towards developing a theory of optimal risk adjustment, but both the markets in which these policies are used and the technology of risk adjustment itself are much more complex than was originally anticipated. For example, we now know that plans with heterogeneous cost structures imperfectly compete alongside each other in the same market. Additionally, risk scores appear to be highly endogenous to the plan a consumer chooses and the contract an insurer designs. Thus, even with policies to limit selection in place, these issues are ongoing. It seems to be an inescapable fact, at least at the current state of knowledge, that risk adjustment and other plan payment policies are unable to capture all relevant dimensions of consumers' expected health care spending.

These complications imply that new theories of optimal (second-best) payment policies need to be developed, along with complementary regulations. Some of this research will focus on alternative methods of calculating risk adjustment payments, along with new structures for subsidies or mandates. But it is also important to expand the range of optimal payment policies to be considered. For example, one approach might consider reinsurance programs that compensate plans based on certain key dimensions of after-the-fact realized costs, but it will be important to focus on dimensions that are least susceptible to moral hazard concerns (Geruso and McGuire 2016; Layton, McGuire, and van Kleef 2016). Another policy alternative might seek to compensate health insurance plans based on certain features of the contracts themselves, rather than the imperfect selection signals generated by risk scores. The long-term success of policies that rely on consumer choice in markets for subsidized but privately provided health insurance depends on research that improves our understanding of how to address the selection issues outlined here.

■ *We thank Keith Ericson, Jon Gruber, Tom McGuire, Joe Newhouse, Dean Spears, and Steve Trejo for reviewing an earlier draft of this essay. We thank Austin Bean for excellent research assistance. We gratefully acknowledge financial support from the National Institute of Mental Health (R01-MH094290), the National Institute on Aging (P01-AG032952), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (P2CHD042849, Population Research Center, awarded to the Population Research Center at the University of Texas at Austin), and the Laura and John Arnold Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or any other funder.*

References

Aizawa, Naoki, and You Suk Kim. 2015. "Advertising and Risk Selection in Health Insurance Markets." Board of Governors of the Federal Reserve System (US) Finance and Economics Discussion Series 2015–101.

Andersen, Martin. Forthcoming. "Constraints on Formulary Design under the Affordable Care Act." *Health Economics.*

Andrews, Michelle. 2015. "New Health Plans Offer Discounts for Diabetes Care." *Kaiser Health News*, November 17. http://khn.org/news/new-health-plans-offer-discounts-for-diabetes-care.

Azevedo, Eduardo, and Daniel Gottlieb. 2017. "Perfect Competition in Markets with Adverse Selection." *Econometrica* 85(1): 67–105.

Bauman, Noam, Jason Bello, Erica Coe, and Jessica Lamb. 2015. "Hospital Networks: Evolution of the Configurations on the 2015 Exchanges."

McKinsey & Company, April. http://healthcare.mckinsey.com/2015-hospital-networks.

Breyer, Friedrich, M. Kate Bundorf, and Mark V. Pauly. 2011. "Health Care Spending Risk, Health Insurance, and Payment to Health Plans." In *Handbook of Health Economics,* edited by Mark V. Pauly, Thomas G. Mcguire, and Pedro P. Barros, 691–762. Amsterdam: Elsevier.

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston. 2014. "How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program." *American Economic Review* 104(10): 3335–64.

Buchmueller, Thomas, and John DiNardo. 2002. "Did Community Rating Induce an Adverse Selection Death Spiral? Evidence from New York, Pennsylvania, and Connecticut." *American Economic Review* 92(1): 280–94.

Bundorf, M. Kate, Jonathan Levin, and Neale Mahoney. 2012. "Pricing and Welfare in Health Plan Choice." *American Economic Review* 102(7): 3214–48.

Cabral, Marika, Michael Geruso, and Neale Mahoney. 2014. "Does Privatized Health Insurance Benefit Patients or Producers? Evidence from Medicare Advantage." NBER Working Paper 20470.

Carey, Colleen. 2017. "Technological Change and Risk Adjustment: Benefit Design Incentives in Medicare Part D." *American Economic Journal: Economic Policy* 9(1): 38–73.

Cochrane, John H. 1995. "Time-Consistent Health Insurance." *Journal of Political Economy* 103(3): 445–73.

Curto, Vilsa, Liran Einav, Jonathan Levin, and Jay Bhattacharya. 2014. "Can Health Insurance Competition Work? Evidence from Medicare Advantage." NBER Working Paper 20818.

Duggan, Mark, and Fiona Scott Morton. 2010. "The Effect of Medicare Part D on Pharmaceutical Prices and Utilization." *American Economic Review* 100(1): 590–607.

Einav, Liran, and Amy Finkelstein. 2011. "Selection in Insurance Markets: Theory and Empirics in Pictures." *Journal of Economic Perspectives* 25(1): 115–38.

Einav, Liran, Amy Finkelstein, and Mark R. Cullen. 2010. "Estimating Welfare in Insurance Markets Using Variation in Prices." *Quarterly Journal of Economics* 125(3): 877–921.

Einav, Liran, Amy Finkelstein, Raymond Kluender, and Paul Schrimpf. 2016. "Beyond Statistics: The Economic Content of Risk Scores." *American Economic Journal: Applied Economics* 8(2): 195.

Ericson, Keith M. Marzilli, and Amanda Starc. 2015. "Pricing Regulation and Imperfect Competition on the Massachusetts Health Insurance Exchange." *Review of Economics and Statistics* 97(3): 667–82.

Finkelstein, Amy. 2004. "Minimum Standards, Insurance Regulation and Adverse Selection: Evidence from the Medigap Market." *Journal of Public Economics* 88(12): 2515–47.

Frank, Richard G., Jacob Glazer, and Thomas G. McGuire. 2000. "Measuring Adverse Selection in Managed Health Care." *Journal of Health Economics* 19(6): 829–54.

Geruso, Michael. Forthcoming. "Demand Heterogeneity in Insurance Markets: Implications for Equity and Efficiency." *Quantitative Economics.*

Geruso, Michael, and Timothy Layton. 2015. "Upcoding: Evidence from Medicare on Squishy Risk Adjustment." NBER Working Paper 21222.

Geruso, Michael, Timothy J. Layton, and Daniel Prinz. 2016. "Screening in Contract Design: Evidence from the ACA Health Insurance Exchanges." NBER Working Paper 22832.

Geruso, Michael, and Thomas G. McGuire. 2016. "Tradeoffs in the Design of Health Plan Payment Systems: Fit, Power and Balance." *Journal of Health Economics* 47: 1–19.

Glazer, Jacob, and Thomas G. McGuire. 2000. "Optimal Risk Adjustment in Markets with Adverse Selection: An Application to Managed Care." *American Economic Review* 90(4): 1055–71.

Glazer, Jacob, and Thomas G. McGuire. 2002. "Multiple Payers, Commonality and Free-Riding in Health Care: Medicare and Private Payers." *Journal of Health Economics* 21(6): 1049–69.

Glazer, Jacob, and Thomas G. McGuire. 2011. "Gold and Silver Health Plans: Accommodating Demand Heterogeneity in Managed Competition." *Journal of Health Economics* 30(5): 1011–19.

Hackmann, Martin B., Jonathan T. Kolstad, and Amanda E. Kowalski. 2015. "Adverse Selection and an Individual Mandate: When Theory Meets Practice." *American Economic Review* 105(3): 1030–66.

Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review* 103(7): 2643–82.

Handel, Ben, Igal Hendel, and Michael D. Whinston. 2015. "Equilibria in Health Exchanges: Adverse Selection versus Reclassification Risk." *Econometrica* 83(4): 1261–1313.

Handel, Benjamin, Igal Hendel, and Michael Whinston. 2017. "The Welfare Effects of Long-Term Health Insurance Contracts." Unpublished paper.

Handel, Benjamin R., Jonathan T. Kolstad, and Johannes Spinnewijn. 2015. "Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets." NBER Working Paper 21759.

Hendren, Nathaniel. 2013. "Private Information and Insurance Rejections." *Econometrica* 81(5): 1713–62.

Hendren, Nathaniel. 2017 "Measuring Ex-ante Welfare in Insurance Markets." NBER Working Paper 23742.

Ho, Kate, and Robin S. Lee. 2017. "Equilibrium Provider Networks: Bargaining and Exclusion in Health Care Markets." Unpublished paper.

Jacobs, Douglas B., and Benjamin D. Sommers. 2015. "Using Drugs to Discriminate? Adverse Selection in the Insurance Marketplace." *New England Journal of Medicine* 372(5): 399–402.

Jacobson, Gretchen, Ariel Trilling, Tricia Neumann, Anthony Damico, and Marsha Gold. 2016. *Medicare Advantage Hospital Networks: How Much Do They Vary?* Menlo Park, CA: Kaiser Family Foundation.

Jaffe, Sonia P., and Mark Shepard. 2017. "Price-Linked Subsidies and Health Insurance Markups." NBER Working Paper 23104.

Kaiser Family Foundation. 2016. *Kaiser Health Tracking Poll: August 2016*. Menlo Park, CA: Kaiser Family Foundation.

Kowalski, Amanda E. 2014. "The Early Impact of the Affordable Care Act, State by State." *Brookings Papers on Economic Activity,* Fall, pp. 277–333.

Lavetti, Kurt, and Kosali Simon. 2016. "Strategic Formulary Design in Medicare Part D Plans." NBER Working Paper 22338.

Layton, Timothy. Forthcoming. "Imperfect Risk Adjustment, Risk Preferences, and Sorting in Competitive Health Insurance Markets." *Journal of Health Economics.*

Layton, Timothy J., Thomas G. McGuire, and Richard C. van Kleef. 2016. "Deriving Risk Adjustment Payment Weights to Maximize Efficiency of Health Insurance Markets." NBER Working Paper 22642.

Layton, Timothy, Ellen Montz, and Mark Shepard. 2017. "Health Plan Payment in U.S. Marketplaces: Regulated Competition with a Weak Mandate." NBER Working Paper 23444.

Mahoney, Neale, and E. Glen Weyl. 2017. "Imperfect Competition in Selection Markets." *Review of Economics and Statistics.* Posted online January 18, ahead of print. doi: 10.1162/REST_a_00661.

Newhouse, Joseph. Forthcoming. "Risk Adjustment with an Outside Option." *Journal of Health Economics.*

Newhouse, Joseph, and Thomas McGuire. 2014. "How Successful Is Medicare Advantage?" *Milbank Quarterly* 92(2): 351–94.

Newhouse, Joseph P., Mary Price, J. Michael McWilliams, John Hsu, and Thomas G. McGuire. 2015. "How Much Favorable Selection Is Left in Medicare Advantage?" *American Journal of Health Economics* 1(1): 1–26.

Polsky, Dan, and Janet Weiner. 2015. *The Skinny on Narrow Networks in Health Insurance Marketplace Plans.* Philadelphia: Leonard Davis Institute of Health Economics.

Rabin, Roni Caryn, and Reed Abelson. 2013. "Health Plan Costs for New Yorkers Set to Fall 50%." *New York Times,* July 16. http://www.nytimes.com/2013/07/17/health/health-plan-cost-for-new-yorkers-set-to-fall-50.html?mcubz=0.

Rothschild, Michael, and Joseph E. Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics* 90(4): 630–49.

Shepard, Mark. 2016. "Hospital Network Competition and Adverse Selection: Evidence from the Massachusetts Health Insurance Exchange." NBER Working Paper 22600.

Simon, Kosali, Sharon Tennyson, and Julie Hudman. 2009. "Do State Cost Control Policies Reduce Medicaid Prescription Drug Spending?" *Risk Management and Insurance Review* 12(1): 39–66.

Song, Zirui, Mary Beth Landrum, and Michael E. Chernew. 2013. "Competitive Bidding in Medicare Advantage: Effect of Benchmark Changes on Plan Bids." *Journal of Health Economics* 32(6): 1301–12.

Tebaldi, Pietro. 2017 "Estimating Equilibrium in Health Insurance Exchanges: Price Competition and Subsidy Design under the ACA." August 14. Unpublished paper.

Van de Ven, Wynand, and Randall P. Ellis. 2000. "Risk Adjustment in Competitive Health Plan Markets." In *Handbook of Health Economics,* edited by A. J. Culyer and J. P. Newhouse, 755–845. Amsterdam: Elsevier.

Veiga, André, and E. Glen Weyl. 2016. "Product Design in Selection Markets." *Quarterly Journal of Economics* 131(2): 1007–56.

Weyl, E. Glen, and André Veiga. 2016. "Pricing Institutions and the Welfare Cost of Adverse Selection." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2344812.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 51–72

The Questionable Value of Having a Choice of Levels of Health Insurance Coverage

Keith Marzilli Ericson and Justin Sydnor

In most health insurance markets in the United States, consumers have substantial choice about their health insurance plan. In general, economists expect competition between insurance providers selling similar products to help hold down insurance costs, provided that people can make informed comparisons across insurers.

Health plans often differ both on the choice of insurance provider and also on a number of other dimensions. For example, plans differ in their *level of coverage,* like deductibles, limits on out-of-pocket payments, and the overall share of medical bills covered by insurance. The health insurance exchanges established by the Patient Protection and Affordable Care Act of 2010 feature four tiers of coverage—platinum, gold, silver, and bronze—which differ in actuarial value. The fraction of the population's medical bills that will be covered by insurance ranges from 90 percent for platinum plans to 60 percent for bronze plans. On HealthCare.gov, the average county has 46 health plans available across these tiers from five different insurers (Department of Health and Human Services 2016). This type of choice over coverage level is present in many other markets as well. For example, many employers offer a choice between a high- or low-deductible health plan, and over

■ *Keith Marzilli Ericson is Associate Professor of Markets, Public Policy, and Law, Questrom School of Business, Boston University, Boston, Massachusetts. Justin Sydnor is the Leslie P. Schultz Professor in Risk Management and Insurance, University of Wisconsin, School of Business, Madison, Wisconsin. Ericson is a Faculty Research Fellow and Sydnor is a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are kericson@bu.edu and justin.sydnor@ wisc.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.51 doi=10.1257/jep.31.4.51

60 percent of people who get health insurance from their employer work in a firm that offers at least two different plan options (Claxton et al. 2016, p. 72).

Giving people choice over health insurance can have positive effects, but it also creates challenges related to both consumer confusion and adverse selection. There is mounting evidence that many people have difficulty understanding the value of insurance coverage, like evaluating the relative benefits of lower premiums versus lower deductibles. Also, in most US health insurance markets, people cannot be charged different prices for insurance based on their individual level of health risk. This creates the potential for well-known problems of adverse selection because people will often base the level of health insurance coverage they choose partly on their health status. As Geruso and Layton review in this issue, avoiding inefficiencies in these situations can require (sometimes imperfect) market regulations.

In this essay, we examine how the forces of consumer confusion and adverse selection interact with each other and with market institutions to affect how valuable it is to have multiple levels of health insurance coverage available in the market. We present an overview of how economists model the value of health insurance for a rational consumer and illustrate ideas using a set of simplified examples with parameters based on US data. We also review evidence on consumer confusion about health insurance and use an example simulation to illustrate the potential effect of consumer confusion on how the market functions.

We highlight a few key points. First, with fully informed consumers and no regulations to limit adverse selection, introducing plans with different levels of coverage can unsurprisingly lead to adverse selection and market unraveling that leaves consumers worse off (on average) than if only a single higher actuarial value plan were available. Having some confused and uninformed consumers can help prevent market unraveling in these situations, and thus raise average consumer welfare (Handel 2013). However, introducing choice across coverage levels can generate large transfers from uninformed consumers to sophisticated consumers. Market institutions, such as risk adjustment, that reduce how adverse selection affects plan prices can make offering choice over coverage levels beneficial to fully informed consumers. Yet these types of regulations are often imperfect and the gains to choice modest. Moreover, even with these regulations, confused consumers can erode any gains from choice.

We also briefly discuss how these issues can affect the value of choice on other dimensions of health insurance and health care. We discuss some policy options for addressing these issues of consumer choice, and some of the broader issues that the problem of consumer errors in making choices in health insurance and health care raises for economists and policymakers.

Modeling the Value of Health Insurance for Informed Consumers

We begin by reviewing the basic building blocks for how economists model the value of health insurance contracts: health risk, the level of coverage in the plans, how those plans are priced, and how consumers value reductions in financial risk. A knowledgeable and informed consumer would take these factors into account in choosing a health insurance policy.

Building Block 1: Health Risk and Medical Spending

Some sources of variation in medical spending can be anticipated ahead of time, but some cannot. Anticipated variation comes from pre-existing conditions and other things that people (or insurers) know. We can think of this as variation between people who can be identified as more- or less-healthy at the time they choose insurance coverage. Unanticipated variation in medical spending is based on unexpected health shocks. For example, a young person with no chronic conditions can expect on average to have low health spending for the year, but might be involved in an accident that generates very large medical bills.

For our example, we want to capture realistic variation in both anticipated and unanticipated medical spending. We started with data from the Medical Expenditure Panel Survey (MEPS), which provides information about individuals' total medical expenditures (and other useful data) for a representative sample of Americans.[1] We focus on survey respondents of working age (18–64) who have health insurance coverage.

Survey respondents are asked, "In general, compared to other people of your age, would you say that your health is excellent, very good, good, fair, or poor?" For our simplified example, we group together people who choose excellent or very good and consider them "healthy" and those who answer good, fair, or poor as "unhealthy." Of course, dividing the population into these two types is a simplification. It ignores additional private information about health spending that an individual might have as well as demographic predictors of spending, such as age. While the question asks the respondents to report his or her health compared to other people of the same age, older individuals are still more likely to report lower health status.

As Table 1 shows, 68 percent of working-age adults are "healthy" and 32 percent are "unhealthy." These two groups differ substantially in their *anticipated* medical spending amounts. Healthy adults on average generate $3,045 in medical bills for the year, while unhealthy adults average more than twice as much.

Within health types, unexpected health shocks will still lead to substantial variation in the amount of medical bills a person has for the year. Table 1 gives a sense of the variability within each type—even unhealthy adults have a 9 percent chance of having no medical spending, and even healthy adults have a 7 percent chance of having over $10,000 in medical spending.

[1] The data is released publicly with a lag, and here we use data from the Agency for Healthcare Research and Quality (AHRQ) 2012 and 2013 Medical Expenditure Panel Survey yearly surveys. Because the MEPS data we use is a few years old and medical costs tend to rise over time, the estimates we provide here will underestimate current medical spending for the population. There is also some evidence (Aizcorbe et al. 2012) that the MEPS data tends to underestimate spending relative to data on claims from employer-sponsored insurance. The basic insights we gain from this exercise, however, are not affected by this issue.

Table 1
Medical Spending for Example of Population with Two Health-Risk Types

Population group	*Percent of population*	*Average medical spending*	*Probability of $0 medical spending*	*Probability of $10,000+ medical Spending*
All adults (18–64)	100%	$4,380	13%	10%
Healthy adults	68%	$3,045	15%	7%
Unhealthy adults	32%	$7,227	9%	18%

Note: Source is authors' calculations from 2012/2013 Medical Expenditure Panel Survey (MEPS) data. We limit the population to those with private insurance coverage for the full year. The split into "healthy" and "unhealthy" is described in the text. MEPS person weights are used to obtain the average spending amounts and the probability of spending thresholds.

When economists analyze health insurance markets, they typically assume that people are aware of the distribution of their possible medical bills for the year and choose their health plan with that information in mind. For our example simulation, when we consider fully informed consumers we assume that our two types of consumers—healthy types and unhealthy types—know the distribution of possible annual medical spending amounts their type could have. We use the observed variation in the data from the Medical Expenditure Panel Survey for each of the health types to create the distribution of possible medical spending each type might have for the year.

Building Block 2: Level of Financial Coverage and Cost-Sharing

The "actuarial value" of a health insurance plan is the percentage of the overall population's medical costs that would be covered by the plan. The person must pay the rest out of pocket, which is called "cost-sharing."

For our example, we consider two different coverage levels: a high and a low actuarial value plan. Each plan is defined by three typical cost-sharing features. The deductible is the amount of money you have to pay *for the year* in medical bills before insurance starts to (partially) cover additional bills. The coinsurance rate is the percentage of *each* bill the individual must pay out-of-pocket once the deductible has been met (the insurance covers the remaining portion). Finally, each plan has a maximum out-of-pocket limit. Once the combination of the deductible and coinsurance payments from the individual for that year hits this level, insurance fully covers all remaining bills. Table 2 describes our two example plans.

The high actuarial value plan, which covers 90 percent of medical costs, is similar to both a fairly generous employer-sponsored plan and to "platinum" plans on the health insurance exchange established by the Patient Protection and Affordable Care Act of 2010. The low actuarial value plan, covering 70 percent of medical costs, is a "high-deductible health plan," similar to the types of plans offered in the silver tier on private health insurance exchanges. High-deductible health plans have also become increasingly common in employer-sponsored plans over the past two decades.

Table 2
Two Examples of Health Plans

Plan type	*Actuarial value*	*Deductible*	*Coinsurance rate*	*Maximum out-of-pocket limit*
High actuarial value	90%	$250	10%	$1,250
Low actuarial value	70%	$2,000	10%	$4,500

Note: The actuarial value of our two example plans is calculated based on the sampling-weighted average amount of spending the plan would cover for people in the Medical Expenditure Panel Survey data.

Of course, real-world choices between insurance plans often include many different coverage levels (for reasons discussed by Geruso and Layton, in this issue). However, we believe that a simple example with two plans is quite helpful for thinking about how consumer choice and adverse selection patterns play out. In particular, the challenges of informed choice and adverse selection would likely be even more problematic with more choice between coverage options. National debates over health care reform during the first half of 2017 included proposals to lift restrictions on the range of coverage levels insurers could offer in the private health insurance market.

Building Block 3: How Insurance Premiums Are Set

Insurance premiums are largely determined by the expected amount of medical bills that will be covered by that insurance. Exactly how expected medical bills are linked to premiums, though, depends on how the market is regulated.

For this example, we focus on health insurance markets with two important regulations: *guaranteed issue*, which means that insurance plans have to sell to anyone who wants to buy, regardless of health status; and *community rating*, which means that insurance plans have to charge everyone in a given plan the same price, regardless of health status. These regulations are similar to circumstances in the employer-sponsored health insurance market and provisions of the 2010 Affordable Care Act.[2]

The premiums for insurance typically include a "load" on top of the medical costs covered by the insurer, which covers markups for insurer profit and administrative costs associated with processing claims. For our example, we assume a load factor of 1.25, implying that insurers charge $1.25 in premiums for each $1 in expected medical bills the insurance covers. This assumption is consistent with regulations from the Patient Protection and Affordable Care Act of 2010 that require insurers

[2] Technically, the Patient Protection and Affordable Care Act of 2010 has "modified community rating," as it permits premiums to vary based on a limited set of factors (for example, geography, age, smoking status) but not health status. Prior to this law, the individual insurance market typically did not have community rating or guaranteed issue, but employer-based insurance did.

to maintain a “medical loss ratio” (medical claim payments divided by premiums) of 80 percent or higher. The load factor is the reciprocal of the medical loss ratio.

For our example, assuming that plans covered the average population, the premiums for the high actuarial value plan would be set at $4,930, while the low actuarial plan would cost $3,810. These premiums come from taking the expected covered spending with each plan (the average medical spending multiplied by the plans’ actuarial value) and then multiplying by the load factor of 1.25. These premiums are roughly in line with the types of premiums seen for real health insurance policies during this time period. They are somewhat lower, though, than is typical for employer-sponsored insurance, probably in part because the Medical Expenditure Panel Survey data tends to report lower total medical spending than data from insurance claims in employer-sponsored insurance (Aizcorbe, Liebman, Pack, Cutler, Chernew, and Rosen 2012).

These premiums are what we would expect if these plans enrolled the same average population. When people have choices over plans, the premiums for plans may be affected by the health status of people who enroll in the plan. We discuss this equilibrium process below.

Building Block 4: Risk Aversion and the Value of Insurance

In a standard model, more generous insurance is valuable primarily because it reduces the financial risk that people face (Einav, Finkelstein, and Levin 2010).[3] For someone who is risk averse, a health insurance plan that covers an additional $1,000 in expected medical bills will actually provide more than $1,000 in value because it also reduces the variation, or the risk, in spending the person faces. However, people differ in the optimal level of coverage they prefer.[4] People who are more risk averse will prefer plans with more coverage. Similarly, people who expect to have higher medical spending will see more value in plans with more coverage.

Economists typically model risk aversion using the idea of concave utility functions that capture the diminishing marginal utility of wealth. With a concave utility function, a person prefers more stable wealth over more variable wealth, even if it means giving up some wealth in expectation. Economists often turn to a few common mathematical functions for the concave utility functions that allow them to quantify the value people get from reducing risk. For our example, we use the constant absolute risk aversion utility function. With this utility function, the parameter r (known as the coefficient of absolute risk aversion) governs how risk averse a person is, with higher r implying more risk aversion. To put the coefficient of absolute risk aversion

[3] We are only discussing the choice between insurance plans of different coverage levels. Having insurance (as compared to being uninsured) may be valuable other reasons, such as access to care (for example, Nyman 1999) and access to negotiated rates for health care services.

[4] Individuals may also vary in how they value medical care. There is an interesting question of how variation in the price elasticity of medical care demand, which can be viewed as related to moral hazard, affects the value of choice (for example, Einav, Finkelstein, Ryan, Schrimpf, and Cullen 2013). There isn’t a simple model we know of that can be used to determine this factor’s effect on the value of choice, and for this essay, we set aside this issue.

into context, consider a person who holds a lottery ticket with a 50 percent chance of winning either $1,000 or nothing. A risk neutral person, with $r = 0$, values that lottery ticket at its expected value of $500. A risk-averse person, however, would be willing to sell the ticket for less than $500, to receive a certain gain and avoid the risk associated with the lottery. A person with $r = 0.001$, for example, would be willing to take as little as $380 for sure instead of the lottery.

There is no agreed upon range of typical risk aversion in the population. For our example, we assume that people vary in their level of risk aversion, which we assume is uniformly distributed between 0 (risk neutral) and $r = 0.001$. This level of variation allows us to model a population with substantial variation in risk aversion and is broadly in line with the range of risk aversion in the health insurance literature (for example, see the distributions compared in Ericson, Kircher, Spinnewijn, and Starc 2015).

We can then calculate the expected utility for both types of consumers—healthy and unhealthy—at different levels of risk aversion for both the high and low actuarial value plans. The expected utility for each plan is the weighted average of their utility for each level of total spending the person might have (premium + out-of-pocket costs) given the distribution of medical spending for their health type. Once we have these expected utilities for each plan, we can then calculate the dollar value of the difference in consumer welfare each person would have for different plans. We do this by calculating how much money could be given to (taken from), a person for the case of welfare gain (loss), to make them indifferent between staying in a baseline plan option versus moving to an alternative plan.[5]

The value of the higher level of coverage in our example is strongly affected by a person's health status. Assuming that plans were available at the premiums set for the average population described above, a risk-neutral healthy type would perceive that moving from the high to low actuarial value plan would increase their welfare by $350. This gain comes because the low actuarial value plan reduces the premium by $1,110, but increases a healthy types' expected out-of-pocket medical costs by only $760. On the other hand, a risk-neutral unhealthy person moving from the high to low actuarial value plan sees a small decrease in welfare: their expected out-of-pocket health care costs rise by $1,160, which is more than the reduction in premiums.

With a higher level of risk aversion, the perceived value of insurance increases. Healthy types with higher levels of risk aversion are worse off in the low actuarial value plan due to the increased financial risk they face, even though their expected total spending will fall. Moreover, unhealthy types with high risk aversion experience large losses (for example, $750 or more) in the low actuarial value plan, even though their expected total spending would increase by only $47. For our example

[5] The constant absolute risk aversion utility function is defined as $u(w) = 1 - e^{-rw}$ for $r > 0$ and $u(w) = w$ for $r = 0$. With this function, the expected welfare of having the low actuarial value plan relative to a benchmark of the high actuarial value plan can be calculated by first calculating the expected utility for each plan and then calculating: $-\frac{1}{r}\ln\left(\frac{EU(low\ AV)}{EU\ (high\ AV)}\right)$.

simulation, averaging over the different people, we find that average consumer welfare would be $127 higher with everyone in the high actuarial value plan than if everyone were in the low actuarial value plan. Thus, if a social planner had to select from only one plan from these two options, the high actuarial value plan would be preferred.

Evidence on Consumer Confusion in Health Insurance Choices

In the workhorse model of insurance choices, people are active deciders with full information who make choices about insurance plans based on their risk-averse expected utility over final wealth outcomes. However, recent empirical research offers evidence against this characterization. We first review some of that evidence, and then examine what happens to the value of choice if consumers are confused when selecting plans.

First, few people have a strong understanding of health insurance plans. For example, Loewenstein et al. (2013) found that in a representative sample, fewer than 14 percent of people could correctly answer a series of multiple-choice questions about key health-insurance terms, including deductible, copay, coinsurance, and maximum out-of-pocket costs. People in this survey were also overconfident, thinking they understood the terms better than they really did. Many people cannot easily calculate how much money they would spend with different plans even if they knew exactly what medical spending they would have. For example, Johnson et al. (2013) found that even with incentives, a majority of people could not identify the cost-minimizing plan from a few choices when given a specific amount of anticipated medical spending and details about premiums and cost-sharing. Even people with high education and income have difficulty understanding health insurance options. In the Johnson et al. (2013) study, MBA students were more likely to select cost-minimizing plans, but even in that group a significant share got it wrong. Handel and Kolstad (2015) found that among employees at a high-paying firm (median income around $125,000), a significant share were confused about details of different plan options.

Second, it is challenging for people to both forecast their possible range of medical needs for the year and also then to know how those needs would map to medical bills. The price of health care services can vary dramatically (Cooper, Craig, Gaynor, and Van Reenen 2015) and providers often cannot tell patients what the price will be in advance (Rosenthal, Lu, and Cram 2013). People also often have distorted perceptions of risk in the context of insurance, sometimes ignoring the possibility of low-probability events altogether and sometimes overreacting to certain salient risks (for example, Johnson, Hershey, Meszaros, and Kunreuther 1993). Although we are not aware of research on how people perceive their health risks when making health insurance decisions specifically, it is likely that they may be subject to similar biases.

Third, individuals are often not active deciders. Many people stick with an initial choice of an insurance plan, even if premiums or other features change dramatically. For example, Handel (2013) finds this pattern in a study of employees selecting plans offered by their employer, and Ericson (2014a) finds it in a study on choices that seniors make between Medicare Part D prescription drug plans. With consumer inertia, even people who originally make an optimal choice may not later be enrolled in the plans that offer the highest expected utility.

Fourth, an abundance of research in psychology and economics shows that people often become overwhelmed when faced with many options, a phenomenon known as "choice overload" (for discussion, see Iyengar and Kamenica 2010). Choice overload can cause people to gravitate toward simple options or to focus on isolated features of products in ways they would not do if there were fewer options available. It can even cause people to disengage entirely and opt not to purchase a product at all. To date, direct evidence of choice overload in health insurance is limited. Bhargava, Lowenstein, and Sydnor (2017) find that people made seemingly suboptimal choices when employers offered them many possible plans, but provide some evidence the number of options per se was not the problem. However, choice overload could be one reason behind consumer inertia in health insurance markets

With these challenges, a number of people may not be able to make choices that maximize their expected utility. For example, Abaluck and Gruber (2011) document that seniors often choose Medicare Part D prescription-drug insurance plans that are not on the efficient frontier: that is, they could purchase cheaper expected cost plans that provide equally good risk protection. Bhargava, Loewenstein, and Sydnor (2017) studied employees' plan choices at a firm where many of the available plans were financially dominated by other options: in one case, a plan with a $500 lower deductible cost an additional $600 in yearly premiums! Yet the majority of employees ended up choosing a dominated plan. Sinaiko and Hirth (2011) also document violations of dominance with a situation where many employees selected a plan with more difficult access to specialists, even though it had the same cost as a more flexible option.

Finally, some people may have some other objective rather than reducing variation in annual health-care spending. Out-of-pocket costs may have different consequences for people than spending on premiums, due to self-control issues, loss aversion, or liquidity constraints. For example, people may recognize that if they face out-of-pocket costs like deductibles and co-pays, they may pull back on valuable health care, like treatment for chronic diseases (Baicker, Mullainathan, and Schwatzstein 2015; Brot-Goldberg, Chandra, Handel, and Kolstad 2017). Liquidity constraints also can increase the value of paying smooth regular premiums over possibly more lumpy out-of-pocket costs, which in some extreme cases can make it rational to purchase a plan that otherwise seems dominated by a different choice (Ericson and Sydnor 2017). Economists are just beginning to incorporate these ideas into their analysis of healthcare markets.

The Value of Choice of Coverage Levels

In this section, we build upon the parameters laid out earlier to illustrate the effect of adding choice over coverage levels under different market institutions and different levels of consumer errors about plan choice. We take a situation where only the high actuarial value plan is available as the benchmark and then explore what happens when the low actuarial value plan is introduced.

The results of our example simulation are summarized in Table 3, which we refer to throughout this section. As described above, we assume that there are healthy and sick people in the market and use the Medical Expenditure Panel Survey to create the distribution of medical spending each type might have for the year. We assume that each person in the market has a level of risk aversion, uniformly distributed from risk neutral ($r = 0$) to substantially risk averse ($r = 0.001$).

Fully Informed versus Uninformed Choice

Fully informed individuals select between the high and low actuarial value options based on which one gives them the higher expected utility, given their health type, risk aversion, and the premiums for the two plans. We assume that fully informed people correctly anticipate the distribution of possible spending, pay attention to the premiums for the plans, and understand how the two plan options will affect their distribution of possible out-of-pocket medical costs.

For uninformed or confused consumers, we consider two ways these consumers might choose: randomly and nonrandomly. In our example, random choosers select each plan with 50 percent probability regardless of that person's underlying risk aversion or health status. For a useful example of a nonrandom error, we will focus on the "inattention heuristic," in which confused consumers select the plan they would have selected if plan premiums were fixed at the levels appropriate for the average population. This error can be thought of as representing a form of inertia: some consumers initially choose plans optimally when they are offered at population-average premiums, but then fail to pay attention if premiums change over time. It can also represent a bias people may show when selecting plans, if those who struggle to understand and compare plan options might naively believe that plans are priced "fairly" for the average population. In this setting, people may select plans based on their relative levels of health risk and risk aversion (what Kamenica 2008 terms "contextual inference"), but without properly incorporating the real differences in premiums into their selection.

There are, of course, many other ways people might choose plans. These two examples of alternative choice processes—random and "inattention heuristic"—however, allow us to discuss some forces that arise when considering how consumer confusion interacts with adverse selection and health insurance market institutions.

In our example, we show the welfare effects of introducing choice for these confused consumers, basing our calculation of welfare on how they would perceive the value if they were fully informed. If someone selects a plan they would not have if they were fully informed, we refer to that as a choice "error." However, we do not

Table 3

Example Simulations of the Effect of Introducing Choice Relative to Baseline with Only the High Actuarial Value Plan Available

			Average change in per-person consumer welfare			
Market environment	*Share of uninformed choosers*	*How the uninformed choose*	*Overall*	*Healthy types*	*Unhealthy types*	*Empirical pattern in equilibrium*
No regulation: premiums reflect average costs of people who enroll in that plan.	0%	—	*−$127*	$25	*−$449*	Market fully unravels so all choose low actuarial value plan.
	50%	Randomly	*−$78*	$31	*−$310*	Plan selection by health type, mitigated in part by random errors. Moderately elevated premium differences between plans. Both plans selected by some informed and some uninformed choosers.
	50%	Nonrandom: inattention heuristic	*−$59*	$151	*−$505*	Plan selection strongly related to health type and large resulting premium differences between plans. Only uninformed select the high actuarial value plan.
Regulation and risk adjustment: premium differences between plans fixed for average population	0%	–	$9	$42	*−$60*	Plan selection strongly related to health type. Low actuarial value plan attracts the unhealthy types as well as healthy types who are highly risk averse.
	50%	Randomly	*−$27*	$27	*−$143*	Plan selection by health type, mitigated in part by random errors.
	50%	Nonrandom: inattention heuristic	$9	$42	*−$60*	Same as the fully informed case because the inattention heuristic is appropriate to the market environment in this case.

Note: Authors' calculations based on simulation described in the text.

model the costs of processing information about health plans, so choosing with a heuristic may be sensible for some people given the challenges of trying to make an informed choice.

The Effects of Choice in Markets with No Regulations to Address Selection Effects

In a competitive insurance market, with no additional regulations, the premiums for an insurance policy will reflect the average covered spending of the people who choose that plan. Plans that attract more unhealthy types will have higher prices. At the same time, people respond to the price of different plans and will tend to switch toward lower-cost plans. Eventually, the market reaches a *competitive equilibrium* in which premiums are equal to a plan's costs (plus load) and no one wants to switch plans. Einav and Finkelstein's (2011) overview article in this journal shows how competitive equilibrium is reached with this sort of pricing.

This equilibrium can "unravel," so that only the plan with the least coverage is actually purchased; for example, this outcome will arise if everyone were fully informed in our example simulation. Suppose plans started out priced for the average population, as described earlier. In that case, all of the unhealthy people want the high actuarial value plan, but among the healthy types, those with below-average risk aversion prefer the low actuarial value plan. Those choice patterns, though, mean that the low actuarial value plan will only have healthy types enrolled, while the high actuarial value plan will have a higher share of unhealthy types than are in the total population. That puts upward pressure on the premiums for the high actuarial value plan and downward pressure for the other plan. As the difference in premiums rises, more people will prefer the low actuarial value plan. Ultimately, premium differences rise to the point that the market fully unravels, with everyone ending up choosing the low actuarial value plan. In this case, introducing the low actuarial value option in this environment has the same end result as only offering that option.

The first row of Table 3 shows the results of this process on consumer welfare relative to the benchmark situations where only the high actuarial value plan was available. Average welfare *falls* by $127 per person when choice is introduced. On average, unhealthy types face losses of $449 per person from this shift, while the healthy on average gain $25. Those who are more risk averse lose more from the shift to low actuarial value plans, so that even healthy types with above average risk aversion lose from introducing choice. The basic tradeoffs at play here are that 1) healthy types would prefer to have less coverage in the market because then they have to do less cross-subsidizing of the unhealthy types and vice versa for unhealthy types (they would prefer to have more coverage, expecting it will be cross-subsidized by the healthy types) and 2) reducing coverage is especially costly for those who are more risk averse.

This stark example involves an insurance "death spiral." In other examples, the market may not unravel completely and there may be less of a welfare loss (or even a possible welfare gain, depending on other parameters) from having everyone enrolled in the low actuarial value plan. However, the basic insight of this example is relevant for considering the value of offering choice. Cutler and Reber (1998), for example, show how a milder form of adverse selection operated when additional health insurance choices were offered to employees of Harvard University. Handel, Hendel, and Whinston (2015) also simulate equilibrium in the state-level health insurance exchanges established by the 2010 Affordable Care Act, and predict unraveling to the lowest level of coverage unless there are regulations and risk adjustments for premiums.

The second two rows of Table 3 show what happens in equilibrium if instead of fully informed consumers, half of consumers are confused and choose either randomly or with the inattention heuristic. Consumer mistakes in selecting health plans can have two effects, as discussed in Handel (2013) and formalized in Handel, Kolstad, and Spinnewijn (2015). On the one side, mistakes lead people to sort less optimally between plans, which can lower welfare. However, random mistakes blunt the force of adverse selection and improve average welfare relative to the case where everyone is fully informed. Even a small fraction of random choosers helps to stabilize premium differences and prevents the market from unraveling. As the fraction of random choosers grows, premium differences will fall toward the population-average level, and average welfare improves.

In Table 3, we see that if half of people choose randomly, the welfare loss from introducing choice is $78 per person, about two-thirds the size of the loss when everyone chooses rationally. With half the people choosing randomly, the difference in premiums between the plans is around $1,760, which is around $600 higher than the premium difference would be with premiums set for the average population, but still low enough for some informed risk-averse unhealthy types to prefer the high actuarial value plan. However, no matter how many random choosers there are, we still find in this example that choice lowers welfare as compared to the benchmark case with only the high actuarial value plan available.

This result highlights a paradox of consumer confusion in health insurance markets where premiums can be affected by adverse selection (first highlighted clearly by Handel 2013). A decision aid that helps people select an optimal insurance plan could have a big return for them. For example, in our simulation, helping a single highly risk-averse unhealthy person avoid the mistake of randomly selecting the low actuarial value plan could provide that person with a few hundred dollars of consumer welfare value at the equilibrium prices that prevail when half of people are confused. However, if everyone started to choose optimally, premiums would change and the equilibrium would unravel, and only the low actuarial value plan could be purchased. The people who benefit the most from having some confused consumers in the market are the informed choosers who are unhealthy and highly risk averse. Those people benefit from having the high actuarial value available and in particular benefit from the healthy types who randomly select that plan with them and help keep its premiums down. Even among the confused consumers, the only ones who really benefit in equilibrium from having everyone become informed are the healthy types who are not very risk-averse. Only those types like the equilibrium where everyone is in the low actuarial value plan.

The third row of Table 3 shows the results if the uninformed consumers follow the nonrandom "inattention heuristic." Again, confused consumers help to ensure that the market does not completely unravel from adverse selection. However, strong selection effects arise. Among those with the inattention heuristic, all of the unhealthy types (along with the more risk-averse healthy types) select the high actuarial value plan. That tends to push up the premium difference between the high and low actuarial value plans. In fact, with half the people choosing based on the heuristic, the equilibrium premium differences between the two plans would be nearly $2,500, which is more than the difference in deductibles between the plans and nearly as large as the difference in maximum out-of-pocket limits. Only the uninformed consumers would be selecting the high coverage plan in this situation. In effect, the premium differences look like the market has unraveled, but uniformed consumers do not realize it, and select the high actuarial value plan even though it is extremely expensive.

This pattern appears to be at play in some employer-sponsored insurance settings in which employees can choose between different levels of coverage. Handel (2013) and Bhargava, Loewenstein, and Sydnor (2017) analyzed data from firms where the difference in premiums between plans with higher and lower levels of coverage had risen to the point where the higher coverage plans were so expensive that they were dominated. Yet a substantial share of employees selected the higher coverage plans, due in part to inattention and inertia (Handel 2013), but also due to active choice processes like the naive sorting by health type we simulate with our inattention heuristic (Bhargava, Loewenstein, and Sydnor 2017).

With these nonrandom errors, the average welfare loss from introducing choice is again lower when there are uninformed consumers relative to the case with only fully informed consumers. With nonrandom errors, however, there is a clearer transfer of welfare from uninformed to informed consumers. In this case, many uninformed consumers are selecting the high actuarial value plan and paying premiums that reflect the fact that many of them are unhealthy types. The informed consumers, especially the healthy types, get a large benefit from selecting the low actuarial value plan, which has an advantageous mix of more healthy types. In effect, the option to choose the low coverage plan allows these informed healthy types to avoid pooling with many of the unhealthy types who naively select the higher coverage plan. We see, for example, in our simulation with half the people choosing with the heuristic, that the average welfare for healthy types is $151 higher per person with choice, while the loss for unhealthy types is over $500 and worse than the case where everyone is fully informed and the market unravels.

Thus, consumer confusion has distributional consequences. When mistakes help to stabilize the market, confused consumers are in effect subsidizing savvier consumers. The individual welfare losses for those making the mistakes may be large, especially when these errors are nonrandom. Moreover, economically vulnerable populations, including those with less education, lower incomes, the elderly, and those with health problems, are all more likely to have problems selecting optimal health insurance plans (for example, Loewenstein et al. 2013; Bhargava, Loewenstein, and Sydnor 2017). The type of consumer welfare calculations we have shown in our example, and which are the norm in much of the economic analysis of health insurance markets, treat a dollar of value the same for all people. However, a dollar has greater utility for someone with lower income, which means the average *social* welfare effects of offering choice might be worse even on average when a substantial share of consumers are confused about their choices.

The Effects of Choice in Markets with Regulations and Risk Adjustments

Many insurance markets have regulations and policies in place that seek to prevent the problems with selection by health type and the consequent market unraveling. Many insurance markets with a consumer choice component, including the Affordable Care Act health insurance exchanges, Medicare Part D, and Medicare Advantage, use a system of "risk adjustments" that transfers money from plans that enroll healthier individuals to plans that enroll sicker individuals. While the details vary by market, risk adjustments tend to target equalizing the average cost of enrollees between plans (for a review, see Van den Ven and Ellis 2000; Geruso and Layton, this issue). For our simulation, we model risk adjustment as perfectly equalizing costs. However, actual risk adjustment tends to be imperfect, which means that the insights from markets with adverse selection remain important even when the market has some risk adjustment.

The bottom panel of Table 3 shows what happens in our simulations when effective risk adjustments are in place. In particular, we assume that the risk-adjustment process stabilizes the difference in premiums between plans at the level appropriate for the population on average (roughly $1,110). The level of the premiums adjusts in equilibrium to be high enough to ensure that the plans collect enough money overall to cover the expected covered medical spending plus loads.

When all the people choosing plans in the market are fully informed, risk adjustment helps make offering choice welfare-enhancing on average, relative to the benchmark case with only the high actuarial value plan available. The risk adjustments keep the premiums between plans stable even though the pool of enrollees for the low actuarial value plan is much healthier (in our example, only healthy individuals choose that plan). In our simulation, we see an average increase in consumer welfare from having choice of $9 per person, with healthy people gaining around $42 per person on average and unhealthy people losing $60 per person.

It may seem surprising that the unhealthy people would lose anything when the low actuarial value plan is introduced, since they all select the benchmark high actuarial value plan and premium differences between plans do not adjust. However, risk adjustments that stabilize premium differences based on the appropriate difference for the average person are not fully efficient. At these premium differences, the healthy types get a larger discount from selecting the high actuarial value plan than is warranted by their reduction in covered spending from reducing coverage. Equivalently, the unhealthy types pay less for higher coverage than the increase in covered spending they receive. This remaining inefficiency with risk adjustment is nearly impossible to avoid and results in the level of premiums (for both plans) rising above the level they would be if everyone were enrolled in a single plan. Ultimately, risk adjustment technology can generate positive welfare gain from choice for informed consumers, but since risk adjustment is not fully efficient, those gains are not guaranteed, may be modest, and may favor the healthy types.

In contrast to the situation with completely unregulated premiums, random mistakes tend to lower consumer welfare when there is risk adjustment. Once risk adjustment reduces the problem of adverse selection, the primary effect of random choice is that people sort less well into the plan that is best for them. In our simulation, with 50 percent of the population choosing randomly, there is now an average loss of $27 per person from introducing choice. Because the gains to choice are so small even in the best case with fully informed consumers, it only takes a small fraction of people making mistakes (about 15 percent in our simulation) for the value of introducing choice to be negative.

As the fraction of those who choose their insurance policy randomly rises, welfare continues to fall, both on average and for each health type. However, random choice will tend to lower welfare most quickly for risk-averse unhealthy types. This group receives an especially strong benefit from additional insurance coverage, and conversely, the welfare costs for this group of wrongly choosing the low actuarial value plan is especially harmful. Thus, expanding choice of coverage levels creates a distributional tradeoff between relatively small gains for many of those who use their added choice wisely (again, given that they are paying a risk-adjusted premium for the health group to which they belong) but the risk of large losses for those in the risk-averse unhealthy group who make random errors.

The final row of Table 3 highlights that unlike random choosing, there are no adverse consequences to uninformed consumers using the inattention heuristic when premiums are controlled by risk adjustment. In this case, the heuristic happens to be perfectly matched to the market environment and people are choosing in the same way they would if they were fully informed. Of course, other forms of nonrandom choice errors could lead to welfare losses even when there are risk adjustments. However, in general, effective risk adjustments will help dampen the negative consequences of nonrandom choice errors in which selection is partly related to one's health status.

Nudges and Other Interventions to Improve Consumer Choice

Behaviorally informed policies can seek to address the consumer confusion and biases that are widespread in health insurance choice. Initial research suggests that such policies can sometimes be effective, but also suggests that they face substantial challenges.

First, *standardizing policy health insurance plan options* within a level of actuarial value can make it easier to compare plans. For example, Ericson and Starc (2016) examined an earlier natural experiment in which health plans on the Massachusetts health insurance exchange were standardized within each tier. Standardization led consumers to choose more generous health insurance plans and to substantial shifts in brands' market shares. However, seemingly small details about the design of choice platforms also affect consumer choice, such as the labels attached to tiers (like calling one level "bronze") and the order in which plans are sorted (Ubel, Comerford, and

Johnson 2015). The HealthCare.gov website which provides information for the state-based health insurance exchanges recently introduced standardized options for each coverage tier called "simple choice" plans, which all have the same deductible and co-pay levels, although nonstandardized options do still remain available. Future studies will likely investigate how this change affects choices.

Second, *providing personalized information* to health insurance shoppers may help their decision-making. There is substantial variation in what is included in these consumer decision support tools (for discussion, see Wong, Polsky, Jones, Wiener, Town, and Baker 2016). Many markets provide out-of-pocket cost calculators that help people estimate how plan choices will affect their expected spending, based on demographics, health status, and/or past claims history. For instance, Medicare Plan Compare sorts Medicare Part D prescription drug insurance plans based on expected costs given the drugs an individual is currently taking. Similarly, for health insurance bought on HealthCare.gov, the site presents expected annual spending amounts for a few representative spending scenarios.

Research on whether out-of-pocket calculators meaningfully affect plan choices is limited, and the results are mixed. Earlier work found that providing personalized information about out-of-pocket costs did induce people to switch plans (Kling, Mullainathan, Shafir, Vermeulen, and Wrobel 2012). However, Abaluck and Gruber (2016) find that providing an out-of-pocket cost predictor at a large employer had little effect on plan choices. Similarly, Ericson, Kingsdale, Layton, and Sacarny (2017) find that a randomized experiment providing people with personalized information about the potential premium savings they could have in the state-level health insurance market induced more people to shop more actively—but did not lead people to switch plans.

Third, *smart defaults* offer a more aggressive approach to nudging consumers, in which the decision-assistance software actually chooses a default plan based on the consumer's information, but the consumer can override that default if desired. Smart defaults have been explored in Medicare Part D (Hoadley, Thompson, Hargrave, and Merrell 2007), and can be used either at the point of initial enrollment or to switch inattentive consumers (Ericson 2014b). Johnson et al. (2013) suggest that it may be necessary to couple calculators with smartly chosen defaults or recommendations to meaningfully improve consumer choices.

Efforts to nudge consumer choice will create tradeoffs of their own. The example of out-of-pocket cost calculators in health insurance can be used to illustrate the concerns. They are typically implemented as an "expected value nudge" that recommends a plan that minimizes a person's total expected health costs (premiums plus out-of-pocket costs for cost sharing). Importantly, most out-of-pocket calculators do not include a measure of risk aversion (although Picwell.com offers a counterexample). As a result, healthy people will typically be nudged to choose low actuarial value plans, even though a healthy person with high risk aversion might prefer a plan with high actuarial value. Such out-of-pocket cost calculators encourage sorting by health status. As noted above, nudges that reduce random errors can worsen adverse selection and lead to market unraveling. Nudges to improve consumer choice are more likely to lead to an overall social gain if accompanied by well-designed risk-adjustment payments to limit the effects of adverse selection.

Finally, policymakers could turn toward explicitly *limiting the amount of choice available* in the market. This is an ongoing debate in many areas of health care policy. For example, the Patient Protection and Affordable Care Act of 2010 regulates the range of options for coverage levels available in the private health insurance exchanges. Plans are required to cover a common set of essential health benefits and there are four allowed levels of actuarial value. These regulations are contentious and there have been proposals to relax them. As our essay has sought to highlight, there are real tradeoffs involved with offering additional choices of health insurance. On one side, a greater range of choice creates opportunity to raise consumer welfare as buyers sort themselves into the options they prefer. On the other side, a greater variety of choice also raises the possibility of adverse selection dynamics as well as losses due to consumer mistakes.

Other Dimensions of Choice: Value and Pitfalls

Choice over coverage levels, in which insurance plans can be ranked by their financial generosity and actuarial value, can be modeled very tractably. This dimension of choice has been most extensively studied. However, health insurance and health care more broadly pose many other choices. Features of insurance plans that are not reflected in actuarial value create many of the same tradeoffs and potentials for interactions between consumer confusion and adverse selection that we highlighted here for choice over coverage levels.

For example, consider two insurance plans. One is a "managed care" plan with strong limitations on the network of doctors and hospitals a patient can see, and a low degree of cost-sharing. Another is a "consumer-directed" plan with a broad network of doctors and hospitals, few limits on access to specialists, but with a high degree of cost-sharing. Many hybrids of these approaches exist in US health insurance markets. On one hand, people can benefit from choosing the breadth of provider network they prefer. On the other hand, such choices create potential for problems with consumer confusion and adverse selection. However, economists know less about and have less of a well-established framework for modeling how people value access to different networks of providers (for discussion, see Ericson and Starc 2015).

Even within types of insurance arrangements—managed care versus consumer directed—consumer confusion is likely to play an important and varied role. Researching the networks included in any insurance plan can be exceedingly difficult. Anticipating how network access might matter in the future is even more difficult. Recent research has also highlighted the potential for "surprise billing" for medical services where patients get charged for out-of-network doctors at hospitals that are in network for their insurance (Cooper, Scott-Morton, and Shekita 2017).

Most people are not well positioned to assess the value of different medical procedures, which can create a conflict of interest between providers and patients for some services that are of questionable value. It is also difficult for patients to assess both the quality of services and the prices they will be charged from different providers. Consumer-driven health plans are increasingly providing people with portals that facilitate "shopping" between plans by providing information on the prices for services and quality ratings for providers. However, the evidence so far suggests that many of these tools are not very effective (for example, Brot-Goldberg et al. 2017).

Finally, a recent literature, not specific to health care, looks at how firms that have the ability to design contracts may seek to exploit consumers by obfuscating the true price or generosity of the contract (Akerlof and Shiller 2015). In some cases, this obfuscation leads to redistribution from less-sophisticated to more-sophisticated households. Gabaix and Laibson (2006) refer to these cases as situations where products have "shrouded attributes" and highlight the example of credit cards, in which consumers who end up carrying costly debt and paying fees partly subsidize savvier and more financially secure consumers who get the convenience benefits and rewards from credit cards. In other cases, firms can make larger profits on socially wasteful products (Heidhues, Kőszegi, and Murooka 2017; for a review of the theoretical literature, see Grubb 2015).

While the literature on firms exploiting consumer confusion in health insurance has been limited to date, some evidence suggests that these insights may apply in the health care setting. For example, Ericson (2014a) showed that insurers offering Medicare Part D plans raised their prices over time on existing plans to take advantage of consumer inertia while simultaneously introducing new plans into the market at lower prices to attract attentive new customers. Similarly, our discussion in this essay highlights the possibility that when employers offer their employees multiple coverage level options, a situation can arise in which healthier and more sophisticated employees can take advantage of cheaper plans that allow them to avoid pooling together with unhealthier employees who are not sophisticated about their plan choice. Moreover, more sophisticated individuals might be able to follow plan rules more closely and get more value out of a given plan (for instance, by making sure to get pre-authorization or by effectively appealing a denied claim).

Discussion

Offering choice over the type of insurance policy or allowing people to select into plans with different networks of doctors may be beneficial. But as this essay has suggested, the additional choice is not an unmixed blessing.

The many and expanding dimensions of consumer choice in health insurance and health care in the US present challenges not only to individuals but also to economic modeling. For example, the Congressional Budget Office (CBO) is tasked with simulating how people will select individual health insurance plans to help inform legislators. For its simulations, the CBO does not use a model based on the classical expected utility approach, in part because this standard economic framework does not capture many of the forces driving actual choice behavior (see for example https://www.cbo.gov/publication/45427, slide 9). Instead, the CBO bases its projections on elasticities measured from data in actual markets, which at least partially capture the effects of inertia, inattention, and consumer biases. Of course, a limitation of this elasticity approach is that elasticities observed in the past may change in different settings or when changes in regulations affect the extent of consumer confusion.

Policymakers and economists also need to wrestle with the fundamental challenge of how we judge what policies make people better off. Economists typically try to answer these questions by observing the choices people make, and then drawing inferences about peoples' preferences. However, consumer confusion means that the choices people make about health plans may not be directly informative about their underlying preferences. Studies that directly measure consumer confusion and use that information to map choice to welfare (for example, Handel and Kolstad 2015) are an important step forward. More work is needed in that direction.

Even the basics of how economists should evaluate welfare in environments where decision-makers have biases and confusion is a contentious issue (Beshears, Choi, Laibson, and Madrian 2008; Bernheim and Rangel 2009). For example, individuals with loss aversion may try to avoid being exposed to out-of-pocket costs. Such people may desire a high actuarial value plan even if the premiums are very expensive, or they may steer away from the combination of a high-deductible insurance plan with a health savings account—even if these options will save them money. Should this form of loss aversion be considered a "mistake" and nudges implemented to guide those with high loss aversion in another direction even if it makes the person miserable? Economists have not yet been able to offer clear guidance to policymakers on this issue.

Both economists and policymakers should pay more attention to how the complexity on many dimensions of modern health insurance in the United States creates confusion for consumers and can erode the benefits of competition. Given the complexity of healthcare and health insurance markets, health economists and health policy experts must in part also be behavioral economists with an eye toward understanding how people process information and decide, and how those forces shape health care markets.

References

Abaluck, Jason, and Jonathan Gruber. 2011. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program." *American Economic Review* 101(4): 1180–210.

Abaluck, Jason, and Jonathan Gruber. 2016. "Improving the Quality of Choices in Health Insurance Markets." NBER Working Paper 22917.

Agency for Healthcare Research and Quality (AHRQ). 2012 and 2013. Medical Expenditure Panel Survey. https://meps.ahrq.gov/mepsweb/.

Aizcorbe, Ana, Eli Liebman, Sarah Pack, David M. Cutler, Michael E. Chernew, and Allison B. Rosen. 2012. "Measuring Health Care Costs of Individuals with Employer-Sponsored Health Insurance in the U.S.: A Comparison of Survey and Claims Data." *Statistical Journal of the IAOS* 28(1–2): 43–51.

Akerlof, George A., and Robert J. Shiller. 2015. *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton University Press.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein. 2015. "Behavioral Hazard in Health Insurance." *Quarterly Journal of Economics* 130(4): 1623–67.

Bernheim, B. Douglas, and Antonio Rangel. 2009. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *Quarterly Journal of Economics* 124(1): 51–104.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian. 2008. "How Are Preferences Revealed?" *Journal of Public Economics* 92(8–9): 1787–94.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor. 2017. "Choose to Lose: Health Plan Choices from a Menu with Dominated Options." *Quarterly Journal of Economics* 132(3): 1319–72.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad. 2017. "What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics." *Quarterly Journal of Economics* 132(3): 1261–318.

Claxton, Gary, Matthew Rae, Michelle Long, Anthony Damico, Bradley Sawyer, Gregory Foster, Heidi Whitmore, and Lindsey Schapiro. 2016. *Employer Health Benefits: 2016 Annual Survey*. Menlo Park, CA*:* Kaiser Family Foundation.

Cooper, Zack, Stuart V. Craig, Martin Gaynor, and John Van Reenen. 2015. "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured." NBER Working Paper 21815.

Cooper, Zack, Fiona Scott Morton, and Nathan Shekita. 2017. "Surprise! Out-of-Network Billing for Emergency Care in the United States." NBER Working Paper 23623.

Cutler, David M., and Sarah J. Reber. 1998. "Paying for Health Insurance: The Trade-off between Competition and Adverse Selection." *Quarterly Journal of Economics* 113(2): 433–66.

Department of Health and Human Services. 2016. "Health Insurance Marketplace Premiums after Shopping, Switching, and Premium Tax Credits, 2015–2016." ASPE Issue Brief, April 1, Office of the Assistant Secretary for Planning and Evaluation. https://aspe.hhs.gov/system/files/pdf/198636/MarketplaceRate.pdf.

Einav, Liran, Amy Finkelstein, and Jonathan Levin. 2010. "Beyond Testing: Empirical Models of Insurance Markets." *Annual Review of Economics* 2(1): 311–36.

Einav, Liran, Amy Finkelstein, Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen. 2013. "Selection on Moral Hazard in Health Insurance." *American Economic Review* 103(1): 178–219.

Einav, Liran, and Amy Finkelstein. 2011. "Selection in Insurance Markets: Theory and Empirics in Pictures." *Journal of Economic Perspectives* 25(1): 115–38.

Ericson, Keith M. Marzilli. 2014a. "Consumer Inertia and Firm Pricing in the Medicare Part D Prescription Drug Insurance Exchange." *American Economic Journal: Economic Policy* 6(1): 38–64.

Ericson, Keith M. Marzilli. 2014b. "When Consumers Do Not Make an Active Decision: Dynamic Default Rules and their Equilibrium Effects." NBER Working Paper 20127.

Ericson, Keith M. Marzilli, Jon Kingsdale, Tim Layton, and Adam Sacarny. 2017. "Nudging Leads Consumers in Colorado to Shop But Not Switch ACA Marketplace Plans." *Health Affairs* 36(2): 311–19.

Ericson, Keith Marzilli, Philipp Kircher, Johannes Spinnewijn, and Amanda Starc. 2015. "Inferring Risk Perceptions and Preferences using Choice from Insurance Menus: Theory and Evidence." NBER Working Paper 21797.

Ericson, Keith Marzilli, and Amanda Starc. 2015. "Measuring Consumer Valuation of Limited Provider Networks." *American Economic Review* 105(5): 115–19.

Ericson, Keith M. Marzilli, and Amanda Starc. 2016. "How Product Standardization Affects Choice: Evidence from the Massachusetts Health Insurance Exchange." *Journal of Health Economics* 50: 71–85.

Ericson, Keith M., and Justin Sydnor. 2017.

"Liquidity Constraints and the Demand for Insurance." Unpublished paper.

Gabaix, Xavier, and David Laibson. 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *Quarterly Journal of Economics* 121(2): 505–40.

Grubb, Michael D. 2015. "Overconfident Consumers in the Marketplace." *Journal of Economic Perspectives* 29(4): 9–36.

Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review* 103(7): 2643–82.

Handel, Ben, Igal Hendel, and Michael D. Whinston. 2015. "Equilibria in Health Exchanges: Adverse Selection versus Reclassification Risk." *Econometrica* 83(4): 1261–313.

Handel, Benjamin R., and Jonathan T. Kolstad. 2015. "Health Insurance for 'Humans': Information Frictions, Plan Choice, and Consumer Welfare." *American Economic Review* 105(8): 2449–500.

Handel, Benjamin R., Jonathan T. Kolstad, and Johannes Spinnewijn. 2015. "Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets." NBER Working Paper 21759.

Heidhues, Paul, Botond Köszegi, and Takeshi Murooka. 2017. "Inferior Products and Profitable Deception." *Review of Economic Studies* 84(1): 323–56.

Hoadley, Jack, Laura Summer, Jennifer Thompson, Elizabeth Hargrave, and Katie Merrell. 2007. *The Role of Beneficiary-Centered Assignment for Medicare Part D.* Washington, DC: Medicare Payment Advisory Commission.

Iyengar, Sheena S., and Emir Kamenica. 2010. "Choice Proliferation, Simplicity Seeking, and Asset Allocation." *Journal of Public Economics* 94(7–8): 530–39.

Johnson, Eric J., Ran Hassin, Tom Baker, Allison T. Bajger, and Galen Treuer. 2013. "Can Consumers Make Affordable Care Affordable? The Value of Choice Architecture." *PLoS ONE* 8(12).

Johnson, Eric J., John Hersehy, Jaqueline Meszaros, and Howard Kunreuther. 1993. "Framing, Probability Distortions, and Insurance Decisions." *Journal of Risk and Uncertainty* 7(1): 35–51.

Kamenica, Emir. 2008. "Contextual Inference in Markets: On the Informational Content of Product Lines." *American Economic Review* 98(5): 2127–49.

Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee C. Vermeulen, and Marian V. Wrobel. 2012. "Comparison Friction: Experimental Evidence from Medicare Drug Plans." *Quarterly Journal of Economics* 127(1): 199–235.

Loewenstein, George, Joelle Y. Friedman, Barbara McGill, Sarah Ahmad, Suzanne Linck, Stacey Sinkula, John Beshears, James J. Choi, Jonathan Kolstad, David Laibson, Bigitte C. Madrian, John A. List, and Kevin G. Volpp. 2013. "Consumers' Misunderstanding of Health Insurance." *Journal of Health Economics* 32(5): 850–62.

Nyman, John A. 1999. "The Value of Health Insurance: The Access Motive." *Journal of Health Economics* 18(2): 141–52.

Rosenthal, Jaime A., Xin Lu, and Peter Cram. 2013. "Availability of Consumer Prices from US Hospitals for a Common Surgical Procedure." *JAMA Internal Medicine* 173(6): 427–32.

Sinaiko, Anna D., and Richard A. Hirth. 2011. "Consumers, Health Insurance and Dominated Choices." *Journal of Health Economics* 30(2): 450–57.

Ubel, Peter A., David A. Comerford, and Eric Johnson. 2015. "Healthcare.gov 3.0—Behavioral Economics and Insurance Exchanges." *New England Journal of Medicine* 372(8): 695–98.

van de Ven, Wynand P. M. M., and Randall P. Ellis. 2000. "Risk Adjustment in Competitive Health Plan Markets." In *Handbook of Health Economics,* vol. 1A, edited by Anthony J. Culyer and Joseph P. Newhouse, 755–845. Amsterdam: Elsevier.

Wong, Charlene A., Daniel E. Polsky, Arthur T. Jones, Janet Weiner, Robert J. Town, and Tom Baker. 2016. "For Third Enrollment Period, Marketplaces Expand Decision Support Tools to Assist Consumers." *Health Affairs* 35(4): 680–87.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 73–102

From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application

Abhijit Banerjee, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton

Randomized controlled trials have been used in economics and other social sciences for decades. A short list of examples familiar to many economists would include the negative income tax experiments (Hausman and Wise 1985), the RAND Health Insurance Experiment (Newhouse 1993), the series of welfare reform experiments in the 1980s and 1990s (Manski and Garfinkel 1992), and work on education such as the Perry Pre-School Project and Project STAR

■ *Abhijit Banerjee is Ford Foundation International Professor of Economics and Director, Abdul Latif Jameel Poverty Action Lab, both at the Massachusetts Institute of Technology, Cambridge, Massachusetts. Rukmini Banerji is CEO of Pratham Education Foundation and Director of the ASER Centre, both in New Delhi, India. James Berry is Assistant Professor of Economics, University of Delaware, Newark, Delaware. Esther Duflo is Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics and Director, Abdul Latif Jameel Poverty Action Lab, both at the Massachusetts Institute of Technology, Cambridge, Massachusetts. Harini Kannan is a Senior Research Manager and Post-Doctoral Fellow, Shobhini Mukerji is the Executive Director of the South Asia regional center, and Marc Shotland is Associate Director of Training in the Research Group, all at various locations of the Abdul Latif Jameel Poverty Action Lab. Michael Walton is Senior Lecturer in Public Policy, Harvard Kennedy School, Cambridge, Massachusetts. Banerjee and Duflo are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts, and members of the Board of Directors, Bureau for Research and Economic Analysis of Development (BREAD). Their email addresses are banerjee@mit.edu, rukmini.banerji@pratham.org, jimberry@udel.edu, eduflo@mit.edu, harini.kannan@ifmr.ac.in, shobhini.mukerji@ifmr.ac.in, shotland@mit.edu, and michael_walton@hks.harvard.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.73 doi=10.1257/jep.31.4.73

(Schweinhart, Barnes, and Weikart 1993; Finn and Achilles 1990). Their use has accelerated dramatically in the past 10 to 15 years in academia, reflecting what Angrist and Pischke (2010) call "the credibility revolution." In terms of establishing causal claims, it is generally accepted within the discipline that randomized controlled trials are particularly credible from the point of view of internal validity (Athey and Imbens 2017). However, as critics have pointed out, this credibility applies to the interventions studied—at that time, on that population, implemented by the organization that was studied—but does not necessarily extend beyond. Some pilot studies these days are enormous, covering many millions of people (we will discuss one such study below). But in the more typical case, critics say, it is not at all clear that results from small "proof-of-concept" studies run by nongovernment organizations can or should be directly turned into recommendations for policies for implementation by governments on a large scale (for example, see Deaton 2010).

In this paper, we begin by exploring six main challenges in drawing conclusions from a localized randomized controlled trial about a policy implemented at scale: market equilibrium effects, spillovers, political reactions, context dependence, randomization or site-selection bias, and piloting bias (implementation challenges at scale). These challenges are widely recognized, and experimental evidence can often be brought to bear on them. We then turn to an example of an educational intervention called "Teaching at the Right Level" that successfully took the steps from a pilot operated by a nongovernment organization in a few slums to a policy implemented at scale by state governments in India (and in population terms, states in India are often larger than most countries in Europe). We will tell the story of how this occurred, and also how this program experienced and dealt with the six above-mentioned challenges.

While external validity of a randomized controlled trial cannot be taken for granted, is it far from unattainable. The journey from smaller-scale internal validity to larger-scale external validity is a process that involves trying to identify the underlying mechanisms, refining the intervention model based on the understanding of these mechanisms and other practical considerations, and often performing multiple iterations of experimentation.

From Proof of Concept to Scalable Policies: Six Challenges

In medical trials, efficacy studies are usually performed first in tightly controlled laboratory conditions. For the same reasons, it often makes sense to verify proof of concept of a new social program under ideal conditions—by finding a context and implementation partner where all the necessary steps for success are likely to be taken (for a formal justification of this argument, see Chassang, Padró i Miquel, and Snowberg 2012). However, the results of such a program tested on a small scale, while informative, are not necessarily a good predictor of what would happen if a similar policy were to be implemented on a large scale. Indeed, it is not uncommon that larger-scale studies fail to replicate results that had been established in small randomized controlled trials elsewhere. In this section, we consider six obstacles that can arise in drawing conclusions from small-scale experiments, especially when the proof of concept is being taken to a larger scale.

Market Equilibrium Effects

When an intervention is implemented at scale, it could change the nature of the market. A small experiment is in many cases consistent with a partial equilibrium analysis: all relative market prices can be assumed to stay constant. By contrast, a large experiment—such as a nationwide policy intervention—is likely to affect wages and the prices of nontradable goods such as land. These price changes might affect both the overall net benefit of the program as well as the identity of the beneficiaries.

For example, a program (like a scholarship) that increases education levels for a small group will only have a minimal effect on overall education levels for the population. But as Heckman, Lochner, and Taber (1998) argue, a large-scale education intervention that produces broad increases in educational attainment across an entire population may thereby decrease the overall return to education. Thus, the results of a small randomized controlled trial of a scholarship program (as in Duflo, Dupas, and Kremer 2017) would potentially be an overestimate of the impact of scholarships on earnings, if such a program were to be scaled up.

In other settings, ignoring the equilibrium effect can lead to underestimation of the overall benefits of a treatment. For example, an intervention that increases the income among some people could lead them to consume more: if part of this consumption is in the form of nontradable goods, this will have a multiplier effect, since those who are supplying those nontradable goods will also benefit. While a small experiment may not capture this effect, it could turn out to be a source of substantial social benefits in a large-scale implementation.

An illustration of the possible pitfalls of ignoring multiplier effect is the analysis of the potential impact of microcredit. Randomized controlled trials consistently find low impact of microcredit on beneficiaries (for a recent review, see Banerjee, Karlan, and Zinman 2015). These experiments are typically based on randomization across villages, neighborhoods, or individuals. But Buera, Kaboski, and Shin (2012) suggest that microcredit may have important general equilibrium effects, and it is possible that those effects operate on a broader scale than just the village. In a nonexperimental study, Breza and Kinnan (2016) examine the sudden collapse of microcredit in Andhra Pradesh, India, following a political backlash. Contrary to the results of the previous randomized studies, they find large negative effects of losing access to microcredit and argue that this was probably the consequence of the cutback in consumption resulting from the credit withdrawal on the rest of the economy. In other words, this is a case where the general equilibrium effect is likely to be much bigger than the effect on the direct beneficiaries.

Andrabi, Das, Ozyurt, and Singh (2017) describe another mechanism for why the general equilibrium effect may be very different from the partial equilibrium effect. In their experiment in Pakistan, in some villages, one randomly selected private school was given a grant to help improve quality. In other villages, all schools received it. The authors find very different effects on the treated schools in the two conditions. When only one school was treated, it improved its facilities at the margin and stole business from other private schools. When all schools were treated, they raised quality more substantially by investing in teachers and expanded capacity at the expense of public schools. The single-school experiment would have entirely missed this effect.

Recent research has taken this concern on board. One approach is to try to build a model to capture the various general equilibrium effects and calibrate it (as in Townsend 2010), making more- or less-heroic assumptions about the many parameters that need to be calibrated. Another approach, which has become popular, now that researchers are able to conduct larger experiments, is to design experiments to estimate those effects directly. At the most recent annual conference of the Bureau for Research and Economic Analysis of Development (the premier network for development economists) in May 2017, three of the eight papers presented described randomized controlled trials designed to assess the equilibrium impact of an intervention (Akram, Chowdhury, and Mobarak 2017; Andrabi et al. 2017; McKenzie and Puerto 2017). The typical design is a two-stage randomization procedure in which the treatment is randomly assigned at the market level in addition to the random assignment within a market. For example, the experiment of Crepon, Duflo, Gurgand, Rathelot, and Zamora (2013) varied the treatment density of a job placement assistance program in France within labor markets, in addition to random assignment of individuals within each market. The results show that placement assistance did benefit those assigned to receive it, but these effects were entirely undone by negative market-level impacts on untreated individuals. This result tempers the conclusion of a large literature on this type of intervention focusing on partial equilibrium effects, which tends to find that the program had significant positive effects (Card, Kluve, and Weber 2010). Muralidharan and Sundararaman (2015) adopt a similar design to evaluate a school voucher program in Andhra Pradesh, and in this case find no evidence of equilibrium effects coming into play.

A number of other experiments were designed to estimate just the full equilibrium effect, by conducting the randomization at the market level and focusing on market-level outcomes. Muralidharan, Niehaus, and Sukhtankar (2016) evaluate the rollout of a smart-card payments system for the National Rural Employment Guarantee Scheme, a workfare program in India. Randomization was conducted at the *mandal* (sub-district) level, allowing estimation of market-level effects across a large number of villages. The intervention increased take-up of the program, and the private sector wages increased in treatment *mandals* as a result. Several other papers estimate the impacts of transfer programs on village-level prices and wages (Cunha, De Giorgi, and Jayachandran 2011; Angelucci and De Giorgi 2009; Attanasio, Meghir, and Santiago 2011).

One potential challenge with the experimental identification of equilibrium effects is that it is not always obvious what the "market" is. For example, Akram, Chowdhury, and Mobarak (2017) evaluate an intervention in rural Bangladesh that provided financial support for temporary migrants and find large effects on the migrants and their households. Implementation was randomized at the village level, as well as within villages, to examine spillover on nonparticipants (which is one type of general equilibrium effect), but the more obvious equilibrium effect in this case seems to be what happens to wages in cities when lots of migrants show up in a city. To address that, the randomization needs to be done at the level of the recipient city. This is conceptually feasible but a different exercise altogether (which this team plans to undertake in future research as the program scales).

One other form of general equilibrium effect receives less attention in the literature but can turn out to be relevant. When a particular intervention is scaled up, more people will be needed to implement it. This may lead to an increase in their wages or in difficulties hiring them, which should be accounted for in the cost–benefit analysis of the program at scale. For example, Duflo, Dupas, and Kremer (2017) exploit the result of their scholarship experiment to calculate the cost per year of making an extra year of secondary school free in Ghana. But once the government decides to implement free secondary schools in Ghana (as Sackey 2017 reports that they have just promised to do), the government will need to hire a large number of secondary schoolteachers. Given the short supply of college-educated workers, this may not be feasible or may be much more expensive than accounted for in the Duflo, Dupas, and Kremer (2017) calculations. The extent to which this is a problem in practice depends on the nature of the intervention and the context. Luckily, it seems researchers tend to be biased towards evaluating programs that do have a chance to be implementable at scale without a significant increase in costs.[1] A more general point is that any evaluation of benefits needs to be coupled with an understanding of the costs if it is to be useful as guidance for policy decisions. The costs will generally be different in the scaled-up version of the program than in the evaluation. Costs may in fact be lower once the program becomes routine—or higher, as in the Ghana case. Fortunately, a more accurate estimate of large-scale costs can often be estimated by collecting costs from versions of the programs that have been implemented at scale elsewhere.

Spillover Effects

Many treatments have spillovers on neighboring units, which implies that those units are not ideal control groups. Some spillovers are related to the technology: For example, intestinal worms are contagious, so if a child is dewormed, this will affect her neighbor. If many of the children in a school are dewormed, this will also affect neighboring schools (Miguel and Kremer 2004). An intervention targeted to some children in a school may also benefit others in the school who were in the control group—perhaps through peer effects or through adjustments in teaching within the school. Other channels of spillover are informational: when a new technology

[1] Banerjee, Duflo, and Kremer (2017) provide some tentative evidence suggesting that researchers are actually good at identifying such interventions before the experiment is conducted.

is introduced (like a long-lasting insecticide-treated bed-net), the first people who are exposed to it may not take it up or use it properly. As more people experience the product, their friends and neighbors will learn about it and moreover, this may have reinforcement effect as neighbors teach each other how to use it better. For example, Dupas (2014) evaluates the impact of free long-lasting insecticide-treated bed-net distribution in Kenya. She finds that when randomly selected households received a highly subsidized bed net in an initial distribution, their neighbors had a higher willingness to pay for a net one year later, suggesting they were learning about the technology.

Economists have long been mindful of the possibility of such spillovers, and even small experiments can be designed to investigate whether they are present. For example, Miguel and Kremer (2004) took advantage of the fact that the number of treatment schools was much higher in some areas than others (just by chance), to estimate the positive spillovers from taking the deworming medicine on those who did not themselves take it. Duflo and Saez (2003) adopt a two-step experimental design to measure information spillovers in retirement savings decisions. But not all spillovers are easy to detect in pilot experiments: in some cases, they may be highly nonlinear. For example, there may need to be enough people using a bed-net before substantial health externalities kick in: Tarozzi et al. (2014) conduct a randomized evaluation of the impact of bed-nets where the randomization was performed at the household level, and find no positive effect, but because very few households in each village received a bed-net, this does not tell us what would happen if they all got (and used) one. Cohen and Dupas (2010) show that calculations on the cost–benefit of free bed-net distribution are highly sensitive to assumptions made about nonlinear spillovers. This is potentially important given that standard models of social learning often embody important nonlinearities or "tipping points."

Political Reactions

Political reactions, including either resistance to or support for a program, may vary as programs scale up. Corrupt officials may be more likely to become interested in stealing from programs once they reach a certain size (Deaton 2010). For example, Kenya's national school-based deworming program, a scale-up based on the results of previous randomized controlled trials, began in 2009 but was halted for several years due to a corruption scandal. The funds for the program had been pooled with other funds destined for primary education spending, and allegations of misappropriation in those pooled funds caused donors to cut off education aid—including support for the deworming program. The program ultimately restarted in 2012 (Sandefur 2011; Evidence Action 2014).

Political resistance to or support for a program may build up when the program reaches a sufficient scale. Banerjee, Duflo, Imbert, Mathew, and Pande (2017) provide an example of political backlash leading to the demise of a promising program in the state of Bihar, India, to reduce corruption in a government workfare program. Even though the experiment was a pilot, it included almost 3,000 villages representing an overall population of 33 million people. The village officials and their immediate superiors at the block- or district-level were dead set against the anticorruption intervention for the obvious reason that it threatened their rents. These officials were successful in lobbying the state government, and the intervention was cancelled, in part because a reduction in corruption was only demonstrated much later.[2]

This pilot of the anticorruption program was much larger than the typical proof-of-concept study, and as a result, the group it reached was large enough to have political influence. A smaller pilot might have had a less-difficult time, but this political counterreaction would have been missed. However, in other cases, pilots can be more vulnerable than scaled-up interventions: because they are subject to review, it is easy to shut them down.

Context Dependence

Evaluations are typically conducted in a few (carefully chosen) locations, with specific organizations. Would results extend in a different setting (even within the same country)? Would the results depend on some observed or unobserved characteristics of the location where the intervention was carried out?

Replication of experiments allows researchers to understand context dependence of programs. Systematic reviews, like those done by the Cochrane Collaboration for health care interventions, collect evidence from replications. Cochrane reviews have been compiled on topics such as water quality interventions (Clasen et al. 2015), mosquito nets (Lengeler 2004), and deworming of schoolchildren (Taylor-Robinson, Maayan, Soares-Weiser, Donegan, and Garner 2015). In economics, the International Initiative for Impact Evaluation maintains a database of systematic reviews of impact evaluations in developing countries that contains more than 300 studies (International Initiative for Impact Evaluation 2017). Several recent studies and journal volumes compile the results from multiple interventions in the same publication. For example, the January 2015 issue of the *American Economic Journal: Applied Economics* was devoted to six experimental studies of microfinance. Although these studies were not conducted in coordination, the overall conclusions are quite consistent across studies: the interventions showed modest increases in business activity but very little evidence of increases in consumption (Banerjee, Karlan, and Zinman 2015). The development of the American Economic Association's registry of randomized trials and public archiving of data, and the greater popularity of systematic meta-analysis methods within economics, should allow similar analyses across many more research questions.[3]

[2] There was, however, an interesting postscript: The results—which came out after the pilot was cancelled in Bihar—indicated a significant decline in rent-seeking and the wealth of public program officials. The anticorruption program was then extended to the same workfare program in all of India (with an explicit reference to the experimental results), and there are discussions to extend it to other government transfer programs.

[3] McEwan (2015) is another example of meta-analysis. He analyzes the results of 77 randomized controlled trials of school-based interventions in developing countries that examine impacts on child learning. While there is some degree of heterogeneity across studies, he is able to classify types of interventions that are consistently most effective based on his random-effects model.

However, to aggregate effects across studies, one has to start from some assumption about the potential distribution of treatment effects (Banerjee, Chassang, and Snowberg 2017). In the economics literature, this is often done without a formal analytical framework, which can lead to misleading results. For example, Pritchett and Sandefur (2015) argue that context-dependence is potentially very important, and that the magnitude of differences in treatment effects across contexts may be larger than the magnitude of the bias generated from program evaluation using retrospective data. They illustrate their point with data from the six randomized controlled trials of microcredit mentioned above. However, as pointed out by Meager (2016), Pritchett and Sandefur's measure of dispersion grossly overstates heterogeneity by conflating sampling variation with true underlying heterogeneity. Meager applies to the same data a Bayesian hierarchical model popularized by Rubin (1981), which assumes that (true) treatment effects in each site are drawn randomly from a normal distribution, and then estimated with error, and finds remarkably homogenous results for the mean treatment effect.

However, once we admit the need for a prior for aggregating results, there is no reason to stick to purely statistical approaches. An alternative is to use the existing evidence to build a theory, which tries to account for why some experiments succeed and others fail—rather than just tallying all the experiments and letting the failures cancel out the successes. The theory can then offer predictions that could be tested in future experiments, or which can feed into the design of a scaled-up intervention. For example, Kremer and Glennerster (2011) consider a range of randomized controlled trials on how price sensitivity affects take-up of preventive health products. They propose a number of alternative theories featuring liquidity constraints, lack of information, nonmonetary costs, or behavioral biases (such as present bias and limited attention). Dupas and Miguel (2017) provide an excellent summary of the evidence from randomized controlled trials on this point and argue that the subsequent evidence supports some aspects of the Kremer–Glennerster framework and rejects others. The point here is that many of those subsequent experiments were designed precisely with the Kremer–Glennerster framework in mind—effectively testing their conjectures—which makes them much more informative.

Randomization or Site-Selection Bias

Organizations or individuals who agree to participate in an early experiment may be different from the rest of the population, which Heckman (1992) calls randomization bias. There are three different possible sources for this concern.

First, organizations (and even individuals within governments) who agree to participate in randomized controlled trials are often exceptional. Glennerster (2017) lists the characteristics of a good partner to work with for a randomized controlled trial, and many organizations in developing countries do not meet the criteria. For example, the organization must be able to organize and implement the randomized implementation, providing relatively uniform implementation in the treatment group while not contaminating the control group. Senior staff must be open to the possibility of the program not working and be willing to have these results publicized. Even within government, willing partners are often particularly competent and motivated bureaucrats. Even when an intervention is not "gold-plated," organizations of individuals with these capabilities may find larger effect sizes than a large-scale program run by a less-stellar organization.[4] This is different from the general equilibrium point made above—even when the personnel to carry out the intervention at scale exists, the key constraint may be that of management capacity, and the difficulty of implementing changes at scale.

Second, a well-understood problem arises when individuals select into treatment. If treatment effects are heterogeneous across these groups, and those who are more likely to benefit are also more likely to be treated, then the estimated effect from the randomized controlled trial applies to compliers (those that respond to treatment), and may not apply to a broader population (Imbens and Angrist 1994). However, randomized controlled trials can be designed to enhance the external validity of experiments when respondents select themselves into treatment (for the theory, see Chassang, Padró i Miquel, and Snowberg 2012; for an application, see Berry, Fischer, and Guiteras 2015).

Third, site-selection bias arises because an organization chooses a location or a subgroup where effects are particularly large. This choice could be for legitimate reasons: nongovernmental organizations have limited resources and will try to work where they think their impact is the greatest, so they go to those areas first. In addition, both the organizations and the researchers, knowing that they are subject to an evaluation, have incentives to choose a site where the program is more likely to work. Organizations who take the trouble to participate in a randomized controlled trial would rather demonstrate success. Furthermore, if researchers anticipate that a study finding significant results is more likely to be published, they may design their studies accordingly. An illustrative case is that of Banerjee, Barnhardt, and Duflo (2015), who find no impact on anemia of free iron-fortified salt, in contrast with previous randomized controlled trials which led to the approval of the product for general marketing. And one reason is that the previous studies targeted adolescent women—and in fact Banerjee, Barnhardt, and Duflo (2015) find substantial treatment effects for that group but not for the average person. Yet fortified salt was approved for sales and distribution to the general population based on the group-specific results.

Several recent papers examine issues of randomization bias across large numbers of studies. Vivalt (2016) compiles data from over 400 randomized controlled trials and examines the relationship between effect size and study characteristics. Studies evaluating interventions run by nongovernment organizations or by researchers tend to find higher effects than randomized controlled trials run with governments, as do studies with smaller sample sizes. Allcott (2015) presents

[4] Allcott (2015) compares microfinance institutions that have partnered in recent randomized controlled trials with a global database of microfinance institutions and finds that partner institutions are older, larger, and have portfolios with lower default risk compared with nonpartner institutions.

the results of 111 randomized controlled trials of the Opower program in which households are presented with information on energy conservation and energy consumption of neighbors. He finds that the first ten evaluations of the intervention show larger effects on energy conservation than the subsequent evaluations and argues that this finding is attributable to differences in both partner utilities and study populations. Blair, Iyengar, and Shapiro (2013) examine the distribution of randomized controlled trials across countries and find that such trials are disproportionally conducted in countries with democratic governments.

Piloting Bias/Implementation Challenges

A large-scale program will inevitably be run by a large-scale bureaucracy. The intense monitoring that is possible in a pilot may no longer be feasible when that happens, or may require a special effort. For example, school reform often requires buy-in from teachers and school principals to be effective. The Coalition for Evidence-Based Policy (2013) reviewed 90 evaluations of US educational interventions commissioned by the Institute of Educational Studies, the research arm of the US Department of Education. They found that lack of implementation by the teachers was a major constraint and one important reason why 79 of 90 these interventions did not have positive effects. Interestingly, these interventions were themselves often quite small scale, despite being scale-ups of other even smaller studies.

Studies rarely document implementation challenges in great detail, but there are some examples. Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur (2015) replicate an intervention in Kenya first evaluated in Duflo, Dupas, and Kremer (2011, 2015), in which a nongovernment organization gave grants to primary school parent–teacher associations to hire extra teachers in order to reduce class sizes. The original Duflo, Dupas, and Kremer (2011, 2015) intervention resulted in significant increases in test scores. Bold et al. (2015) evaluate two versions of the program: one run by a nongovernment organization, which produced very similar results to the Duflo, Dupas, and Kremer (2011, 2015) evaluation, and a government-run version, which did not produce significant gains. Analysis of process data finds that government implementation was substantially weaker: the government was less successful in hiring teachers, monitored the teachers less closely, and was more likely to delay salary payments. The authors also suggest that political reactions—particularly the unionizing of the government contract teachers—could have also dampened the effects of the government-led implementation.

A number of studies have found differences between implementation by nongovernment organizations and governments. Barrera-Osorio and Linden (2009) evaluate a program in Colombia in which computers were integrated into the school language curriculum. In contrast with a previous intervention led by a nongovernment organization in India (Banerjee, Cole, Duflo, and Linden 2007), the authors find negligible effects of the program on learning, which they attribute to the failure of the teachers. Banerjee, Duflo, and Glennerster (2008) report on an experiment where incentives were provided for verified attendance in government health clinics in India. Although a similar incentive scheme had previously been proven effective when implemented in education centers run by a nongovernment organization in the same area (Duflo, Hanna, and Ryan 2012), there were no long-term effects on attendance in government health centers due to staff and supervisors exploiting loopholes in the verification system. Banerjee, Chattopadhyay, Duflo, Keniston, and Singh (2014), working with the police leadership in Rajasthan, India, to improve the attitudes of the police towards the public, find that the reforms that required the collaboration of station heads were never implemented.

In an interesting counterexample, Banerjee, Hanna, Kyle, Olken, and Sumarto (2016) study the distribution of identity cards entitling families to claim rice subsidies in Indonesia. In the pilot, the Indonesian government was meant to distribute cards containing information on the rice subsidy program to beneficiary households, but only 30 percent of targeted households received these cards. When the program was scaled up to the whole country, the mechanism for sending cards was changed and almost everybody did finally get a card. In this case, the government was less effective at running a pilot program and more effective with full implementation. This dynamic may be more general than one might at first expect: pilots face their own challenge because they impose new ad hoc procedures on top of an existing system. Once a bureaucracy takes over and puts a routine in place, implementation can become more systematic.

As the discussion in this section has emphasized, the issue of how to travel from evidence at proof-of-concept level to a scaled-up version cannot be settled in the abstract. The issue of context-dependence needs to be addressed through replications, ideally guided by theory. General equilibrium and spillover effects can be addressed by incorporating estimation of these effects into study designs, or by conducting large-scale experiments where the equilibrium plays out. Randomization and piloting bias can be addressed by trying out the programs on a sufficient scale with the government that will eventually implement it, documenting success and failure, and moving from there.

In the next section, we illustrate how these issues play out in practice by describing the long journey from the original concept of a specific teaching intervention in India, through its multiple variants, to the eventual design and evaluation of two "large-scale" successful incarnations implemented in government schools that are now in the process of being scaled up in other government systems.

A Successful Scale-up: Teaching at the Right Level

In India, as in many developing countries, teachers are expected to teach a demanding curriculum, regardless of the level of preparation of the children. As a result, children who get lost in early grades may never catch up (Muralidharan 2017). In response, Pratham, an Indian nongovernmental organization, designed a deceptively simple approach, which has come to be called "teaching at the right level" (TaRL). Pratham credits literacy expert Abul Khair Jalaluddin for developing the first incarnation of the pedagogy (Banerji, Chavan, and Rane 2004). The basic idea is to group children, for some period of the day or part of the school year, not according to their age, but according to what they know—for example, by splitting the class, organizing supplemental sessions, or reorganizing children by level—and match the teaching to the level of the students.

From Bombay Slums to 33 Million Children

The partnership between researchers and Pratham started with a "proof of concept" randomized controlled trial of Pratham's *Balsakhi* Program in the cities of Vadodara and Mumbai, conducted in 2001–2004 (Banerjee et al. 2007). In this program, third- and fourth-grade students identified as "lagging behind" by their teachers were removed from class for two hours per day, during which they were taught remedial language and math skills by paid community members (*balsakhis*) hired and trained by Pratham. Their learning levels (measured by first- and second-grade-level tests of basic math and literacy) increased by 0.28 standard deviations.

Pratham next took this approach from the relatively prosperous urban centers in West India into rural areas, and in particular into rural areas of Northern India. By 2004, Pratham worked in 30 cities and nine rural districts (Banerji, Chavan, and Rane 2004). As Pratham increased the scale of its program, the key principle of teaching children at the appropriate level remained, but one core feature of its model changed for the sake of financial viability: they were forced to rely largely on volunteers rather than paid teachers. These volunteers worked outside the school running their own learning-improvement classes and were much less closely supervised after the initial two-week training. To facilitate this change, the pedagogy became more structured and more formal, with an emphasis on frequent testing. Whether the intervention would continue to work well with a new programmatic design, organizational change, and new contexts was an open question. A new randomized evaluation was therefore launched to test the volunteer-based model in the much more challenging context of rural North India.

This second randomized controlled trial was conducted in rural Jaunpur district of Uttar Pradesh in 2005–2006: this was a test of the volunteer-led, out-of-school model Pratham called "Learning to Read." The results were very positive: after accounting for the fraction of students who attended, treatment-on-the-treated estimates show that attending the classes made children who could read nothing at baseline 60 percent more likely to progress to letters at endline. For children who could read letters at baseline, the classes resulted in a 26 percent higher likelihood of reading a story, the highest level on the test, at endline (Banerjee, Banerji, Duflo, Glennerster, and Khemani 2010).

This second study established that the pedagogical idea behind the *balsakhi* program could survive the change in context and program design, but it also revealed new challenges. There was substantial attrition among the volunteers, and many classes ended prematurely. Also, because the program targeted children outside of school, take-up was far from universal. Only 17 percent of eligible students were treated. Most concerning, the treated students did not come disproportionately from the bottom end of the distribution—those who were unable to recognize letters or numbers, and who needed it the most.

Nevertheless, in 2007, building on the success of the Learning to Read intervention, Pratham rolled out its flagship "Read India" Program. Within two years, the program reached over 33 million children. To reach all of the children who needed remedial education, Pratham started collaborating with state governments in running the program. But the efficacy of the government's implementation of the program was again an open question. In the remainder of this section, we present the results of the series of experiments aimed to develop a scalable policy in government schools based on the Pratham methodology.

A First Attempt to Scale-Up with Government

Starting in 2008, Pratham and the Abdul Latif Jameel Poverty Action Lab, commonly known as J-PAL, embarked on a series of new evaluations to test Pratham's approach when integrated with the government school system. Two randomized controlled trials were conducted in the states of Bihar and Uttarakhand over the two school years of 2008–2009 and 2009–2010. Although the evaluation covered only a few hundred schools, it was embedded in a full scale-up effort: as of June 2009, the Read India program was being run in approximately 40,000 schools in Bihar and 12,000 schools in Uttarakhand, representing the majority of schools in each state (Kapur and Icaza 2010).

In the first intervention (evaluated only in Bihar during June 2008), remedial instruction was provided during a one-month summer camp, run in school buildings by government schoolteachers. Pratham provided materials and training for these teachers and also trained volunteers who supported teachers in the classroom. The government schoolteachers were paid extra by the government for their service over the summer period.

The other three interventions were conducted during the school year. The first model (evaluated only in Bihar) involved the distribution of Pratham materials with no additional training or support. The second variant of the intervention included materials, as well as training of teachers in Pratham methodology and monitoring by Pratham staff. Teachers were trained to improve teaching at all levels through better targeting and more engaging instruction. The third and most-intensive intervention included materials, training, and volunteer support. The volunteer part of the materials-training-volunteers intervention in Bihar was a replication of the successful Learning-to-Read model evaluated in Jaunpur, in which the volunteers conducted out-of-school learning camps that focused on remedial instruction for students directed to them by teachers. As part of the materials-training-volunteers intervention in Uttarakhand, however, volunteers worked in schools and were meant to support the teachers. In both states, about 40 villages were randomly assigned to each treatment group.

The main outcome measures in the Bihar and Uttarakhand evaluations, as with the others presented later in this section, are performance on simple language and math tests developed by the ASER Centre, Pratham's research arm. In language, children are classified based on whether they can recognize letters, read words, read a paragraph, or read a short story. In math, the levels include single-digit and double-digit number recognition, double-digit subtraction, and division of a double-digit number by a single digit. In the results that follow, we assign an integer score between zero and four based on the highest level the child can perform.

To complement the randomized controlled trial, we collected extensive process data and partnered with political scientists who, through interviews, collected details of the relationship between Pratham and the government.[5]

The language and math results of the evaluations in Bihar and Uttarakhand (presented in Table 1) were striking and mostly disappointing. The materials-alone and materials-plus-training interventions had no effect in either Bihar or Uttarakhand. The materials-training-volunteers treatment in Uttarakhand had no detectible impact either. However, in the materials-training-volunteers results in Bihar, we found a significant impact on reading and math scores, quite comparable to the earlier Jaunpur results. Since the materials-plus-training intervention seemed to make no difference, we interpret this as further evidence that, like in Jaunpur, Pratham's pedagogical approach also worked in this new context when implemented by volunteers outside school hours. However, when the volunteers were made part of the in-school team, as in Uttarakhand, they either became absorbed as regular teachers, teaching the curriculum rather than implementing Pratham's pedagogy, or did not show up at all. The failure of schools to utilize the volunteers as intended may be why the Uttarakhand intervention did not work.

At this point, one might have been tempted to assume that the key distinction is between government teachers and private volunteers as implementers (along the lines of Bold et al. 2015). However, this interpretation is belied by the Bihar summer camp results, which show significant gains in language and math despite being implemented by the government schoolteachers. Based on the fraction of children who attended the summer camp, the treatment-on-the-treated results show that the camp improved reading scores by about 0.5 levels in just a few weeks. This finding suggests the possibility that government teachers were in fact able to deliver remedial education if they did focus on it, but this did not happen during the school year.

Some process data and the qualitative information bolster this interpretation. Table 2 (panels A and B) shows selected process measures across the two experiments. The situations were very different in the two states (Kapur and Icaza 2010). In Bihar, Pratham had an excellent relationship with the educational bureaucracy, from the top rungs down to district- and block-level administrators. As a result, the basic inputs of the program were effectively delivered: over 80 percent of the

[5] Banerjee, Banerji et al. (2016) provide more details on the evaluation design and the results of these two experiments as well as the two further experiments described in the next subsection. Kapur and Icaza (2010) provide a detailed account of the working of the partnership between Pratham and the government at various levels in Bihar and Uttarakhand. Sharma and Deshpande (2010) present a qualitative study based on interviews with parents, teachers, and immediate supervisors of the teachers.

Table 1

Language and Math Results—Bihar and Uttarakhand

	Test Score (0–4)	
	Language	*Math*
A. Bihar—Summer Camp		
Treatment	0.12**	0.085*
	(0.059)	(0.050)
Control group mean	2.2	2.1
Observations	2,839	2,838
B. Bihar—School Year		
Materials	0.027	0.051
	(0.061)	(0.051)
Materials, Training	0.064	0.017
	(0.059)	(0.049)
Materials, Training, Volunteer Support	0.20***	0.13***
	(0.054)	(0.046)
Control group mean	1.8	1.8
Observations	6,490	6,490
C. Uttarakhand		
Materials, Training	0.030	0.038
	(0.053)	(0.042)
Materials, Training, Volunteer Support	–0.012	0.0091
	(0.044)	(0.043)
Control group mean	2.2	2.0
Observations	5,645	5,646

Note: Pratham and the Abdul Latif Jameel Poverty Action Lab (J-PAL) conducted randomized controlled trials testing the Pratham pedagogical approach when integrated with the government school system in the states of Bihar and Uttarakhand. In the first intervention (Panel A), remedial instruction was provided during a one-month summer camp run in school buildings by government schoolteachers. Pratham provided materials and training for these teachers and also trained volunteers who supported teachers in the classroom. The other three interventions (Panels B and C) were conducted during the school year: The first model (evaluated only in Bihar) involved the distribution of Pratham materials with no additional training or support. The second intervention included materials, as well as training of teachers in Pratham methodology and monitoring by Pratham staff. The third and most-intensive intervention included materials, training, and volunteer support. In Bihar, the volunteers conducted out-of-school learning camps (during the school year) that focused on remedial instruction for students directed to them by teachers. As part of the materials-training-volunteers intervention in Uttarakhand, however, volunteers worked in schools and were meant to support the teachers. Standard errors in parentheses (clustered at the level of randomization). Test scores are computed on an integer scale from 0 (nothing) to 4 (can read a story) in language and 0 (nothing) to 4 (can perform division) in math. Regressions control for baseline scores as well as gender age, and grade at baseline.

*, **, and *** mean significance at the 10, 5, and 1 percent levels, respectively.

Table 2
Selected Process Results

	Percent of schools (# of schools in parentheses)		
	Teachers trained	*Pratham materials used*	*Classes grouped by ability*
A. Bihar—School Year			
Control	1.4 (63)	0.8 (59)	0.0 (60)
Materials	5.6 (64)	33.6 (63)	1.6 (63)
Materials, Training	84.4 (66)	62.5 (64)	3.8 (65)
Materials, Training, Volunteer Support	84.7 (68)	69.2 (65)	0.0 (65)
B. Uttarakhand			
Control	18.9 (41)	3.8 (39)	14.1 (39)
Materials, Training	29.4 (40)	26.3 (40)	10.0 (40)
Materials, Training, Volunteer Support	53.8 (39)	38.5 (39)	5.1 (39)
C. Haryana			
Control	0.5 (200)	0.5 (199)	0.0 (199)
Teaching at the Right Level (During TaRL classes)	94.7 (126)	81.0 (126)	91.3 (126)
Teaching at the Right Level (Other times)	94.0 (155)	1.3 (149)	2.0 (149)
D. Uttar Pradesh			
Control		0.0 (108)	
Materials		30.7 (111)	
Four 10-Day Camps	89.9 (122)	90.6 (122)	79.4 (122)
Two 20-Day Camps	87.8 (120)	84.2 (120)	83.5 (120)

Note: The Bihar school-year and Uttarakhand evaluations consisted of three interventions. The first model (evaluated only in Bihar) involved the distribution of Pratham materials with no additional training or support. The second intervention included materials, as well as training of teachers in Pratham methodology and monitoring by Pratham staff. The third and most-intensive intervention included materials, training, and volunteer support. In the Haryana intervention, efforts were made to promote organizational buy-in, including the creation of a system of academic leaders within government to guide and supervise the teachers as they implemented the Pratham methodology; the program was implemented during a dedicated hour of the school day; and all children in grades 3–5 were reassigned to achievement-based groups and physically moved from their grade-based classrooms to classrooms based on levels. The Uttar Pradesh interventions used the in-school "learning camps" model, with learning camps administered primarily by Pratham volunteers and staff during school hours when regular teaching was temporarily suspended. When a school was observed multiple times, the average is used for that school.

teachers were trained, they received the material, and they used the materials more than half the time. In Uttarakhand, key state personnel changed just before the evaluation period and several times afterwards. There was infighting within the educational bureaucracy, and strikes by teachers and their supervisors (unrelated to the program). The local Pratham staff were demoralized and turned over rapidly. As a result, only between 29 and 54 percent of teachers got trained (for only three days each), and only one-third of the schools used the materials, which they received very late. In many cases, there was either no volunteer or no teacher in the school during the monitoring visits.

The process data also show that the key component of Pratham's approach, the focus on teaching at the children's level, were generally not implemented in schools in either state. One consistent lesson of the earlier studies is that the pedagogy worked when children grouped in a way that the teaching could be targeted to the deficiencies in their training. This happened systematically in the volunteer classes, and this also happened in the Bihar summer camps because their express purpose was to focus on remedial skills. But in regular classes in Bihar, for example, only between 0 and 4 percent of the classes were observed to be grouped by levels.

Thus, the challenge for Pratham was how to get government teachers to not only use materials and deliver the pedagogy, but also how to incorporate the targeted teaching aspect of the model into the regular school day. As we see from Bihar, independent training by Pratham by itself was insufficient to get teachers to do this, even with consistent support by the bureaucracy. The summer camp in Bihar, however, produced a large effect. Therefore, it is possible for governments to "teach at the right level." Why don't they do so during the regular school day?

In *Poor Economics*, Banerjee and Duflo (2011) discuss this resistance and point out that Teaching at the Right Level is not even being implemented in private schools, which are subject to a high level of competition and are certainly not lacking in incentives, despite the fact that most children in those schools are also not at grade level. They propose the hypothesis that teachers and parents must put more weight on covering the grade-level curriculum than on making sure that everyone has strong basic skills. Similarly, the qualitative interviews conducted in the Read India scale-up experiments revealed that teachers believed the methods proposed by Pratham were effective and materials were interesting, but they did not think that adopting them was a part of their core responsibility. Paraphrasing the teachers they interviewed in Bihar, Sharma and Deshpande (2010) write: "[T]he materials are good in terms of language and content. The language is simple and the content is relevant. … However, teaching with these materials require patience and time. So they do not use them regularly as they also have to complete the syllabus."

If this hypothesis is correct, it suggests two main strategies: either convince the teachers to take Teaching at the Right Level more seriously by working with their superiors to build it into their mission; or cut out the teachers altogether and implement a volunteer-style intervention, but do it in the school during school hours, so as to capture the entire class rather than just those who opt to show up for special after-school or summer classes. These ideas guided the design of the next two interventions.

Getting Teachers to Take the Intervention Seriously

In 2012–2013, Pratham, in partnership with the Haryana State Department of Education, adopted new strategies to embed the Teaching at the Right Level approach more strongly into the core of teaching/learning in primary schools; in particular, they were interested in how to get teachers to view it as a "core responsibility."

Several methods were used to promote organizational buy-in. First, all efforts were made to emphasize that the program was fully supported and implemented by the government of Haryana, rather than an external entity. In the earlier experiment in Bihar and Uttarakhand, despite the fact that this was a government initiative, teachers did not perceive it as such, in part because they rarely got that message from their immediate supervisors and the responsibility of monitoring the teachers was left to Pratham staff. In Haryana, a system of academic leaders within the government was created to guide and supervise teachers as they implemented the Pratham methodology. As part of the interventions, Pratham gave four days of training and field practice to "Associate Block Resources Coordinators," who were then placed in groups of three in actual schools for a period of 15–20 days to carry out daily classes and field-test the Pratham methodology of grouping by level and providing level-appropriate instruction. Once the practice period was over, these Coordinators, assisted by Pratham staff, in turn trained the teachers that were in their jurisdiction. Second, the program was implemented during a dedicated hour during the school day. Beginning in the 2011–2012 school year, the government of Haryana mandated that all schools add an extra hour of instruction to the school day. In regular schools, the normal school day was just longer. Within Teaching at the Right Level schools, the extra hour was to be used for class reorganization and teaching remedial Hindi classes using the Pratham curriculum. This change sent a signal that the intervention was government-mandated, broke the status quo inertia of routinely following the curriculum, and made it easier to observe compliance. Third, during the extra hour, in Teaching at the Right Level schools, all children in grades 3–5 were reassigned to achievement-based groups and physically moved from their grade-based classrooms to classrooms based on levels, as determined by a baseline assessment done by teachers and the coordinators. Once classes were restructured into these level-based groups, teachers were allocated to the groups for instruction. This removed teacher discretion on whether to group children by achievement.

This new version of the program was evaluated in the school year 2012–2013 in 400 schools, out of which 200 received the program. The results were this time positive, as shown in Table 3: Hindi test scores increased by 0.2 levels. This intervention did not target math.

Because the objective of this study was to develop a model that could be adopted at scale, we also incorporated extensive process monitoring into our study design,

Table 3
Language and Math Results—Haryana and Pradesh

	Test Score (0–4)	
	Language	*Math*
A. Haryana		
Teaching at the Right Level	0.20***	−0.0069
	(0.023)	(0.019)
Control group mean	2.4	2.2
Observations	11,963	11,962
B. Uttar Pradesh		
Materials	0.045	0.053**
	(0.030)	(0.027)
Four 10-Day Camps	0.95***	0.81***
	(0.030)	(0.028)
Two 20-Day Camps	0.82***	0.73***
	(0.031)	(0.029)
Control group mean	1.5	1.7
Observations	17,254	17,265

Note: In the Haryana intervention, efforts were made to promote organizational buy-in, including the creation of a system of academic leaders within government to guide and supervise the teachers as they implemented the Pratham methodology; the program was implemented during a dedicated hour of the school day; and all children in grades 3–5 were reassigned to achievement-based groups and physically moved from their grade-based classrooms to classrooms based on levels. The Uttar Pradesh interventions used the in-school "learning camps" model, with learning camps administered primarily by Pratham volunteers and staff during school hours when regular teaching was temporarily suspended. Standard errors are in parentheses (clustered at the level of randomization). Test scores are computed on an integer scale from 0 (nothing) to 4 (can read a story in language, and 0 (nothing) to 4 (can perform division) in math. Regressions control for baseline test scores, as well as gender, age, and grade at baseline.
*, **, and *** mean significance at the 10, 5, and 1 percent levels, respectively.

including regular surprise visits to the schools. The third panel of Table 2 shows that about 95 percent of teachers in the treatment group attended training, compared with virtually no teachers in the control group. Most importantly, grouping by ability was also successful in Haryana, where it had largely failed in Bihar and Uttarakhand: over 90 percent of schools were grouped by learning levels during Teaching at the Right Level classes. In addition, teachers in Haryana used Pratham materials in 81 percent of Teaching at the Right Level classes, whereas much lower rates were observed in Bihar and Uttarakhand. Interviews with teachers, headmasters, and department administration suggested that the monitoring and mentoring role played by Associate Block Resource Coordinators was critical. Indeed, 80 percent of schools reported a visit from a Coordinator in the previous 30 days. Of those who reported a visit, 75 percent said that the Coordinator spent over an hour in the school, and 95 percent said that the Coordinator observed a class in progress.

Using the Schools, But Not the Teachers: In-School Learning Camps

In areas where the teaching culture is very weak, it may be too difficult or costly to involve the teachers in this alternative pedagogy. Instead, it may make sense to use an outside team, which can sidestep the teachers but still take advantage of the school infrastructure and the fact that the children are already present at school. The risk in going down this path, as we had seen in Uttarakhand before, was that the volunteers would be absorbed by the system and not implement the desired pedogogy.

To address this issue, Pratham, with the permission of the district administration, developed the in-school "Learning Camps" model. Learning Camps are intensive bursts of teaching/learning activity using the Pratham methodology and administered primarily by Pratham volunteers and staff during school hours when regular teaching is temporarily suspended. These camps were confined to periods of 10 or 20 days each (and a total of 50 days a year). In that sense, they were more similar to the original volunteer Learning-to-Read model (where volunteers ran "sessions" of 2–3 months) than to previous in-school experiences, except that they were within school premises during school hours. On "camp" days, children from grades 3–5 were grouped according to their ability level and taught Hindi and math for about 1.5 hours each by Pratham staff and Pratham-trained local village volunteers.

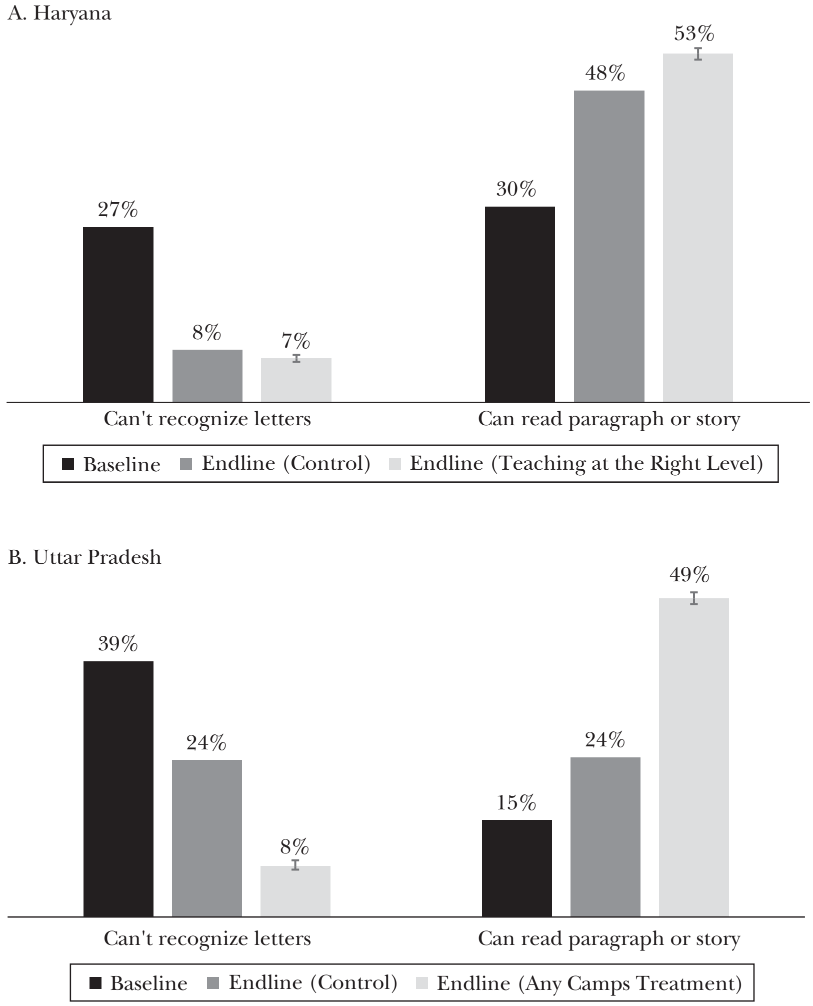
The model was tested in a randomized evaluation in Uttar Pradesh in the year 2013–2014: a sample of schools was selected and randomly divided into two camp treatment groups, a control group, and a materials-only intervention, with approximately 120 schools in each group. The learning camp intervention groups varied the length of the camp rounds, with one group receiving four 10-day rounds of camp, and the second receiving two 20-day rounds. Each intervention included an additional 10-day camp during the summer.

The two interventions had similar impacts, with test score gains of 0.7 to 1.0 levels, on average (as shown in Table 3).

It is useful to pause and consider how large these effects are. Figure 1 summarizes the results in Haryana and Uttar Pradesh. At baseline, 27 percent of children in Haryana could not even recognize letters, and 30 percent could read a paragraph or story (since the studies are randomized, control group and treatment group students are similar, so we present the pooled data for the baseline statistics). In Uttar Pradesh, 39 percent of the children could not recognize letters, and only 15 percent could read a paragraph or story. The difference between the two states was very large. At endline, there was little progress in the control group in Uttar Pradesh: 24 percent of children could still not recognize letters, and only 24 percent could read a paragraph or a story. But in the treatment group, only 8 percent could not recognize letters, and 49 percent could read a paragraph or a story. Thanks to these 50 days of intervention, they had fully caught up to the control group in Haryana

Figure 1

Distribution of Student Competency in Language: Baseline and Endline, by Treatment Status



Note: In the Haryana intervention, efforts were made to promote organizational buy-in, including the creation of a system of academic leaders within government to guide and supervise the teachers as they implemented the Pratham methodology; the program was implemented during a dedicated hour of the school day; and all children in grades 3–5 were reassigned to achievement-based groups and physically moved from their grade-based classrooms to classrooms based on levels. The Uttar Pradesh interventions used the in-school "learning camps" model, with learning camps administered primarily by Pratham volunteers and staff during school hours when regular teaching was temporarily suspended. Whiskers represent the 95 percent confidence interval between intervention and control groups.

(where, at endline, 48 percent could read a paragraph or story and 8 percent could not recognize letters), and had almost reached the level of the treated children in Haryana (where 53 percent of the treatment children could read a story). This reflects in part the abysmal performance of the school system in Uttar Pradesh. But the fact that the children actually reach the Haryana level in Uttar Pradesh also demonstrates the relative ease with which apparently daunting learning gaps can be closed.

As with the other evaluations, a systematic process-monitoring survey collected data on attendance, evidence of learning by "grouping," activities during "camp" sessions, teaching practices of volunteers, involvement of schoolteachers, and their perception of "camp" activities. There was strong adherence to key program components in Uttar Pradesh (Table 2, panel D). During camp days, use of Pratham materials was observed in over 80 percent of classes in both the 10-day and 20-day camp interventions. Critically, about 80 percent of classes in both treatments were observed to be grouped by achievement.

It took five randomized control trials and several years to traverse the distance from a concept to a policy that actually could be successful on a large scale. Today, the teacher-led "Haryana" model has been implemented in 107,921 schools across 13 states of India, reaching almost 5 million children. The in-school volunteer led model has been implemented in 4,210 schools across India, reaching over 200,000 children.

General Lessons about Scaling Up

Of the potential scale-up issues we identified, which ones turned out to be relevant in the Teaching at the Right Level example, and beyond that, what should be taken into consideration when designing an experiment with the view of ultimate scale up?

Equilibrium effects were not really a threat in this context, despite the size of the scale up in which the evaluations were embedded, since our outcome of interest was human capital, where there is no strategic interdependence. We did not explicitly study *spillovers* (although some could have occurred between teachers).

The interventions were repeatedly stress-tested for *context dependence* by moving the program from urban India to rural Uttar Pradesh, and then to Bihar, Uttarakhand, Haryana, and back to Uttar Pradesh again. This shows that the pedagogy that Pratham developed can improve basic learning levels in both reading and math across a variety of contexts. Moreover, there is supporting evidence from Ghana, where a successful replication of the Teaching at the Right Level approach was organized with the government (Innovations for Poverty Action 2015), and in Kenya, where students performed better when grouped by ability (Duflo, Dupas, and Kremer 2011). The results both in India, alongside results from similar tests worldwide, made it clear that many children clearly needed remedial education. In terms of understanding the magnitude of the need for remedial education, it is striking that the *intention-to-treat* effect of the camps program in Uttar Pradesh (estimated over all children in the schools) are as high as the *treatment-on-the-treated* results were in the early Learning to Read evaluation in Jaunpur (estimated only for those that attended after-school classes) (Banerjee et al. 2010). This finding suggest that the high early results were not driven by a subpopulation with high marginal returns in the original experiment.

Although political issues did arise in Uttarakhand, they were more due to turnover and infighting than to issues with Pratham, and there were no direct adverse *political reactions* to the program in its scaled-up version. However, such resistance could arise elsewhere. An attempt to pilot the project in the state of Tamil Nadu was not successful after the government officials displayed strong resistance. The backstory here is that Pratham has become such an important large player in the India educational scene that it cannot be seen as just another partner organization. In Tamil Nadu, Pratham was viewed as the group that had exposed the less-than-stellar performance of the state-run schools. Also, the Tamil Nadu government had their own pedagogical approach called "Activity Based Learning," which it was not keen to subject to scrutiny.

Although most of the attention on the challenge of scalability in the recent literature has been on the equilibrium effects and context dependence, it appears these issues were not particularly relevant here. The key obstacle to Teaching at the Right Level was the difficulty of implementing at scale. The first successes were clearly affected by a combination of *randomization bias* and *implementation challenges* when moving to scale. Pratham was one of the first organizations to partner with researchers to evaluate its programs (before J-PAL even existed), and may be rare in its combination of scale and purpose. It is conceivable that moving from Pratham to *any* other partner, not just the government, would have been difficult. Even within Pratham it was harder to find a good and enthusiastic team in Uttarakhand than in Bihar (Kapur and Icaza 2010). The fundamental challenge was to integrate the core concept of the program in the schools' day-to-day workflow, and this relied on organizational innovations beyond the basic concept of Teaching at the Right Level. In particular, achieving the alignment between pedagogy and initial learning levels required an explicit organizational effort to ensure that children were assessed, grouped, and actually taught at the right level. This did not occur automatically within the existing government school system but was achieved by articulating a narrow set of program goals, ensuring there was specific time for the program, and properly supervising implementation.

One way to interpret the series of Teaching at the Right Level studies is as a process of persuasion at scale: the experimental approach played not only an evaluation role but also an instrumental role in fostering acceptance of the policy by the government. In other words, we can see this effort as trying to answer the question: "How do you get a bureaucracy to make a common-sense change that has a very strong chance of being beneficial—like not totally ignoring students who have fallen behind and instead offering them a path to catching up?" From that perspective, the experimental approach is a little like opening a jammed door with a pry-bar. First you stick the bar in a little crack and get a little traction. Then you move to another location and get a little more traction. When you've got a little more purchase, you can jam in a bigger pry-bar and really tug hard. From this perspective, choosing where to pry, and finding organizations willing to experiment, and choosing other places to pry, and then finding government partners willing to participate, is all a way of prying open the door. At some point, the leverage is great enough that you can throw the door open. Sequential experimentation becomes a political economy tool for getting momentum for policy change.

More generally, what should practitioners and researchers keep in mind when designing randomized evaluations with a view to identifying policies that will work at scale? Perhaps the key point is to remember what small pilot experiments are good for and what they are not good for. Formulation of a successful large-scale public policy begins with the identification of a promising concept, which requires elaborating a model of the mechanism at play. Small-scale experiments can identify these concepts, both by pinpointing the sources of specific problems and testing approaches of dealing with them. Fully understanding the key mechanism behind successful interventions is often likely to take more than one experiment. In the case of education, early experiments by Glewwe, Kremer, and Moulin (2009) and the initial *balsakhi* results (Banerjee, Cole, Duflo, and Linden 2007) helped identify the core problem of the mismatch between what gets taught and what the children need to learn, but the results could have been explained by other factors (for example, in the *balsakhi* study, class size went down, and the instructor was younger and closer to the students). Based on this work, Duflo, Dupas, and Kremer (2011) designed an experiment that specifically investigated the potential of matching children by level, disentangling it from the effect of being assigned different kinds of teachers (for example, those who may be closer to the students and have better incentives), and found that it indeed matters. If the objective is to design or test a model, the researcher can ignore most of the concerns that we talked about in this paper. Something valuable will be learnt anyway. This is the equivalent of what is sometimes called "stage one" in venture capital investing.[6]

It would of course be dangerous to advocate a policy scale-up based exclusively on results from an investigation of this sort. The importance of all the issues we discussed earlier needs to be evaluated, which typically requires additional experimental work. We now consider them in turn (though not in the order in which we first discussed them).

Context dependence can be assessed by replications, either of the same experiments or of related experiments (that is, by experiments that test programs inspired by the same general idea). To assess whether a program is ready to be scaled up, or should be evaluated again first (perhaps starting on a smaller scale), policymakers

[6] This staged approach, inspired by venture capital funding process, is now explicitly adopted by some aid organizations, such as the US Agency for International Development's Development Ventures and the Global Innovation Fund (US Agency for International Development 2017; Global Innovation Fund 2017).

should ideally be able to rely on an aggregation of all the existing reliable evidence, randomized or not. Many are skeptical as to whether needed replications would be undertaken, but this skepticism does not seem warranted. With the proliferation of experiments in the last decade or so, there starts to be a critical mass of work on many key issues. Programs that appear to be particularly promising are more likely to be replicated. For example, Banerjee, Duflo et al. (2015) present six separate evaluations of an asset transfer program developed by the Bangladeshi Rural Advancement Committee that is being implemented around the world.

Once a program has passed the proof-of-concept test and is chosen to be scaled up, the next step is to develop an implementable large-scale policy version, and to subject it to a stage-two trial, meaning a larger trial that will confront and document the problems that the program would have at scale.[7] Designing this intervention typically requires combining an understanding of the mechanism underlying the concept with insight into how the particular government (or other large-scale implementer) works as an organization, which we have referred to elsewhere as getting "inside the machine" (Banerjee 2007) or as fixing the "plumbing" (Duflo 2017). For such trials to be informative, a number of critical design issues need to be addressed, which is what we turn to next.

To address the possibility of *randomization bias*, the organization that implements a stage-two trial must be the organization that will eventually implement it at scale, if it were to be scaled up. Within this organization, it must be implemented by the regular staff, not by a group of external experts. It also needs to be run in an area that is representative of where it would be scaled up eventually. For example, Muralidharan and Sundararaman (2015) randomly chose districts to run their market-level private voucher experiments.

For researchers, a strong temptation in a stage-two trial will be to do what it takes "to make it work," but the risk of *implementation challenges* means that it is important to think about how far to go in that direction. On the one hand, trial and error will be needed to embed any new intervention within an existing bureaucracy. Anything new is challenging, and at the beginning of a stage-two trial, considerable time needs to be spent to give the program a fair shot. On the other hand, if the research team embeds too much of its own staff and effort and ends up substituting for the organization, not enough will need to be learnt about where implementation problems might emerge. Our suggestion is to pilot a potential implementation system with some external staff support initially, and then to move progressively towards a more hands-off approach, but to continue to monitor processes carefully in at least a representative sample of locations.

When an intervention that can work at scale in the right organization has been successfully developed, it can be deployed at scale to evaluate the full effect of the intervention, including any spillover and market-level effects. A number of studies have been designed to estimate *equilibrium effects* by randomizing at the level of the

[7] In some cases, it will make sense to go straight to a fairly large stage-two trial, because the experiment does not even make sense on a small scale.

relevant market. Theory (and common sense) can guide the key design questions: On what variables (if any) do we expect to see equilibrium effects? What is the nature of those effects? Are we moving down a demand curve? Do they arise because of competition? What is the relevant market?

There are situations where relevant market equilibrium effects cannot be experimentally estimated—for example, because the entire country would be the right market. In the case of free secondary schooling in Ghana, for example, we expect that secondary school graduates will have a national market essentially (they can move to Accra, and they compete nationally for teacher and nurse training slots). In that case, the best a researcher can do is to combine the partial equilibrium results with some modeling and known elasticities, and exploit the understanding of the context to make some predictions about possible market equilibrium. In Ghana, the cohort that was subject to the experiment of Duflo, Dupas, and Kremer (2017) graduated as part of a "double cohort" (because the length of secondary school was brought down from four to three years after this cohort matriculated). Therefore, the authors conclude that the partial equilibrium impacts within that double cohort are probably a good approximation of what would happen if free secondary school doubled the share of graduates, at least in the short run. It is also useful to try to identify situations where one would expect market equilibrium effects to be too small to matter (consider, for example, the case of a preschool math programs that can be run by existing teachers).

Although conceptually distinct, *spillover effects* can be evaluated experimentally in the same way as market equilibrium, by randomizing at the appropriate level (and randomizing in two steps if one is particularly interested in the spillover themselves, not just the total effect). One open issue that has proven difficult is the identification of nonlinear spillovers. Conceptually, this requires randomization of treatment intensity at several points and then comparison of treated and untreated units in each treatment intensity. Crepon et al. (2013) adopts this design in their experiment on the French labor market (they treat 25, 50, and 75 percent of the units). Similarly, Banerjee et al. (2014) treat 25, 50, 75, or 100 percent of the police officers in police stations in Rajasthan. In practice, the Crepon et al. (2013) study lacks enough statistical precision to identify differential spillover effects (despite working at the scale of half of France). Banerjee et al. (2014) find a nonlinearity in overall effect of the treatment (there is no impact in treating 25 percent of the police officers, and the effect is the same when 50, 75, and 100 percent of the officers are trained), but they do not specifically track spillovers.

Finally, implementing the scale-up with the organization that would finally implement, within their standard operating procedures, and at a scale sufficient to detect market equilibrium effects will also give a chance for any potential *political backlash* to manifest itself. As mentioned above, this happened in the Banerjee, Duflo, Imbert, Mathew, and Pande (2017) study of anticorruption reforms in India. When backlash happens, it is worth exploring whether some changes in potentially inessential program details (perhaps some side payments to the aggrieved parties) are available. It is also important to try to anticipate the backlash and create a constituency for the reform from the start. Finally, the potential for political backlash may provide an argument for not doing too many pilots, since large-scale programs are less likely to be scotched.

■ *Thanks to Richard McDowell, Harris Eppsteiner, and Madeline Duhon for research assistance; to Tamayata Bansal, Sugat Bajracharya, Anupama Deshpande, Blaise Gonda, John Firth, Christian Larroulet, Adrien Lorenceau, Jonathan Mazumdar, Manaswini Rao, Paribhasha Sharma, Joseph Shields, Zakaria Siddiqui, Yashas Vaidya, and Melanie Wasserman for field management; to Diva Dhar for supervision; and to Shaher Bhanu Vagh for the educational test design and analysis. Special thanks to the staff of Pratham for their openness and engagement, and to the William and Flora Hewlett Foundation, the International Initiative for Impact Evaluation, the Government of Haryana, and the Regional Centers for Learning on Evaluation and Results for their financial support and commitment.*

References

Akram, Agha Ali, Shyamal Chowdhury, and Ahmed Musfiq Mobarak. 2017. "Effects of Emigration on Rural Labor Markets." Unpublished paper

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130(3): 1117–65.

Andrabi, Tahir, Jishnu Das, Selcuk Ozyurt, and Niharika Singh. 2017. "Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools." Unpublished paper.

Angelucci, Manuela, and Giacomo De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?" *American Economic Review* 99(1): 468–508.

Angrist, Joshua, and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24(2): 3–30.

Athey, Susan, and Guido Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments,* edited by Abhijit Vinayak Banerjee and Esther Duflo, 323–85. ScienceDirect.

Attanasio, Orazio P., Costas Meghir, and Ana Santiago. 2011. "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA." *Review of Economic Studies* 79(1): 37–66.

Banerjee, Abhijit. 2007. *Making Aid Work.* Cambridge, MA: MIT Press.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India." NBER Working Paper 22746.

Banerjee, Abhijit V., Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2010. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in India." *American Economic Journal: Economic Policy* 2(1): 1–30.

Banerjee, Abhijit, Sharon Barnhardt, and Esther Duflo. 2015. "Movies, Margins and Marketing: Encouraging the Adoption of Iron-Fortified Salt." NBER Working Paper 21616.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg. 2017. "Decision Theoretic Approaches to Experiment Design and External Validity." In *Handbook of Economic Field Experiments,* edited by Abhijit Vinayak Banerjee and Esther Duflo, 141–74. ScienceDirect.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2014. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training." NBER Working Paper 17912.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122(3): 1235–64.

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.*

Banerjee, Abhijit, Esther Duflo, and Rachel Glennerster. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Public Health Care System." *Journal of the European Economic Association* 6(2–3): 487–500.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Pariente, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science* 348(6236).

Banerjee, Abhijit, Esther Duflo, Clement Imbert, Santhosh Mathew, and Rohini Pande. 2017. "E-governance, Accountability, and Leakage in Public Programs: Experimental Evidence from a Financial Management Reform in India." CEPR Discussion Paper 1176.

Banerjee, Abhijit, Esther Duflo, and Michael Kremer. Forthcoming. "The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy." In *The State of Economics, The State of the World*, edited by Kaushik Basu.

Banerjee, Abhijit, Rema Hanna, Jordan Kyle, Benjamin A. Olken, and Sudarno Sumarto. 2016. "Tangible Information and Citizen Empowerment: Identification Cards and Food Subsidy Programs in Indonesia." https://economics.mit.edu/files/11877.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics* 7(1): 1–21.

Banerji, Rukmini, Madhav Chavan, and Usha Rane. 2004. "Learning to Read." *India Seminar*, April. http://www.indiaseminar.com/2004/536/536%20rukmini%20banerji%20%26%20et%20al.htm.

Barrera-Osorio, Felipe, and Leigh L. Linden. 2009. "The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia." World Bank Policy Research Working Paper 4836.

Berry, James, Greg Fischer, and Raymond P. Guiteras. 2015. "Eliciting and Utilizing Willingness-to-Pay: Evidence from Field Trials in Northern Ghana." CEPR Discussion Paper 10703.

Blair, Graeme, Radha K. Iyengar, and Jacob N. Shapiro. 2013. "Where Policy Experiments Are Conducted in Economics and Political Science: The Missing Autocracies." http://scholar.princeton.edu/sites/default/files/jns/files/blair_iyengar_shapiro_rcts_20may13.pdf.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2015. "Interventions and Institutions: Experimental Evidence on Scaling up Education Reforms in Kenya." Unpublished paper.

Breza, Emily, and Cynthia Kinnan. 2016. "Measuring the Equilibrium Impacts of Credit: Evidence from the Indian Microfinance Crisis." http://faculty.wcas.northwestern.edu/~cgk281/Eqm_impacts_credit.pdf.

Buera, Francisco, Joseph Kaboski, and Yongseok Shin. 2012. "The Macroeconomics of Microfinance." NBER Working Paper 17905.

Card, David, Jochen Kluve, and Andrea Weber. 2010. "Active Labour Market Policy Evaluations: A Meta-Analysis." *Economic Journal* 120(548): F452–77.

Chassang, Sylvain, Gerard Padró i Miquel, and Erik Snowberg. 2012. "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments." *American Economic Review* 102(4): 1279–1309.

Clasen, Thomas F., Kelly T. Alexander, David Sinclair, Sophie Boisson, Rachel Peletz, Howard H. Chang, Fiona Majorin, and Sandy Cairncross. 2015. "Interventions to Improve Water Quality for Preventing Diarrhoea." *Cochrane Database of Systematic Reviews* 10.

Coalition for Evidence-Based Policy. 2013. *Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive versus Weak or No Effects.* Washington, DC: Coalition for Evidence-Based Policy.

Cohen, Jessica, and Pascaline Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics* 125(1): 1–45.

Crepon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128(2): 531–80.

Cunha, Jesse, Giacomo De Giorgi, and Seema Jayachandran. 2011. "The Price Effects of Cash versus In-Kind Transfers." NBER Working Paper 17456.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2): 424–55.

Duflo, Esther. 2017. "The Economist as Plumber." NBER Working Paper 23213.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101(5): 1739–74.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School Governance, Teacher Incentives and Pupil–Teacher Ratios." *Journal of Public Economics* 123: 92–110.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2017. "The Impact of Free Secondary Education: Experimental Evidence from Ghana." https://web.stanford.edu/~pdupas/DDK_GhanaScholarships.pdf.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School. " *American Economic Review* 102(4): 1241–78.

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 118(3): 815–42.

Dupas, Pascaline. 2014. "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment." *Econometrica* 82(1): 197–228.

Dupas, Pascaline, and Edward Miguel. 2017. "Impacts and Determinants of Health Levels in Low-Income Countries." In *Handbook of Economic Field Experiments*, edited by Abhijit V. Banerjee and Esther Duflo, 3–93. ScienceDirect.

Evidence Action. 2014. "Kenya Deworming Results Announced: 6.4 Million Children Worm-Free and Ready to Learn." *Evidence Action*, August 4. http://www.evidenceaction.org/blog-full/kenya-deworming-results-announced-6-4-million-children-worm-free-and-ready-to-learn.

Finn, Jeremy D., and Charles M. Achilles. 1990. "Answers and Questions about Class Size: A Statewide Experiment." *American Educational Research Journal* 27(3): 557–77.

Glennerster, Rachel. 2017. "The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency." In *Handbook of Economic Field Experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo, 175–243. ScienceDirect.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1(1): 112–35.

Global Innovation Fund. 2017. *Stages of Financing*. Washington, DC: Global Innovation Fund.

Hausman, Jerry A., and David A. Wise. 1985. *Social Experimentation*. University of Chicago Press.

Heckman, James. 1992. "Randomization and Social Programs." In *Evaluating Welfare and Training Programs*, edited by Charles Manski and Irwin Garfinkel. Cambridge, MA: Harvard University Press.

Heckman, James J., Lance Lochner, and Christopher Taber. 1998. "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents." *Review of Economic Dynamics* 1: 1–58.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–75.

Innovations for Poverty Action. 2015. *Targeted Lessons to Improve Basic Skills*. New Haven, CT: Innovations for Poverty Action.

International Initiative for Impact Evaluation. 2017. "Systematic Reviews." International Initiative for Impact Evaluation. http://www.3ieimpact.org/en/evidence/systematic-reviews/ (accessed June 27, 2017).

Kapur, Avani, and Lorenza Icaza. 2010. "An Institutional Study of Read India in Bihar and Uttarakhand." Unpublished paper.

Kremer, Michael, and Rachel Glennerster. 2011. "Improving Health in Developing Countries: Evidence from Randomized Evaluations." In *Handbook of Health Economics*, edited by Mark V. Pauly, Thomas G. McGuire, and Pedro P. Barros, 201–315. ScienceDirect.

Lengeler, Christian. 2004. "Insecticide-Treated Bed Nets and Curtains for Preventing Malaria." *Cochrane Database of Systematic Reviews* 2.

Manski, Charles F., and Irwin Garfinkel. 1992. *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard University Press.

McEwan, Patrick J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-analysis of Randomized Experiments." *Review of Educational Research* 85(3): 353–94.

McKenzie, David, and Susana Puerto. 2017. "Growing Markets through Business Training for Female Entrepreneurs." World Bank Policy Research Working Paper 7793.

Meager, Rachael. 2016. "Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomized Experiments." https://economics.mit.edu/files/11443.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1): 159–217.

Muralidharan, Karthik. 2017. "Field Experiments in Education in Developing Countries." In *Handbook of Economic Field Experiments*, edited by Abhijit V. Banerjee and Esther Duflo, 323–85. ScienceDirect.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2016. "Building State Capacity: Evidence from Biometric Smartcards in India." *American Economic Review* 106(10): 2895–2929.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *Quarterly Journal of Economics* 130(3): 1011–66.

Newhouse, Joseph P. 1993. *Free for All? Lessons from the RAND Health Insurance Experiment.* Cambridge, MA: Harvard University Press.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments when Context Matters." *American Economic Review* 105(5): 471–75.

Rubin, Donald B. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Education and Behavioral Statistics* 6(4): 377–401.

Sackey, Ken. 2017. "Free SHS to Commence September 2017—President Akufo-Addo." *Ghana News Agency*, February 11. http://www.ghananewsagency.org/education/free-shs to-commence-september-2017-president-akufo-addo-113179.

Sandefur, Justin. 2011. "Held Hostage: Funding for a Proven Success in Global Development on Hold in Kenya." *Center for Global Development*, April 25.

Schweinhart, Lawrence J., Helen V. Barnes, and David P. Weikart. 1993. *Significant Benefits: The High/Scope Perry Preschool Study through Age 27.* Ypsilanti, MI: High/Scope Press.

Sharma, Paribhasha, and Anupama Deshpande. 2010. "Teachers' Perception of Primary Education and Mothers' Aspirations for Their Children—A Qualitative Study in Bihar and Uttarakhand." Unpublished paper.

Tarozzi, Alessandro, Aprajit Mahajan, Brian Blackburn, Dan Kopf, Lakshmi Krishnan, and Joanne Yoong. 2014. "Micro-loans, Insecticide-Treated Bednets, and Malaria: Evidence from a Randomized Controlled Trial in Orissa, India." *American Economic Review* 104(7): 1909–41.

Taylor-Robinson, David C., Nicola Maayan, Karla Soares-Weiser, Sarah Donegan, and Paul Garner. 2015. "Deworming Drugs for Soil-Transmitted Intestinal Worms in Children: Effects on Nutritional Indicators, Haemoglobin, and School Performance." *Cochrane Database of Systematic Reviews* 7.

Townsend, Robert. 2010. "Financial Structure and Economic Welfare: Applied General Equilibrium Development Economics." *Annual Review of Economics* 2(1): 507–46.

US Agency for International Development (USAID). 2017. "Development Innovation Ventures." Webpage. https://www.usaid.gov/div (accessed June 27, 2017).

Vivalt, Eva. 2016. "How Much Can We Generalize from Impact Evaluations?" http://evavivalt.com/wp-content/uploads/2014/12/Vivalt_JMP_latest.pdf.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 103–124

Experimentation at Scale

Karthik Muralidharan and Paul Niehaus

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 15 years. These experiments have advanced our understanding within a broad range of topics including education, health, governance, finance (credit, savings, insurance), and social protection programs, as summarized in 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, large-scale evaluations can directly inform large-scale spending decisions. 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 indicate whether they will work. Randomizing these rollouts can generate direct evidence on policy questions that are inarguably of interest—after all, such programs 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.

■ *Karthik Muralidharan and Paul Niehaus are both Associate Professors of Economics, University of California at San Diego, La Jolla, California. Both authors are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. They are also both Affiliated Professors with the Abdul Latif Jameel Poverty Action Lab, Cambridge, Massachusetts. Their email addresses are kamurali@ucsd.edu and pniehaus@ucsd.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.103 doi=10.1257/jep.31.4.103

Second, scale can help to improve "external validity," or the accuracy with which the estimates of impact from a randomized controlled trial 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 to which policymakers want to generalize their results; 2) program effects may differ when implemented at smaller scale (say, by a highly motivated nonprofit 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 relitigate these well-known issues, but instead to highlight one way in which the field experimental literature can make (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 nonrepresentative 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 (and likely worse) when the scale of the operation increases. Third, experiments can be randomized in large units such as villages or regions. This procedure 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 just mentioned: 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 listed 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 will be best, as proposed in this same symposium by Banerjee, Banerji, Berry, Duflo, Kannan, Mukerji, Shotland, and Walton. In our closing section, we discuss these tradeoffs, including some of the

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

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.

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 economics journals (the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*) from January 2001 to July 2016. We restricted our focus to experiments framed as (development) 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. 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. Our substantive conclusions are not sensitive to how we categorize borderline cases. Appendix A1, available with this paper at http://e-jep.org, describes the protocol for the exercise in more detail and Table A1 provides a full list of studies included and excluded. These totals 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 economics journals.

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 countrywide 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

Table 1
Summary Statistics: Program Evaluation Randomized Control Trials 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	490	10,885	46,418	681,918	2,715,917	26
Units treated	289	5,340	29,325	13,564	18,224	29
Clustered randomization?	0	1	1	0.62	0.49	29
Mean size of randomization unit	1	26	99	167	477	28

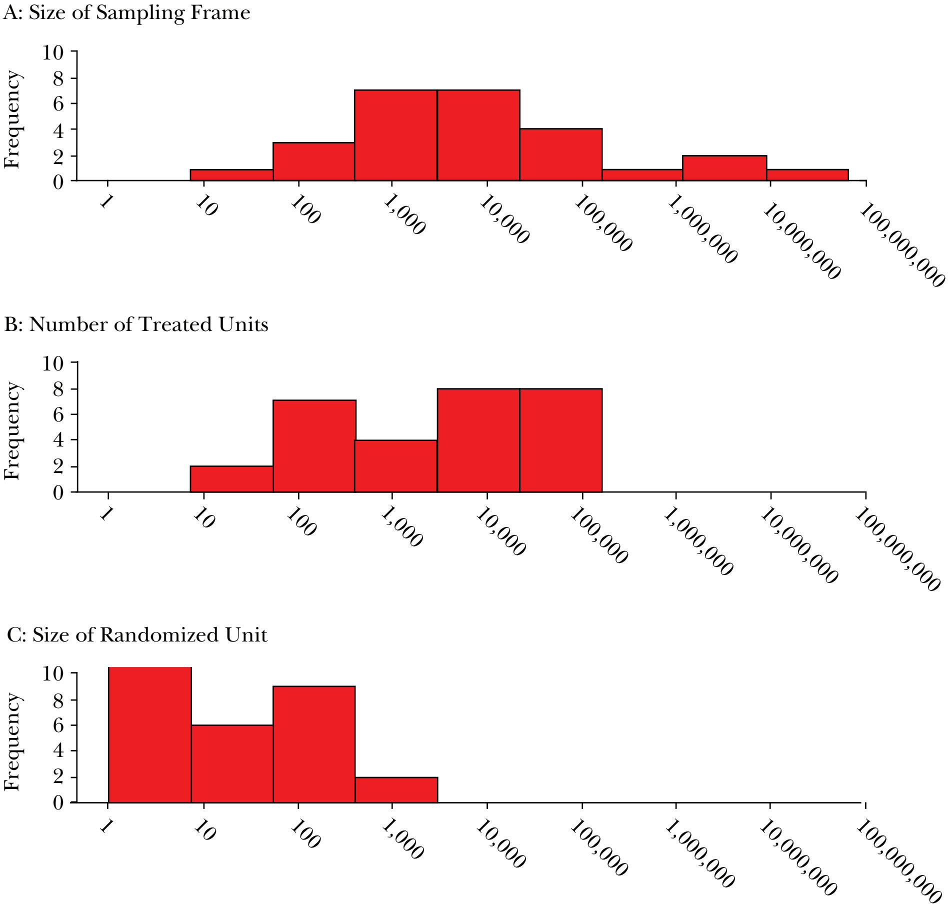
Note: This table reports summary statistics for measures of experimental scale for randomized controlled trials published in the *American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics,* and *Review of Economic Studies* between January 2001 and July 2016 that we categorized as primarily "program evaluations" (as opposed to mechanism experiments). Counting metrics are in "primary units of analysis," which we define as the level at which the studies' primary outcomes are measured (for example, 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 the 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).

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 1A plots the distribution of the size of the sampling frame on a logarithmic scale. 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 as the unit at which the outcome(s) we believe are most important for the study's thesis are measured.[2]

Generally speaking, the samples in the studies we reviewed are representative of small populations. Only 31 percent of the studies report that the sampling frame was itself drawn from a larger population (Table 1, row 1). Among the 26

[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, or other groups 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 would typically be students).

Figure 1
Distributions of Measures of Experimental Scale



Note: Figure shows the distribution of key attributes of "program evaluation" randomized control trials published in selected economics journals between January 2001 and July 2016. The x-axis has a logrithmic scale. For more detail on sample and variable construction, refer to notes for Table 1.

studies that report the size of their sampling frame, the median frame contains 10,885 units, while the 75th percentile frame contains 46,418 units (Table 1 row 2). These figures are obviously modest compared to tens or hundreds of millions of low-income people in the countries where the studies are run (typically to evaluate antipoverty programs). There are notable exceptions to this rule, however, which we discuss further below. For instance, Alatas et al. (2012) perform an experimental comparison of different methods for targeting welfare benefits to the poor in Indonesia on a representative sample of three large provinces; their study results are representative of a population of over 50,000,000 people.

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 (like governments) than the ones that initially developed and tested an intervention (like nongovernment organizations). 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, Kimenyi, Mwabu, Ng'ang'a, and Sandefur (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 effect when scaled up further and run by the Kenyan government. In a nonexperimental 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 a nongovernmental organization as opposed to a government body. More broadly, the productivity literature finds wide dispersion in the productivity of firms (for example, Hsieh and Klenow 2009) and plants (for example, Bloom, Eifert, Mahajan, McKenzie, and Roberts 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.

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 Figure 1B 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 that is modest relative to the full-scale implementation envisioned for those policies.

[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. The extent of adherence may depend on the patient, how the physician explains treatment to the patient, what monitoring protocols are put in place, and other factors.

The Scale of Units Randomized

The size of the units randomized may matter because of "spillovers," which in this situation refers to mechanisms through which an individual's outcomes depend not only on the person's 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 that would have occurred from the untreated to the treated had the former been treated as well (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, De Giorgi, and Jayachandran (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, one can alleviate this concern by using larger units of randomization in the experiment. For instance, 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 at larger geographical scales will reduce bias by increasing the spatial segregation of control from treatment units. Control units will be affected by spillovers from fewer nearby treated units, and treated units will be affected by spillovers from more nearby treatment 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 paper's 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 Figure 1C plots the full distribution of the latter (on a logarithmic scale).

We find that randomization is commonly "clustered": 62 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. In fact, the largest mean unit of randomization we identified, in the study with the largest randomized units, was just 2,500 households (in Björkman and Svensson 2009). The median design featured 26 units per cluster. 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—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 we found no evidence of a shift in either direction: none of the relationships we estimated were either statistically significant or economically meaningful.

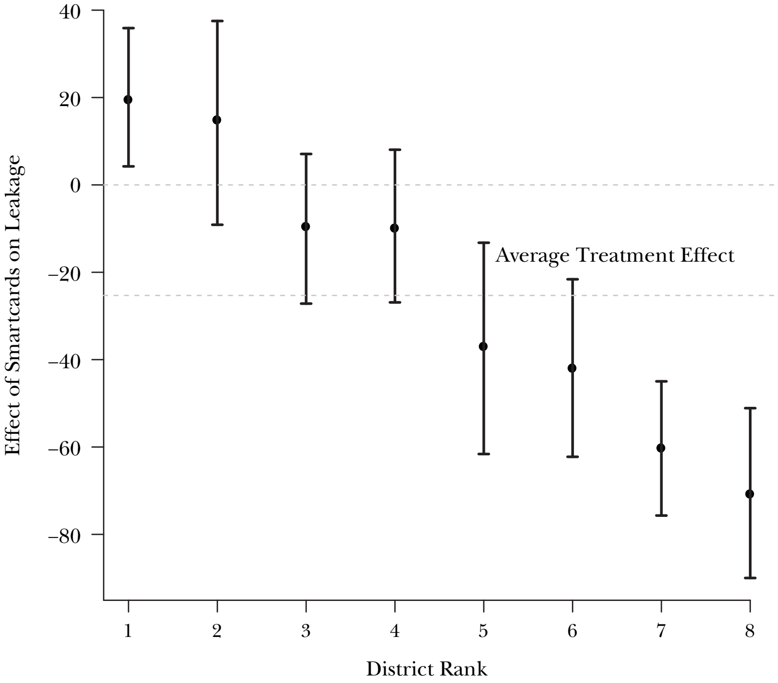
Experimenting at Scale: Some Examples

While impact evaluations have typically been small, a number of exceptions 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 biometrically authenticated payment system ("Smartcards") into two large anti-poverty programs in rural Andhra Pradesh (Muralidharan, Niehaus, and Sukhtankar 2016; Muralidharan, Niehaus, and Sukhtankar 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 20 million people, for example, and in clusters of 62,000 people.

Experiments in (Nearly) 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.

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 eight districts in the experimental sample of Muralidharan, Niehaus, and Sukhtankar (2016). Error bars show the 90 percent conffidence interval generated through a block bootstrap.

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 beneficiaries—across 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 subdistricts and sampled villages/households (sampled with replacement)

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. (2012) 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, Muralidharan, Pradhan, and Rogers (forthcoming) 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 for 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 nonrepresentative sample of the same size; the additional costs 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 nongovernment 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 official in India who told one of us that it was "not worth

and plot the distribution of treatment effects that would be obtained from such a study sample. As both Figure A1 (Panel B) and Table A4 (row 2) show, the resulting estimates would be much less precise, and a 90 percent 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 Pritchett and Sandefur 2013). 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 nonstudy 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 Table A4 and Figure A1 (panel C), 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.

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 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 other populations, 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 having (say) run it in a single district. 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 10,000.

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 of experimentation at scale is the Progresa-Oportunidades program 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 (Gertler and Boyce 2003; Rivera, Sotres-Alvarez, Habicht, Shamah, and Villapando 2004; Schultz 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 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.

[6] One approach to this challenge is to conduct multisite experiments where the same/similar program is experimentally evaluated in multiple locations. Such an approach is exemplified by Banerjee, Duflo, Imbert, Mathew, and Pande (2015), who report results on the impact of a "graduation" program in reducing poverty across six different countries. However, that paper does not report the representativeness of the study populations within each country.

A third example is De Ree et al. (forthcoming) 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 (because 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 effects 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 conditional cash transfer programs 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 also 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.

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 and to measure spillovers was crucial to estimating policy parameters accurately.

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 also 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 underestimated 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 effect 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 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 finding was important in itself, as the possibility of spillovers had been widely conjectured in the earlier school choice literature.

A final example is the Smartcard evaluation, which randomly assigned subdistricts of Andhra Pradesh to treatment and control categories. Since a subdistrict 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

[7] The estimates 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.

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 subdistrict 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, De Giorgi, and Jayachandran (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, see Muralidharan 2017).

Some Practical Considerations

Running large-scale experiments has merit, but it 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?

When to Go Big, and How to Do Small Well

Not all experiments should be "big"; certainly, balance is needed. 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 institutional 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 the headquarters of a nongovernment organization or the office of a research unit.

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 researcher's 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 researcher's 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 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 (as in Muralidharan, Singh, and Ganimian 2017), 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 than 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 (for example, by a government). Some 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

[8] This smaller-scale evaluation might itself be the first step in an optimal sequence of experimentation, as discussed in the paper in this symposium by Banerjee, Banerji, Berry, Duflo, Kannan, Mukerji, Shotland, and Walton. 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.

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, Finan, and Rossi (2013), who study the impact of varying the salary offered on the quality of public employees recruited in Mexico; and d) Khan, Khwaja, and Olken (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 the effects on behaviors which would be likely to affect prices in general equilibrium, and then forecasting the likely effects. 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 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 the 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 nongovernment organization GiveDirectly, which studies the effects of capital inflows equivalent to about 15 percent of GDP in treated communities in Kenya (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.

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 that 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 that 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 the Abdul Latif Jameel Poverty Action Lab (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 indicate these to the research community, for example, in the acknowledgements to papers. This is the model in the natural sciences, where the 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 Smartcard

[9] These issues are not restricted to field-experimental research. Similar changes may be needed to support work in teams working with administrative datasets 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.

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, on-boarding 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 significant risk, as it can be difficult to hold the government to an 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 we 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 Tamil Nadu 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, which has yielded several high-quality papers that have influenced both research and policy (Alatas et al. 2012; Alatas et al. 2016; Banerjee, Hanna, Kyle, Olken, and Sumarto 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 a nongovernment or government organization in order to publish strong research has weakened incentives for objectivity. While this issue is hardly a new one, we wish to highlight safeguards that we have found important in practice. First, researchers should use Memorandums of Understanding and pre-analysis plans judiciously as a means of protecting themselves against pressure to shade or spin their analysis as the research findings become apparent. 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 programs 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 defunding) 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 quotation from the former Indonesian Finance Minister).[10]

In conclusion, the past 15 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 randomized controlled trials for policy evaluation. Our goal in this paper has been to demonstrate one practical way to combine the credibility and transparency of randomized controlled trials with greater policy relevance, which is to run experiments at a larger scale.

We believe that this approach is fruitful to pursue both because large-scale randomized controlled trials 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

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

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 nontreated 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.

■ *We thank David Lagakos, Abhijeet Singh, Gordon Hanson, and all the JEP editors 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.*

References

Acemoglu, Daron. 2010. "Theory, General Equilibrium, and Political Economy in Development Economics." *Journal of Economic Perspectives* 24(3): 17–32.

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

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

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130(3): 1117–65.

Banerjee, Abhijit, Esther Duflo, Clement Imbert, Santhosh Mathew, and Rohini Pande. 2015. "Can E-governance Reduce Capture of Public Programs? Experimental Evidence from a Financial Reform of India's Employment Guarantee." International Initiative for Impact Evaluation Impact Evaluation Report 31.

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

Björkman, Martina, and Jakob Svensson. 2009. "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda." *Quarterly Journal of Economics* 124(2): 735–69.

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

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2013. "Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education." Center for Global Development Working Paper 321.

Cunha, Jesse M., Giacomo De Giorgi, and Seema Jayachandran. 2015. "The Price Effects of Cash versus In-Kind Transfers." Federal Reserve Bank of New York Staff Report 735.

Dal Bó, Ernesto, Frederico Finan, and Martín A. Rossi. 2013. "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service." *Quarterly Journal of Economics* 128(3): 1169–1218.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of*

Economic Literature 48(2): 424–55.

Deaton, Angus, and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." NBER Working Paper 22595.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. Forthcoming. "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia." *Quarterly Journal of Economics.*

Duflo, Esther, and Abhijit Banerjee, ed. 2017. *Handbook of Field Experiments*, vol. 1. Handbooks in Economics. Amsterdam: Elsevier.

Fryer, Roland G., Jr. 2017. "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." Chap. 2 in *Handbook of Economic Field Experiments*, Vol. 2, edited by Esther Duflo and Abhijit Banerjee. Amsterdam: Elsevier.

Gertler, Paul J., and Simone Boyce. 2003. "An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico." Paper presented at Royal Economic Society Annual Conference, Coventry, UK, April 7–9.

Gueron, Judith M. 2017. "The Politics and Practice of Social Experiments: Seeds of a Revolution." In *Handbook of Field Experiments*, vol. 1, edited by Abhijit Vinayak Banerjee and Esther Duflo, 27–69. Amsterdam: Elsevier.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9(2): 85–110.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *Quarterly Journal of Economics* 124(4): 1403–48.

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

Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken. 2016. "Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors." *Quarterly Journal of Economics* 131(1): 219–71.

Kleven, Henrik Jacobsen, Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez. 2011. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." *Econometrica* 79(3): 651–92.

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

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

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2016. "Inputs, Incentives and Complementarities in Primary Education: Experimental Evidence from Tanzania." https://economia.uniandes.edu.co/images/archivos/pdfs/CEDE/SeminariosCEDE/2016/Mauricio_Romero.pdf.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1): 159–217.

Muralidharan, Karthik. 2017. "Field Experiments in Education in Developing Countries." In *Handbook of Economic Field Experiments*, Vol. 2, edited by Abhijit Vinayak Banerjee and Esther Duflo, 323–85. Amsterdam: Elsevier.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2016. "Building State Capacity: Evidence from Biometric Smartcards in India." *American Economic Review* 106(10): 2895–2929.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2017. "General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence from India." Unpublished paper.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro Ganimian. 2017. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." NBER Working Paper 22923.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *Economic Journal* 120(546): F187–F203.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119(1): 39–77.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. "Contract Teachers: Experimental Evidence from India." NBER Working Paper 19440.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *Quarterly Journal of Economics* 130(3): 1011–66.

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

Pritchett, Lant, and Justin Sandefur. 2013. "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix." Working Paper 336, Center for Global Development.

Rivera, Juan A. Daniela Sotres-Alvarez, Jean-Pierre Habicht, Teresa Shamah, and Salvador Villalpando. 2004. "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." *Journal of the American Medical Association* 291(21): 2563–70.

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74(1): 199–250.

Tarozzi, Alessandro, Aprajit Mahajan, Brian Blackburn, Dan Kopf, Lakshmi Krishnan, and Joanne Yoong. 2014. "Micro-loans, Insecticide-Treated Bednets, and Malaria: Evidence from a Randomized Controlled Trial in Orissa, India." *American Economic Review* 104(7): 1909–41.

Vivalt, Eva. 2015. "Heterogeneous Treatment Effects in Impact Evaluation." *American Economic Review* 105(5): 467–70.

Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature

Omar Al-Ubaydli, John A. List, Danielle LoRe, and Dana Suskind

Many economists would be surprised to learn that patients adhere to the medications that physicians prescribe as little as 50 percent of the time (McDonald, Garg, and Haynes 2002). Clinical non-adherence is more than just an inconvenience to medical practitioners—it represents wasted resources and causes medical problems to evolve into forms that are even more expensive to treat. This has driven medical researchers to investigate rigorously ways of improving patient adherence. Their findings are of interest to economists who study interventions and wish to ensure that the inferences they draw from small-scale studies apply at larger scales, too.

More specifically, many experimental studies in economics are evaluations of a modification to an individual's behavior, based on the researchers' belief that such a modification will benefit the individual or confer benefits upon society. In the event that these beliefs are supported by the generated data, the broader goal is for large

■ *Omar Al-Ubaydli is a Researcher at the Bahrain Center for Strategic, International and Energy Studies, Manama, Bahrain, an Affiliated Associate Professor of Economics at George Mason University, Fairfax, Virginia, and an Affiliated Senior Research Fellow at the Mercatus Center, Arlington, Virginia. John List is Kenneth C. Griffin Distinguished Service Professor of Economics, University of Chicago, Chicago, Illinois, and a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Danielle LoRe is a Medical Student at the University of Chicago Pritzker School of Medicine. Dana Suskind is Professor of Surgery and Pediatrics at the Duchossois Center for Advanced Medicine, University of Chicago, Chicago, Illinois. Al-Ubaydli is the corresponding author at omar@omar.ec.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.125 doi=10.1257/jep.31.4.125

groups of individuals to then adopt the proposed intervention autonomously, or to have policymakers promote its adoption via methods such as subsidies, awareness campaigns, legislation, and so on.

For example, in an effort to provide policymakers with recommendations on how to improve agricultural productivity, Duflo, Kremer, and Robinson (2011) investigate the benefits accrued to Kenyan farmers who invest in fertilizer. Likewise, Field and Pande (2008) study the effect of loan repayment frequency on client default in microfinance by experimentally manipulating repayment schedules, with an eye to supplying creditors with scientific information about optimal loan structure.

These studies typically result in a recommendation to policymakers about how to affect broad-based behaviors outside of the experiment. The studies also show why people were deviating from the optimal behavior prior to study. For example, perhaps they simply did not realize that they could do better by modifying their behavior, as in the Kenyan farmer case (Duflo, Kremer, and Robinson 2011). Or infeasibility could be the driver, as in the case of studies investigating novel repayment schedules, or it could be something beyond the agent's financial means, such as Fryer's (2011) study of financial incentives in schools.

While scientists typically have clear advice for policymakers, a common occurrence is that such research programs are never scaled, or when they are scaled, the size of the measured treatment effect diminishes substantially relative to the that found in the original study. This is a common phenomenon known in the literature as "voltage drop," but this type of predictable change is not accounted for in benefit–cost analysis. The papers cited above have not, to the best of our knowledge, exhibited voltage drop—we mentioned them because they are archetypal experimental papers where voltage drop is a concern.

In terms of actual examples of voltage drop, consider significant public health threats, such as HIV, tuberculosis, and malaria. Despite the demonstrated effectiveness of drug therapies on transmission in trials and in small-scale settings, prevalence in developing countries remains high. For example, Global Fund to Fight AIDS, Tuberculosis, and Malaria sought to raise and disburse money to poor countries in an effort to provide therapies (Lu, Khan, and Murray 2006). However, half the funds never reached clinics due to inability of health services to manage the funds effectively (Garrett 2007), resulting in significantly weaker "effective" treatment effects. This voltage drop was caused by financial and resource constraints that were never overcome.

A related example includes initiatives to decrease rates of transmission and promote safe sex practice, such as providing condoms to a community. Such practices have faced voltage drops when scaling, due to variations in community beliefs and values (Campbell and Mzaidume 2002). In a review of barriers to HIV intervention implementation, stigmatism of prevention and treatment, power dynamics within society, and plateauing of health education messaging were all identified as decreasing efficacy of these provisions (Chopra and Ford 2005).

The experimental literature is littered with such examples. In a general sense, the issue revolves around a query of this form: "I just found a 0.2 standard deviation effect in my experiment. Should I expect to observe such an effect when scaled to a city, state, or country?" Voltage drop can occur for many reasons, and it constitutes an example of a particularly vexing public policy challenge that we denote "the scaling problem." Threats to scalability can be divided into three areas: statistical inference; representativeness of the experimental population; and representativeness of the experimental situation (Al-Ubaydli, List, and Suskind 2017).

The statistical inference class of scaling problem relates to inferential errors by scholars and policymakers seeking to apply insights gleaned from a small-scale study to a larger-scale setting. These errors typically relate to a failure to adjust correctly for the fact that a published statistical finding about the relationship between two variables is merely one of numerous, contemporaneous, investigations into the same relationship conducted by other research teams working quasi-independently (Maniadis, Tufano, and List 2014). This inference problem is related to the broader problem of publication bias (Young, Ioannudis, and Al-Ubaydli 2008).

The "representativeness of the experimental population" class of scaling problem refers to the challenge of ensuring that the subject pool in a small-scale study is representative of the larger population targeted by policymakers seeking to scale the findings. For example, does an experimental investigation of a proposed intervention that improves saving habits, which is conducted on college students, yield results that apply to the entire population, which is mostly composed of people who are not currently enrolled in college? Note that this class of scaling problem is not specific to experimental data—it applies to naturally occurring data, too. However, there exists a lively debate about the susceptibility of various data types to this type of scaling problem (Al-Ubaydli and List 2015b; Deaton and Cartwright 2016).

While both statistical inference and representativeness of the population are important, the focus of our discussion is on representativeness of the experimental situation. The experimental situation is quite rich and includes many important considerations. For instance, the next simple example of "program drift" illustrates one set of reasons for the scalability problem within the experimental situation. Consider Early Head Start home visiting services, one of the largest federally funded early childhood interventions in the world. The program demonstrated significantly improved school readiness for children aged up to three years old, improved family economic self-sufficiency, and parenting practices through high-quality efficacy trials (Paulsell, Avellar, Martin, and Del Grosso 2010). However, variation in quality of home visits was found at larger scale, with home visits for "at risk" families involving more distractions and less time on child-focused activities, causing the delivery of a different program than what had been studied. Lower proportion of time on child-focused activities and lower parental engagement was associated with diminished effectiveness for both child and parent outcomes as well as higher dropout rates (Raikes, Green, Atwater, Kisker, and Constantine 2006; Roggman, Cook, Peterson, and Raikes 2008). General equilibrium effects also fall under this class of scaling problem, whereby fidelity to the original small-scale design at a larger-scale setting results in interactions with other variables, in turn causing treatment effects to structurally change.

Closely related to delivery of the wrong program, or the wrong dosage of the program, is that no program at all is received. For instance, individual non-adoption of treatment represents a serious consideration when assessing the efficacy of public programs. Even when non-adoption apparently contradicts the best interests of agents, it is often found within government programs or after findings from research studies are made public. For example, many scholars are puzzled by the persistent reluctance of consumers to purchase energy-efficient lightbulbs despite the manifest cost savings that they offer and the absence of any notable downside to their usage (Allcott and Taubinsky 2015). Likewise, when new technologies are advanced as governmental policies, such as agricultural, financial, or time-saving computer technologies (like renewing one's passport online), the level of adoption is typically far less than most anticipate.

Medical practitioners have for centuries been facing an isomorphic problem: patient non-adherence to prescribed medications, specifically those that manifestly serve the patient's interests as opposed to cutting-edge, experimental medicines without established efficacy or medicines that require invasive administration with high risk of complications or intolerable side effects. This problem has spawned a large literature in the medical sciences regarding the best practices for improving medication adherence, and the results are directly relevant to economists seeking to tackle the narrow component of the scaling problem considered in this paper. In this study, we explore the findings of medical practitioners and present a series of recommendations tailored to the environments typically studied by economists.

One reasonable response to the problem of non-adherence when scaling is to interpret this as a failure to design the original, small-scale study properly. Thus, rather than proposing ways to enhance adherence, we could focus on how to design the original experiment such that it correctly captures the expected level of adherence in the larger scale. We regard the two approaches as complementary. We choose to focus on techniques for boosting adherence because low adherence in the general population undermines treatment effects and therefore the effectiveness of policy. Moreover, empirical researchers often gather experimental data precisely because the enhanced control possible in the experimental setting allows for cheaper and more powerful estimates of treatment effects, as a precursor to more effective policy. In other words, the low adherence levels observed in large-scale field settings should not be taken as an inescapable constraint; part of the study should involve considerations to boost adherence.

Scaling and Researcher Control over Non-Adoption in Economics Experiments

Controlled experiments, be they laboratory or field, have become mainstream in economics only during the last 30 years, and therefore scaling issues, including the specific one under consideration in this study, are relatively recent problems for economists. Economics experiments typically involve unusual levels of control over the options available to agents. In the case of laboratory experiments as well as most field experiments (with the exception of natural field experiments), researcher control is close to absolute, allowing scholars to severely restrict choice sets: for example, *cooperate* or *defect* in a prisoner's dilemma game (Andreoni and Miller 1993), or *get paid to go to the gym* or *do not get paid* (Charness and Gneezy 2009).

Even in the case of natural field experiments, researchers often select natural environments that offer enhanced levels of control. For example, Shearer (2004) experimentally compared piece-rate and flat compensation schemes in a natural setting by finding a rare case of a company that uses both compensation schemes in its operations for the same type of work, and would therefore be willing to experimentally (and covertly) modify the compensation scheme offered to its workers.

Elevated levels of control are a key reason for the attractiveness of experimental methods when seeking to evaluate the consequences of a proposed modification to agents' behavior. For example, in the Duflo, Kremer, and Robinson (2011) fertilizer experiment, a first step toward demonstrating the superiority of fertilizer, and therefore convincing farmers of the benefits of using fertilizer, is heavily subsidizing the fertilizer or the delivery thereof. Similarly, Gosnell, List, and Metcalf (2016) were able to alert airline pilots and their managers to the possibility of enhancing fuel efficiency by offering pilots elevated levels of feedback, and by providing them the opportunity to earn money for charity.

Many of these evaluative studies result in a conclusion of the form: "upon being made aware of our findings, agents should autonomously adopt the behavior that was investigated," based on the premise that a primary barrier to the previous adoption of the behavior was informational. Yet in practice, economists often find lukewarm receptions for their findings among the agents who should respond by updating their behavior—whether the audience is policymakers, firms, or laypeople. When Ferraro and Price (2013) demonstrated that providing social comparisons in utility bills yields significant improvements in conservation at a trivial cost, utility companies the world over should have, in principle, expressed interest in deploying similar policies, yet this has not occurred. Likewise, Hossain and List (2012) discovered that the productivity-enhancing effects of providing Chinese factory workers with financial incentives led to a net increase in profits, and that this effect was larger when the incentives were presented in a negative frame—the most novel component of the experiment. According to neoclassical economics, merely publicizing this finding should lead to substantial enthusiasm for the adoption of such methods, and the scaling of the result. To our knowledge, this has not occurred widely.

Admittedly, in Ferraro and Price (2013) and Hossain and List (2012), and more generally in the case of the thousands of other economics experiments conducted, some of the reluctance among agents to modify behavior is due to uncertainty over the generalizability of the finding in question—what works for a Georgia water company might not work for a Slovenian electricity provider. It may work in China, but does it work in Toledo?

However, we can be confident that part of the non-adoption can be classified as purely irrational behavior, as many of the relevant studies investigate the promotion of actions that people should already be undertaking themselves. This includes Charness and Gneezy's (2009) use of financial incentives to induce greater exercise and Allcott and Taubinsky's (2015) attempts at increasing usage of energy-efficient lightbulbs. Moreover, in laboratory and field experiments (except natural field experiments), purely irrational behavior may be temporarily suppressed by experimenter demand effects or by the artificial restrictions on choices available to participants (Levitt and List 2007), accentuating the discrepancy between agents' willingness to modify their behavior in the study and in the natural environments ultimately targeted by the researchers.

This state of affairs poses a problem for policymakers seeking to scale empirical findings. If policymakers rely purely on publication of the results, then adoption will be impaired by irrational non-adoption. Alternatively, should the policymakers try to replicate the methods used in the original study, then they will face a host of structural scaling problems, such as the rising marginal cost of program administration, heterogeneity in the population, intransigence by stakeholders who are invested in the prevailing mode of behavior, and a litany of other issues discussed more fully in the rest of this symposium and in Al-Ubaydli, List, and Suskind (2017).

To illustrate these issues with a concrete irrational non-adherence example from the economics literature, consider the small-scale study conducted by Fryer, Levitt, and List (2015). Using a sample of 257 families from Chicago, the authors studied the effect of providing parents with financial incentives to engage in behaviors designed to increase early childhood skills via a parent academy that delivered training sessions. The study found large and statistically significant effects; in particular, over 80 percent of parents attended at least one training session, and over 40 percent attended all sessions, which is a crucial link in the causal chain under investigation. Inspired by these findings, the UK Education Endowment Foundation launched a parenting academy and a study structured similar to that in Chicago, but involving over 2,500 children spread across a larger geographical area. The larger-scale program found that only 60 percent of parents attended at least one session, and only 11 percent attended all sessions. Unsurprisingly, with such weak attendance, the study found no evidence of a positive effect of the interventions (in fact, the absence of an effect was true even when controlling for attendance).

Examples as clear as this are rare in economics, simply because this system of small-scale experimental research leading to large-scale policy implementation is a recent addition to the discipline. We anticipate that such problems will increase in frequency over the coming years as a larger volume of the profession's research resources are dedicated to this system for delivering policy insights. In this spirit, we envision nonprofit and for-profit firms, governmental bodies from local to federal, and supernational authorities as strong demanders of information on the causal effects of interventions.

If this is indeed the case, then there are considerable benefits associated with devising methods to deal with irrational non-adoption, and the first step is to obtain a better understanding of the underlying causes. The large behavioral economics literature offers many convincing explanations; based on our experience, we draw attention to the following likely sources.

First, humans experience *psychological switching costs* (Klemperer 1987; Carroll, Overland, and Weil 2000) and tend to dislike modifying their behavior for reasons independent of any material cost associated with changes in behavior. This may be a manifestation of an overarching tendency for humans to exhibit path dependence, which commonly surfaces in the form of the endowment effect (Kahnemann, Knetsch, and Thaler 1991). It may also reflect a propensity to herd: that is, to avoid deviating from the manner in which peers are behaving (Chang, Cheng, and Khorana 2000). In the context of scaling small-scale experimental results, psychological switching costs and habit formation constitute a barrier to the organic modification of behavior prescribed by a study.

Second, humans often display *hyperbolic discounting* (Laibson 1997), that is, when the cost is borne up front, they can indefinitely delay decisions that serve their interests because of an irrational fixation on reaping short-term rewards. This model of decision-making is used to explain apparently irrational patterns of credit card usage, such as borrowing at a high interest rate while simultaneously depositing money in a checking account (Telyukova and Wright 2008), as well as irrationality in pensions and savings decisions (Thaler and Benartzi 2004). Thus, even when a small-scale experiment demonstrates the benefits of a modifying behavior, agents may still exhibit reluctance toward organically adopting the change if it requires an up-front cost, despite the back-loaded benefits more than offsetting the up-front costs.

Third, when *complexity is combined with limited cognitive capabilities* (Simon 1972), humans may sometimes wish to take a certain course of action in the pursuit of their interests but be prevented from correctly modifying their behavior by limited cognitive abilities. For example, many who succumb to the Allais (1990) paradox (which involves choices between different sets of gambles) lack the intellectual capacity to understand the potential sub-optimality of their actions. In the context of insurance, consumers exhibit significant difficulty in making rational assessments of the premiums and deductibles offered in contracts (Watt, Vazquez, and Moreno 2001). Similarly, people may have systemically incorrect beliefs about the consequences of actions (Caplan 2002). In the context of small-scale experiments, not all agents are equipped with the cognitive tools necessary to appreciate the benefit of a prescribed change in behavior, or to acquire and process the information required to make a sound judgment.

These explanations are not intended to be exhaustive; our aim is simply to illustrate that the economics literature provides us with rich refinements to the baseline neoclassical model of decision-making that can account for why people sometimes seemingly refuse to pick up the proverbial dollar bills from the sidewalk. While economists have investigated a broad range of appropriate countermeasures, when it comes to the problem of getting people to modify their behavior for their own benefit, we can draw important lessons from attempts to address one version of this problem in medicine.

An Isomorphic Problem: Medication Non-Adherence

Clinicians take great care to ensure that the medications they prescribe to their patients serve their patients' interests. This includes social norms such as the Hippocratic oath (Orr, Pang, Pellegrino, and Siegler 1997), a sophisticated system of oversight by clinical peers and administrators (Farnan et al. 2012), and the threat of legal action in the event that clinicians fail to serve patients' best interests (Studdert et al. 2006). Moreover, clinical units that underperform suffer adverse commercial consequences, as consumers care about reputation (Hibbard, Stockard, and Tusler 2005).

While no system is perfect, medical practitioners should be considered highly motivated to provide sound prescriptions that patients can trust. Despite these favorable conditions from the perspective of patients, typical adherence rates for prescribed medications are around 50 percent (McDonald, Garg, and Haynes 2002), a figure that is of grave concern for clinicians because it diminishes the benefits of the treatments. This non-adherence rate has broad implications, including raising the costs of healthcare, prolonging patient discomfort, allowing disease progression, and biasing assessments of the effectiveness of treatments (Vervloet et al. 2012).

What drives so many patients to behave consistently in a manner that apparently contradicts their best interests?

Causes of Medication Non-Adherence

The medical literature has identified two primary classes of cause for medication non-adherence: intentional and unintentional (Kripalani et al. 2007; Vervloet et al. 2012; Dayer, Heldenbrand, Anderson, Gubbins, and Martin 2013). Interestingly, these two causes overlap with the causes pinpointed for irrational non-adoption discussed in the economics literature described above.

Intentional non-adherence, which refers to willful cost–benefit analysis by the patient, usually results from the patient attaching significant discomfort to the medication, and assessing that the purported benefits from the medication do not justify the discomfort. For example, a liquid medicine may have a disagreeable taste or may need to be administered via a painful injection. Intentional non-adherence may be exacerbated by systematically inaccurate beliefs on the effects of a medication, such as when patients prefer anecdotal evidence, or the advice of a celebrity, to the results of formal studies.

One such example helps to illustrate this mechanism at work—recall that the decline in vaccination rates and concurrent rise in vaccine-preventable diseases align with an anti-vaccine movement that, despite significant scientific evidence to the contrary, was fueled by personal stories and celebrity endorsement. At the time, such information outweighed pro-vaccine information on user-friendly outlets such as YouTube (Venkatraman, Garg, and Kumar 2015).

There are, of course, situations where intentional non-adherence can be a rational personal choice. In terminal disease and end-of-life care, it is straightforward to make the case that the small gain in lifespan provided by a medication is objectively not worth the decrease in quality of life. However, for most situations, such non-adherence corresponds to economists' model of hyperbolic discounting, possibly combined with cognitive limitations and/or biased beliefs.

Unintentional non-adherence is a major hurdle for clinicians and patients. It covers simple forgetfulness, as well as failures to comply with treatment plans resulting from regimen complexity, which can stem from quantity of medications and frequent dosage times or complicated, multistep administration of medicine (as with inhalers); as a result, treatment may be carried out incorrectly and thus ineffectively (Lavorini et al. 2008). Physical problems, such as sleeping through a scheduled treatment appointment due to fatigue, or lacking the mobility to adhere to a regimen, are also included in this class of cause.

Clinicians have devised a diverse range of interventions to address these factors, and have tested them using randomized control trials. Before evaluating these interventions, it is worth considering what can be inferred about medication adherence from naturally occurring variation in treatment features and background variables. Summarizing the literature broadly, McDonald et al. (2002) conclude that compliance is at best weakly related to sociodemographic factors, including age, sex, race, intelligence, and education.

Interestingly, McDonald et al. (2002) also found that patients with physical disabilities caused by the disease being treated were more likely to adhere to the prescribed regimen. Clearly, it is difficult to ascertain the degree to which such results generalize to the domains most frequently encountered by economists, but this finding may reflect simple cost–benefit dynamics: that is, those who benefit most from a treatment make the most effort to adhere. For example, Alcott and Taubinsky (2015) report that a significant proportion of non-adoption of energy-efficient lightbulbs might potentially be attributable to the fact that the financial returns of switching—while being positive in net terms—were too small to justify the act. Furthermore, and as expected, natural increases in the cost/complexity/duration of treatment plans are also associated with diminished compliance.

Improving Medication Adherence: Methods

Improving medication adherence is a central problem in the medical sciences, spawning thousands of papers and dozens of meta-studies. Peterson, Takiya, and Finley (2003) provide a useful categorization of the types of interventions that practitioners have evaluated in formal trials, many of which should be instantly recognizable to experimental economists who conduct small-scale studies with the goal of scaling their results to larger populations.

One important class of studies is *educational interventions*, whereby the medical team attempts to plug any informational lacunae that the patient may be suffering.

For instance, they may provide instructions on how to take a specific medicine and address misconceptions that the patient might have regarding the treatment's effectiveness or its side effects. Educational interventions vary along many dimensions, such as the medium (oral, visual, written), the delivery method (in person, telephone, electronic, printed), the professional delivering the intervention (physician, nurse, pharmacist), the frequency (one-off, weekly, monthly), the location (home, hospital, community center, remote), and the number of participants (one-to-one, group).

Related to educational interventions are *counseling and accountability interventions*, whereby members of the clinical team follow-up with the patient on the treatment to ensure that it is being taken as prescribed. In addition, under this approach, informational and psychological support are provided as the need arises. These interventions may feature monitoring devices, such as remote blood pressure sensors, that assist clinicians in gathering accurate information about a patient's adherence to best provide personalized counseling.

Medical practitioners have also investigated *interventions that support the patient's independent adherence efforts*, such as self-monitoring devices, including simple pillboxes that help patients track how many pills they have ingested, as well as more sophisticated electronic aids that measure vital signs. The advent of mobile telephone technology has greatly enhanced the opportunity to make use of automatic reminders, including text messages and notifications from smartphone applications.

An intermediate form of intervention, familiar to economists studying microfinance, is *involving family members in counseling and educational sessions*. Family members can directly assist in the delivery of treatment (for example, by injecting a patient who might otherwise be reluctant to inject themselves), provide reminders and emotional support, and give clinicians richer feedback on the degree of adherence and on the sources of non-adherence.

A final intervention class that—to the best of our knowledge is scarcely deployed in the medical non-adherence literature—is *using financial incentives*. While clinicians regularly advocate subsidizing treatment plans up to the point of free provision, reflecting a tacit acceptance of the importance of financial considerations in the patient's adherence calculus, there appears to be very little appetite for actually *paying* patients to take medicines as prescribed. This holds even if a reasonable cost–benefit case can be made in terms of the medical authorities avoiding more expensive treatments further down the road arising from non-adherence at present (Guiffrida and Torgerson 1997).

Several reasons have been suggested for this comparative rarity of financial incentives for medical adherence. For example, one concern may be due to clinician awareness of the debate regarding extrinsic versus intrinsic incentives (Deci, Koestner, and Ryan 1999; Benabou and Tirole 2003) and the fear that extrinsic incentives may diminish intrinsic motivation. Alternatively, there may be fears that positive financial incentives could induce spurious claims for the need for treatment. Yet another possibility based on sunk cost reasoning could account for clinician reluctance: if the individual pays a positive amount for medication, then they are more likely to use the product than if they receive it for free.

A field experiment due to Ashraf, Berry, and Shapiro (2010) suggests that this is might be the case for households using Clorin to treat drinking water: although they find no evidence that people who paid lower prices consume the product less than those paying higher amounts, they find some evidence suggesting that those who pay nothing use it least. Notably, some clinicians have successfully used financial incentives to encourage general healthy behavior, such as smoking cessation (Halpern et al. 2015) and weight loss (Volpp et al. 2008).

The diversity in methods adopted by clinicians to address non-adherence is partially a response to the diversity of conditions treated, and consequently the diversity of treatments. In our experience, economists sometimes view clinical medical trials as a binary situation: take the drug or don't. But in fact, medical conditions vary in a large number of dimensions: whether they are one-time or chronic, the nature of the discomfort that they induce, the time profile of the condition's effects upon the patient, the efficacy of the treatments, the results of noncompliance, and so on. Of course, economic environments feature parallel levels of context-specificity. Our point here is that when economists consider the medical literature on non-adherence, they should be aware that issues of context arise here, too.

Improving Medication Adherence: Traditional Results

What have clinicians learned and surmised based on randomized control trials designed to improve medication adherence? We will focus in this section on what we call "traditional" meta-studies, which covers most meta-studies conducted up to around 2012. These studies exclude investigations of smartphone applications and other mobile telephone-based methods of improving medication adherence, which represent more recent technological innovations. In the next section, we focus on the effectiveness of modern mobile telephones in clinicians' quest to enhance medication adherence. Because our ultimate focus is applying these results to the environments that typically interest economists, which are quite distinct from those considered in the medication non-adherence literature, we focus on providing readers with qualitative results. Those interested in a quantitatively rigorous meta-analysis should consult the meta-studies cited here. We primarily draw upon the work of McDonald, Garg, and Haynes (2002), Peterson, Takiya, and Finley (2003), Kripalani, Yao, and Haynes (2007), Haynes, Ackloo, Sahota, McDonald, and Yao (2008), Zullig, Peterson, and Bosworth (2013), and Nieuwlaat et al. (2014).

An overarching—and somewhat disappointing—conclusion from this literature is that the methods considered exhibit a high degree of context-specificity in their effectiveness, making it difficult to arrive at general conclusions. As mentioned above, this is the result of the huge diversity in medical conditions, and in the treatments that clinicians prescribe in the pursuit of better health outcomes. Consequently, this sobering conclusion should not be considered anomalous. Interestingly, such results parallel the arguments in Levitt and List (2007) concerning the generalizability of experimental results from the lab.

Overall, a slight majority of studies find no significant effect of the interventions being investigated on medication adherence. In fact, it is quite common for some very expensive interventions, such as face-to-face meetings with specialist physicians, to result in no statistically discernable effect upon medication adherence. Yet, it is important to note that experimental power may be a culprit here—most studies are not reporting a treatment effect estimate of precisely zero.

Among the approximately 40 percent of studies that do detect a statistically significant effect, the magnitude is somewhat modest, falling in the range of 4–11 percent. Importantly, detected effects tend to shrink further when one focuses on the relationship between the adherence intervention and clinical outcome, rather than the intermediate relationship between the adherence intervention and the rate of adherence. Moreover, there is no general pattern regarding the comparative effectiveness of narrow interventions, such as focus groups versus email reminders.

Particularly in the case of long-term, chronic medical conditions however, there is a tendency for the most effective interventions to be those based on complex combinations of the basic classes, such as educational sessions at the start, counseling sessions throughout the treatment plan, and a selection of reminder methods, such as telephone calls from nurses and pillboxes.

Finally, we should highlight that the value of studies of medication adherence is limited by a series of flaws in the data-gathering and analysis process. First, datasets tend to be small and experimental designs underpowered, and authors of survey articles typically urge scholars to pay more attention to established best practices in sample size determination. Second, the somewhat inevitable dependence upon self-reported measures of medication adherence, especially in the traditional studies that predate the era of smartphones and remote monitoring, is a considerable source of noise that impedes precise inference. Third, publication bias—the tendency for journal editors to systematically favor studies that report significant results—is a source of upward bias in detected treatment effects, though (as with issues of experimental design), appropriate coordination between scholars and journal editors can eliminate this problem (Young, Ionnidis, and Al-Ubaydli 2008).

Improving Medication Adherence: Smartphone Results

The traditional literature suggests that organic medication adherence rates can be quite modest, and that exogenous interventions tend to have a small effect at best on patient adherence to prescribed medications. Given the dramatic effects that smartphones have had on the nature of many services delivered to consumers, such as banking, dating, ridesharing, and media, there is a sense of optimism that they can also contribute to higher rates of medication adherence.

In particular, smartphone applications have several novel and attractive attributes (Dayer et al. 2013): constant accessibility, the ability to act as a repository of patient- and medication-specific information; a source of education for patients about adherence; and interoperability with existing systems for prescriptions and medical records. Critically, the cost of these features is potentially many orders lower than that of the next-best alternative. For example, if a patient has to convey self-reported adherence to a clinician in oral or written form, there is a considerable time cost of recording the data in the patient's medical file, as compared to the instant integration that a smartphone can offer.

It is too early for the literature to provide rigorous quantitative assessments of the effectiveness of smartphone interventions targeting medication adherence. Existing studies, such as Dayer et al. (2013), focus more on qualitative conjectures about the likely effectiveness of various features. They emphasize the positive role of several features, including the ability to sync adherence data with records housed on servers of healthcare providers. In addition, the ability to track missed and taken doses, not just via patient self-reporting, but also via direct cable or wireless link to various treatments can be important. Further, the ability to provide detailed instructions on complex medications—the value of which can be enhanced by linking to databases that give patients a broader background on their medical conditions and treatments—is invaluable. Finally, such features can address multilingualism, as the technical vocabulary associated with clinical settings can be a challenge for the millions of migrants that live in a country with a language different than their mother tongue.

Another piece of evidence lending insights into the potential efficacy of smartphone technology is Vervloet et al. (2012). They report that, across several studies, electronic reminder devices—which operate in a manner similar to pager systems, allowing for automated or visual reminders—have a substantial, positive, and robust effect on medication adherence. This is interesting evidence because it serves to highlight the potential role that smartphones can play, seeing as electronic reminder devices are in many regards rendered obsolete by smartphones, which can perform all of the same functions, as well as many additional ones, described above.

As an illustration, one smartphone application formally evaluated in a randomized control trial was the WellDoc diabetes management application, which displays medication regimen, provides feedback on patients' blood glucose levels, and tailored management through evidence-based algorithms (Quinn et al. 2008). Compared to a control group, patients showed a significant decrease in HbA1c, a clinical measure of diabetes management, and its success was salient enough to convince insurance companies to subsidize the application, allowing it to be prescribed as part of the treatment.

Nevertheless, the overwhelming majority of quantitative studies of smartphones, such as Strandbygaard, Thomsen, and Backer (2010), Petrie, Perry, Broadbent, and Weinman (2012), Huang et al. (2013), and Finistis, Pellowski, and Johnson (2014), are forced to focus on the simplest intervention that smartphones permit—text message reminders. The comprehensive surveys by Vervloet et al. (2012) and Sarabi, Sadoughi, Orak, and Bahaadinbeigy (2016) concluded that there is robust evidence that text message reminders improve medication adherence, especially in chronic conditions or in patient populations requiring complex medication regimens, such as HIV, asthma, and diabetes. There is evidence that tailoring the messages to the patient yields a larger effect on medication adherence (Kreuter, Farrell, Olevitch, and Brennan 2000), as do interactive messages—for example, ones that require a reply from the patient.

An additional finding from the literature is that reminders are systematically more effective for those who are unintentionally non-adherent, such as older patients who have memory problems, or adolescents who might be preoccupied with their social lives. In contrast, reminders are found to be ineffective for the intentionally non-adherent, which further illustrates the need for smartphone capabilities that go beyond reminders to most fully address non-adherence. Many patients have an increasing interest in accessing health information on smartphone apps over internet sources (Smith 2015), which provides a promising avenue for dissemination of scientific evidence in a user-friendly manner to combat intentional non-adherence due to systematically inaccurate beliefs.

As a whole, the consensus regarding smartphones at this point is that there are sound reasons for optimism, but there remains a need for the accumulation of further evidence. There is the possibility that in the long-term, such interactivity through smartphones might backfire by creating user fatigue, especially in light of the variety of stimuli that smartphones offer (Dennison, Morrison, Conway, and Yardley 2013). One possibility is that future consumers might use a number of wearable electronic devices, perhaps including a device dedicated to medications or to fitness more broadly, rather than bundling so much of their personal information technology into a smartphone. In fact, in clinical settings, the risks posed by smartphones, including their acting as a distraction for patients and clinicians, have driven some medical researchers to propose strict regulations on smartphone usage in conjunction with using specialized alternatives, such as electronic reminder devices (Gill, Kamath, and Gill 2012).

What Can Economists Learn from Medicine?

In the narrow domain of irrational non-adoption, there is much that economists can potentially learn from medical researchers, as the latter group has been rigorously studying an isomorphic problem for decades. In addition, these studies are in a setting where the stakes are significantly higher than those typically encountered by economists. Several insights can be gained from the medical adherence literature.

First, economists should not assume that merely demonstrating the superiority of an alternative mode of behavior to an agent—even from the perspective of the agent's interests—is sufficient for the agent to organically modify their behavior. Admittedly, almost all economists understand this point, as confirmed by the existence of a large and heavily cited behavioral economics literature. However, our sense is that this lesson is sometimes forgotten when economists are solicited by policymakers interested in scaling up their observed findings. Maybe it is due to the understandable excitement of seeing one's research have a profound effect on society—a rare event in the life of most economists. Alternatively, it could be the result of the difficulty of explaining such subtleties to the nonspecialist policymakers and senior civil servants who expressed interest in scaling the findings. Yet exercising restraint is critical at such junctures.

Clinicians have struggled with patient medication adherence rates that average around 50 percent for decades, due to reasons ranging from willful noncompliance stemming from systematically biased beliefs about the effects of medication, to an inadvertent failure to comply caused by forgetfulness or an inability to follow complex treatment plans. Non-adherence can sometimes lead to significant financial costs and physical discomfort, and in some cases, it can cause death, yet these nominally strong incentives are still too weak to motivate "rational" behavior. These reasons for non-adherance closely correspond to the suite of behavioral models that are becoming more common in economics, and the medical literature should make economists more willing to estimate the structural parameters in behavioral models as they seek to scale their results more effectively (DellaVigna, List, and Malmendier 2012; Allcott and Taubinsky 2015).

Second, even as economists acknowledge the apparent prevalence of irrational non-adoption, in their search for countermeasures, they should not expect to find any silver bullets, due to the high levels of context-specificity exhibited by such interventions. This affirms the principle that generalizing results is an imprecise art at best, and that there is no substitute for systematic, incremental research in field contexts that are as close as possible to the target domain. This insight has interesting parallels to the generalizability debate concerning lab experiments (Al-Ubaydli and List 2015a) and calls upon theorists to bridge the important gap between experiment and practice by creating models of generalizability that enhance our fundamental understanding of the where's, why's, and how's of scaling. We trust that this will come down to an understanding of behavioral primitives and how those can be affected as well as learning about features of the environment that attenuate or exacerbate effects of treatment.

This somewhat disheartening result should not be confused with economists having little to learn from the medical literature; learning that interventions that are expected to work are actually unlikely to work constitutes useful information. It potentially saves resources, and allows them to be directed to interventions that are more likely to yield improvements in adherence.

Third, one of the more robust conclusions to emerge from the medical literature is that improvements in adherence are usually the result of complex interventions that combine education, monitoring, and the involvement of other stakeholders. Thus, economists seeking to scale the results of their studies should—after acknowledging the possibility of significant irrational non-adoption—consider skipping simple interventions and going straight to multipronged approaches, especially those that involve exploiting the omnipresence of smartphones, and their interoperability with the electronic systems that underlie the desired modification to behavior.

Moreover, during the evaluation stage, researchers should be careful to focus on the effects of the interventions on the final outcomes associated with the original proposal for modifying behavior, rather than myopically fixating on the effect of the interventions on adoption, which can sometimes be a red herring. Doctors want patients to get better, not to take pills; equivalently, economists should want people to experience superior outcomes, rather than to modify their behavior as an end in and of itself. In this manner, a bit of backward induction at the design stage can go a long way.

More generally, both clinical researchers and economists stand to gain much from applying experimental design best practices. This includes ensuring that their studies have appropriate sample sizes and sufficient power—see List, Sadoff, and Wagner (2011) for simple rules of thumb for optimal experimental design. In addition, replication and sound inference should be emphasized as countermeasures to the problem of publication bias (Young, Ionnidis, and Al-Ubaydli 2008; Maniadis, Tufano, and List 2014).

Finally, we note one area where economists have shown more initiative than medical researchers: the deployment of financial incentives to improve adoption. There are reasons to be skeptical about the effects of such interventions in the clinical domain, but to the best of our knowledge, the tangible evidence in the medical domain is still limited. Also, one medical study conducted by economists (Charness and Gneezy 2009) is cause for tentative optimism about the beneficial role that financial incentives can play in getting people to overcome the cognitive biases that impede modifications to behavior. Thus, while recommending that economists focus on complex interventions rather than simple ones, we make an exception for the use of financial incentives. These should be systematically compared to nonfinancial alternatives whenever possible for each target context.

■ *We are grateful for comments by Gordon Hanson, Enrico Moretti, and Timothy Taylor.*

References

Al-Ubaydli, Omar, and John A. List. 2015a. "On the Generalizability of Experimental Results in Economics." Chapter 20 in *Handbook of Experimental Economic Methodology*, edited by G. R. Frechette and A. Schotter. Oxford Scholarship.

Al-Ubaydli, Omar, and John A. List. 2015b. "Do Natural Field Experiments Afford Researchers More or Less Control Than Laboratory Experiments?" *American Economic Review* 105(5): 462–66.

Al-Ubaydli, Omar, John A. List, and Dana L. Suskind. 2017. "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results." *American Economic Review* 107(5): 282–86.

Allais, Maurice. 1990. "Allais Paradox." In *Utility and Probability*, edited by John Eatwell, Murray Milgate, and Peter Newman, 3–9. Springer.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105(8): 2501–38.

Andreoni, James A., and John H. Miller. 1993.

"Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence." *Economic Journal* 103(418): 570–85.

Ashraf, Nava, James Berry, and Jesse M. Shapiro. 2010. "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia." *American Economic Review* 100(5): 2383–2413.

Benabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70(3): 489–520.

Campbell, Catherine, and Yodwa Mzaidume. 2002. "How Can HIV Be Prevented in South Africa? A Social Perspective." *British Medical Journal* 324(7331): 229–32.

Caplan, Bryan. 2002. "Systematically Biased Beliefs about Economics: Robust Evidence of Judgmental Anomalies from the Survey of Americans and Economists on the Economy." *Economic Journal* 112(479): 433–58.

Carroll, Christopher D., Jody Overland, and David N. Weil. 2000. "Saving and Growth with Habit Formation." *American Economic Review* 90(3): 341–55.

Chang, Eric C., Joseph W. Cheng, and Ajay Khorana. 2000. "An Examination of Herd Behavior in Equity Markets: An International Perspective." *Journal of Banking and Finance* 24(10): 1651–79.

Charness, Gary, and Uri Gneezy. 2009. "Incentives to Exercise." *Econometrica* 77(3): 909–31.

Chopra, Mickey, and Neil Ford. 2005. "Scaling up Health Promotion Interventions in the Era of HIV/AIDS: Challenges for a Rights Based Approach." Health Promotion International 20(4): 383–90.

Dayer, Lindsey, Seth Heldenbrand, Paul Anderson, Paul O. Gubbins, and Bradley C. Martin. 2013. "Smartphone Medication Adherence Apps: Potential Benefits to Patients and Providers." *Journal of the American Pharmacists Association*: 53(2): 172–81.

Deaton, Angus, and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." NBER Working Paper 22595.

Deci, E. L., R. Koestner, and R. M. Ryan. 1999. "A Meta-analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychology Bulletin* 125(6): 627–68.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127(1): 1–56.

Dennison, Laura, Leanne Morrison, Gemma Conway, and Lucy Yardley. 2013. "Opportunities and Challenges for Smartphone Applications in Supporting Health Behavior Change: Qualitative Study." *Journal of Medical Internet Research* 15(4).

Duflo, Esther, Michael Kremer, and Jonathan Robinson. 2011. "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya." *American Economic Review* 101(6): 2350–90.

Farnan, J. M., L. A. Petty, E. Georgitis, S. Martin, E. Chiu, M. Prochaska, and V. M. Arora. 2012. "A Systematic Review: The Effect of Clinical Supervision on Patient and Residency Education Outcomes." *Academic Medicine* 87(4): 428–42.

Ferraro, Paul J., and Michael K. Price. 2013. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics* 95(1): 64–73.

Field, Erica, and Rohini Pande. 2008. "Repayment Frequency and Default in Microfinance: Evidence from India." *Journal of the European Economic Association* 6(2–3): 501–09.

Finitsis, David J., Jennifer A. Pellowski, and Blair T. Johnson. 2014. "Text Message Intervention Designs to Promote Adherence to Antiretroviral Therapy (ART): A Meta-analysis of Randomized Controlled Trials." *PLOS One* 9(2): e88166.

Fryer, Roland G., Jr. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics* 126(4): 1755–98.

Fryer, Roland G., Jr., Steven D. Levitt, and John A. List. 2015. "Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights." NBER Working Paper 21477.

Garrett, Laurie. 2007. "The Challenge of Global Health." *Foreign Affairs* 86(1): 14–38.

Gill, Preetinder S., Ashwini Kamath, and Tejkaran S. Gill. 2012. "Distraction: An Assessment of Smartphone Usage in Health Care Work Settings." *Risk Management and Healthcare Policy* 5: 105–14.

Giuffrida, Antonio, and David J. Torgerson. 1997. "Should We Pay the Patient? Review of Financial Incentives to Enhance Patient Compliance." *BMJ* 315(7110): 703–07.

Gosnell, Greer K., John A. List, and Robert Metcalfe. 2016. "A New Approach to an Age-Old Problem: Solving Externalities by Incenting Workers Directly." NBER Working Paper 22316.

Halpern, S. D., Benjamin French, Dylan Small, Kathryn Saulsgiver, Michael O. Harhay, Janet Audrain-McGovern, George Loewenstein, Troyen A. Brennan, David A. Asch, and Kevin G. Volpp. 2015. "Randomized Trial of Four Financial-Incentive Programs for Smoking Cessation." *New England Journal of Medicine* 372(22): 2108–17.

Haynes, R. B., E. Ackloo, N. Sahota, H. P. McDonald, and X. Yao. 2008. "Interventions for Enhancing Medication Adherence." *Cochrane*

Database Systematic Reviews 16(2).

Hibbard, Judith H., Jean Stockard, and Martin Tusler. 2005. "Hospital Performance Reports: Impact on Quality, Market Share, and Reputation." *Health Affairs* 24(4): 1150–60.

Hossain, Tanjim, and John A. List. 2012. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Science* 58(12): 2151–67.

Huang, H. L., Y. C. Li, Y. C. Chou, Y. W. Hsieh, F. Kuo, W. C. Tsai, S. D. Chai, B. Y. Lin, P. T. Kung, and C. J. Chuang. 2013. "Effects of and Satisfaction with Short Message Service Reminders for Patient Medication Adherence: A Randomized Controlled Study." *BMC Medical Informatics and Decision Making* 13(1): 127.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1991. "The Endowment Effect, Loss Aversion, and Status Quo Bias: Anomalies." *Journal of Economic Perspectives* 5(1): 193–206.

Klemperer, Paul. 1987. "Markets with Consumer Switching Costs." *Quarterly Journal of Economics* 102(2): 375–94.

Kreuter, Mathew, David Farrell, Laura Olevitch, and Laura Brennan. 2000. *Tailoring Health Messages: Customizing Communication with Computer Technology.* New York: Routledge.

Kripalani, S., X. Yao, and R. B. Haynes. 2007. "Interventions to Enhance Medication Adherence in Chronic Medical Conditions: A Systematic Review." *Archives of Internal Medicine* 167(6): 540–50.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112(2): 443–77.

Lavorini, F., A. Magnan, J. C. Dubus, T. Voshaar, L. Corbetta, M. Broeders, R. Dekhuijzen, J. Sanchis, J. L. Viejo, P. Barnes, C. Corrigan, M. Levy, and G. K. Crompton. 2008. "Effect of Incorrect Use of Dry Powder Inhalers on Management Patients with Asthma and COPD." *Respiratory Medicine* 102(4): 593–604.

Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21(2): 153–74.

List, John A., Sally Sadoff, and Mathis Wagner. 2011. "So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design." *Experimental Economics* 14(4): 439–57.

Lu, C., C. M. Michaud, K. Khan, and C. J. Murray. 2006. "Absorptive Capacity and Disbursements by the Global Fund to Fight AIDS, Tuberculosis and Malaria: Analysis of Grant Implementation." *Lancet* 368(9534): 483–88.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review* 104(1): 277–90.

McDonald, H. P., A. X. Garg, and R. B. Haynes. 2002. "Interventions to Enhance Patient Adherence to Medication Prescriptions: Scientific Review." *Journal of the American Medical Association* 288(22): 2868–79.

Nieuwlaat, R., N. Wilczynski, T. Navarro, N. Hobson, R. Jeffery, A. Keepanasseril, T. Agoritsas, N. Mistry, A. Iorio, S. Jack, B. Sivaramalingam, E. Iserman, R. A. Mustafa, D. Jedraszewski, C. Cotoi, and R. B. Haynes. 2014. "Interventions for Enhancing Medication Adherence." *Cochrane Database Systematic Review* 20(11).

Orr, R. D., N. Pang, E. D. Pellegrino, and M. Siegler. 1997. "Use of the Hippocratic Oath: A Review of Twentieth Century Practice and a Content Analysis of Oaths Administered in Medical Schools in the U.S. and Canada in 1993." *Journal of Clinical Ethics* 8(4): 377–88.

Paulsell, Diane, Sarah Avellar, Emily Sama Martin, and Patricia Del Grosso. 2010. *Home Visiting Evidence of Effectiveness Review: Executive Summary.* Princeton, NJ: Mathematic Policy Research.

Peterson, A. M., L. Takiya, and R. Finley. 2003. "Meta-analysis of Trials of Interventions to Improve Medication Adherence." *American Journal of Health-System Pharmacy* 60(7): 657–65.

Petrie, K. J., K. Perry, E. Broadbent, and J. Weinman. 2012. "A Text Message Programme Designed to Modify Patients' Illness and Treatment Beliefs Improves Self-Reported Adherence to Asthma Preventer Medication." *Journal of Health Psychology* 17(1): 74–84.

Quinn, C. C., S. S. Clough, J. M. Minor, D. Lender, C. C. Okafor, and A. Gruber-Baldini. 2008. "WellDoc Mobile Diabetes Management Randomized Controlled Trial: Change in Clinical and Behavioral Outcomes and Patient and Physician Satisfaction." *Diabetes Technology and Therapeutics* 10(3): 160–68.

Raikes, Helen, Beth L. Green, Jane Atwater, Ellen Kisker, and Jill Constantine. 2006. "Involvement in Early Head Start Home Visiting Services: Demographic Predictors and Relations to Child and Parent Outcomes." *Early Childhood Research Quarterly* 21(1): 2–24.

Roggman, Lori A., Gina A. Cook, Carla A. Peterson, and Helen H. Raikes. 2008. "Who Drops out of Early Head Start Home Visiting Programs?" *Early Education and Development* 19(4): 574–99.

Sarabi, Roghayeh Ershad, Farahnaz Sadoughi, Roohangiz Jamshidi Orak, and Kambiz Bahaadinbeigy. 2016. "The Effectiveness of Mobile Phone Text Messaging in Improving Medication Adherence for Patients with Chronic Diseases: A

Systematic Review." *Iranian Red Crescent Medical Journal* 18(5).

Shearer, Bruce. 2004. "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment." *Review of Economic Studies* 71(2): 513–34.

Simon, H. A. 1972. "Theories of Bounded Rationality." In *Decision and Organization*, edited by C. B. McGuire and Roy Radner, 161–76. Amsterdam: North Holland Publishing Company.

Smith, Aaron. 2015. *U.S. Smartphone Use in 2015.* Washington, DC: Pew Research Center.

Strandbygaard, U., S. F. Thomsen, and V. Backer. 2010. "A Daily SMS Reminder Increases Adherence to Asthma Treatment: A Three-Month Follow-up Study." *Respiratory Medicine* 104(2): 166–71.

Studdert, David M., Michelle M. Mello, Atul A. Gawande, Tejal K. Gandhi, Allen Kachalia, Catherine Yoon, Ann Louise Puopolo, and Troyen A. Brennan. 2006. "Claims, Errors, and Compensation Payments in Medical Malpractice Litigation." *New England Journal of Medicine* 354(19): 2024–33.

Telyukova, Irina A., and Randall Wright. 2008. "A Model of Money and Credit, with Application to the Credit Card Debt Puzzle." *Review of Economic Studies* 75(2): 629–47.

Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112(1): S164–87.

Venkatraman, A., N. Garg, and N. Kumar. 2015. "Greater Freedom of Speech on Web 2.0 Correlates with Dominance of Views Linking Vaccines to Autism." *Vaccine* 33(12): 1422–25.

Vervloet, M., A. J. Linn, J. C. van Weert, D. H. de Bakker, M. L. Bouvy, and L. van Dijk. 2012. "The Effectiveness of Interventions Using Electronic Reminders to Improve Adherence to Chronic Medication: A Systematic Review of the Literature." *Journal of American Medical Informatics Association* 19(5): 696–704.

Volpp, K. G., L. K. John, A. B. Troxel, L. Norton, J. Fassbender, and G. Loewenstein. 2008. "Financial Incentive-Based Approaches for Weight Loss: A Randomized Trial." *Journal of the American Medical Association* 300(22): 2631–37.

Watt, Richard, Francisco J. Vazquez, and Ignacio Moreno. 2001. "An Experiment on Rational Insurance Decisions." *Theory and Decision* 51(2–4): 247–96.

Young, Neal S., John P. A. Ioannidis, and Omar Al-Ubaydli. 2008. "Why Current Publication Practices May Distort Science." *PLOS Medicine* 5(10): e201.

Zullig, L. L., E. D. Peterson, and H. B. Bosworth. 2013. "Ingredients of Successful Interventions to Improve Medication Adherence." *Journal of the American Medical Association* 10(24): 2611–12.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 145–162

How Food Banks Use Markets to Feed the Poor

Canice Prendergast

Imagine that someone gave you 300 million pounds of food and asked you to distribute it to the poor—through food banks—all across the United States. The nonprofit Feeding America faces this problem every year. The food in question is donated to Feeding America by manufacturers and distributors across the United States. As an example, a Walmart in Georgia could have 25,000 pounds of excess tinned fruit at one of its warehouses and give it to Feeding America to distribute to one of 210 regional food banks. How should this be accomplished?

This is a problem where regular markets are off the table: Feeding America does not sell food to the food banks. Instead, Feeding America has to find some other way to satisfy its desire for food to go where it is needed most. One way would be to simply assign food to each food bank. This is how most nonprofits allocate resources. However, a field of economics—often associated with the Nobel Prize winning contributions of Al Roth—has been aimed at designing mechanisms so that outcomes in such nonmarket settings can better reflect what consumers want. This area of research has made enormous advances, both theoretical and practical, in problems such as the allocation of children to schools, kidneys to patients, and medical students to hospital residencies (for examples, see Abdulkadiroglu, Pathak, and Roth 2005, 2009; Budish and Cantillon 2012; Roth 1984, 2008; Roth and Peranson 2004; Roth, Sönmez, and Ünver 1999). This paper tells the story of

■ *Canice Prendergast is the W. Allen Wallis Distinguished Service Professor of Economics, University of Chicago Booth School of Business, Chicago, Illinois. His email is canice.prendergast@chicagobooth.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.145 doi=10.1257/jep.31.4.145

a market design innovation at Feeding America in 2005. Specifically, the author was part of a group that designed a mechanism through which Feeding America transitioned from a centralized allocation system, in which food banks (implicitly) queued for food, to a system in which they bid daily in auctions for truckloads of food using a "fake" currency that the nonprofit designed.

A central focus of the market design literature is designing allocation procedures or "mechanisms" that credibly reveal the preferences of consumers. There are typically two ways to do this. The first is to ask consumers to rank a set of possible outcomes. For example, a student lists which school she likes best, second best, and so on on—or a budding medical resident ranks hospitals. Through appropriate choice of mechanism, the consumer finds it in her interest to report truthfully, and is then efficiently "matched" to an outcome. The second way is to use a more standard market setting where goods have prices, but where participants can only use a specialized currency to buy the goods in the allocation system. Perhaps the most celebrated example of this approach is Radford's (1945) classic description of how cigarettes acted as currency in German prisoner-of-war camps towards the end of World War II. Specialized currencies were also common during the US Depression: the typical case being a company town where wages paid in scrip could be redeemed in the company store.[1] Another commonly cited example here is course choice among MBA students who are given a fixed budget of points to bid for oversubscribed courses (Budish 2011; Budish and Cantillon 2012). A small number of universities have used similar mechanisms for allocating undergraduate courses. Here students face a budget much like the standard market setting but that budget can only be used on courses.

As one might imagine, many of those involved in food banks are skeptical of markets in general, and initially many had severe reservations about a market-based approach. As one example, John Arnold, a member of the redesign group who was for many years Director of the Feeding America Western Michigan Food Bank said to me once near the start of the process: "I am a socialist. That's why I run a food bank. I don't believe in markets. I'm not saying I won't listen, but I am against this." This paper describes how 300 million pounds of donated food are now allocated between regional food banks each year and how the initial reservations of some involved in the design process were overcome. I will use basic economic ideas to show how a market—with appropriate safeguards—was constructed to allow food banks to effectively express their preferences in ways not possible under a (well-intentioned) centralized system (for a more formal analysis, see Prendergast 2017).

[1] Other specialized currencies included community initiatives where (often unemployed) individuals provided services and goods for each other, using currencies that they created. Gatch (2008) discusses the range of such initiatives. Such local communitarian currencies continue to exist in small scale today, such as "Brixton pounds" in London (described at http://brixtonpound.org).

Before 2005

Until 2005, Feeding America had a method of allocating resources that is fairly common among not-for-profits: a "wait your turn" system, where it gave out food based on a food bank's position in a queue. The queue was determined by the amount of food that a food bank had received compared to a measure of need called the "Goal Factor," which is (roughly) the number of poor in a food bank's area compared to the national average. The formula is more nuanced than a simple head count, as it distinguishes between usage rates for those below the poverty line, between 100 and 125 percent of the poverty line, and between 125 and 185 percent.

When a food bank's position in the queue was high enough, it would receive a call or email from Feeding America to say that it had been assigned a "load." The load had to be collected from the donor, and food banks were (and remain) liable for transportation costs. The food bank had 4–6 hours to say "yes" or "no." After a food bank was offered food, its position in the queue would be recalculated, as its measure of food received relative to need would change. If it turned down the offer, the load would go to the next food bank in the queue. This mechanism had been used since the late 1980s, and it allocated 200–220 million pounds of food each year from 2000 to 2004. Feeding America did not distinguish much between different kinds of food, so that each food bank on average got a similar product mix from them (though randomly a food bank could get lucky or unlucky in whether it would get food that was popular among participants).

The objective of this centralized assignment mechanism was to offer an equal number of pounds of food to each *client* of the food bank (this outcome occurred because a 1 percent increase in Goal Factor meant 1 percent more clients, and the mechanism gave 1 percent more food). On average, the mechanism indeed accomplished this goal: regression results show that a 1 percent increase in Goal Factor was associated with a 1.01 percent increase in pounds of food. This way of handing out food works well if all food banks *should* get the same amount and kind of food. But by 2004, Feeding America had concerns that this was not the case.

The problem arises because Feeding America allocates only about one-quarter of all the food that food regional banks receive, with the rest coming directly to the food banks from manufacturers, distributors, grocery stores, and so on. But Feeding America knows little about the other three-quarters of what food banks have. Some food banks—sometimes called the "food rich"—have better contacts with potential donors and have larger amounts of food than the "food poor," who have little access to distributors and manufacturers. Moreover, food banks vary not only in how much food they have, but also in what kind. For example, a food bank's existing inventory of other food may already be heavily weighted towards dairy products, and its residual needs are for other kinds of food.

In this context, the queuing system faces two major problems. First, a food bank might get food that a different food bank values more. If a food bank already has enough for its clients at a point in time, any extra may even go to waste. This concern is exacerbated by spoilage issues: for example, fresh produce is often only donated close to its expiration date. An example that routinely cropped up with the committee was when the Idaho Food Bank was offered potatoes even though they already had a warehouse full of them. Another reason for spoilage is the need for refrigeration. For example, sending eggs or milk to a food bank that lacks excess refrigeration capacity (because its fridges are currently close to full) likely results in those products not being used. Food-rich banks are often unable to efficiently use more of the staples like produce and dairy products for the reason above. However, additional stocks of some highly valued (and relatively rare) foods, like cereal and pasta, are always of use. On the other side, the food-poor banks are less fussy about what kind of food they get (though they still like cereal and pasta more), as they don't have enough of anything. As a result, an equal mix of food across food banks is unlikely to be efficient. Feeding America was aware of this issue, but did not know much about actual food bank inventories. While it may have suspected that, say, Los Angeles had more food per client from other sources than did Idaho, it lacked any data on which to base a (politically legitimate) policy that could respond to this situation.

The second problem with the old queuing system is that Feeding America could only offer food to one food bank at a time. A typical scenario was that a distributor had a truckload of excess food sitting in its warehouse or dock, and offered it to Feeding America. If the donation was accepted by Feeding America, it would contact the food bank at the top of the queue and offer it to them. The food bank had four to six hours to say "yes" or "no," and some of that time was inevitably spent on practical details like checking existing inventory, seeing if transportation to pick up the donation was feasible, and so on. Another food bank would be offered the load only if the first food bank demurred. This process implied that Feeding America could only offer the load to a small number of food banks before either the donor would become upset over the load being left on its dock for a long time, or the food would spoil.[2]

The Choice System

With this backdrop, Feeding America put together a committee to make recommendations on the redesign of its allocation system. The group consisted of eight food bank directors, three staff from Feeding America, and four University of Chicago faculty.[3] The group quickly realized that food banks had such variety in needs that it would be difficult to design any efficient system with Feeding America

[2] An additional, smaller, problem is due to randomness in what kind of food comes up when it is a food bank's turn. For example, suppose that a food bank gets lucky and is assigned cereal twice in a month, while another food bank gets produce twice. (As we will see, cereal is much more valuable to food banks than fresh produce.) While this result does generate inequality, it seems a relatively small issue compared to the other two.

[3] The committee consisted of John Alford, John Arnold, Al Brislain, Bill Clark, Phil Fraser, Maria Hough, Mike Halligan, Brenda Kirk, Rob Johnson, Susannah Morgan, Steve Sellent, Roger Simon, Harry Davis, Don Eisenstein, Robert Hamada, and the author.

deciding what was best for individual food banks. After considerable discussion about alternatives and practical details, Feeding America introduced what is called the Choice System, a market-based mechanism with food banks bidding on truckloads of food. This system involves twice-daily first-price auctions, the ability to borrow and save, fractional bidding, the possibility of negative prices for loads that are not wanted by food banks, and the capacity for food banks to put their own food up for auction on the system. As we will see, the Choice System seems to have alleviated many of the issues above.

Feeding America was well aware that having food banks pay for food (rather than simply giving it to them) would more credibly reveal demand. Indeed, these food banks distribute food to local food pantries and soup kitchens and often require *them* to buy food from the food bank. However, Feeding America feared that with a market-based system, a food bank's budget would be based on its fund-raising skills and whether it was based in an area that was wealthier or denser. Given such differences, using money could exacerbate inequality across food banks. This concern was sufficiently important to Feeding America that it used the queuing system to ensure that the poorest areas are offered adequate food compared to their richer counterparts, despite the obvious drawbacks of such a system.

The redesign group met for over a year before converging on the Choice System. A central feature was the creation of a specialized currency called *shares* that are used to purchase food. By using fake money, Feeding America could set a food bank's budget for food based on measures of need rather than fund-raising capacity.

However, when the idea of a "market" was introduced as an alternative to waiting in line, it met with considerable resistance. Food banks exist to serve the marginalized, who are often those that the market economy has left behind. The preferences of food bank directors often reflect that concern about marginalization. But as the committee discussions progressed, it became clear that many of the concerns of the food bank directors on the design committee about a market-based system were not of a broadly philosophical kind, but rather originated in a fear that the details of markets often benefit the strong at the expense of the weak. As a result, many of the more detailed features described below were particularly aimed at ensuring that *smaller* food banks, typically with fewer resources and manpower, would not be harmed relative to their larger counterparts, where there are often dozens of workers or volunteers.

Budgets

Remember that a primary concern was that access to food should depend on need. This goal was implemented by allocating initial budgets of shares to the food banks in proportion to Goal Factor, thereby aligning capacity to spend with Feeding America's perception of a food bank's need. Shares could not be traded for real money nor used for anything other than the items on the auction market described below. Balances did not depreciate, nor was there an interest rate on savings. Budgets are replenished each evening by redistributing the spent shares to the food banks according to the rules described below.

Demand

On any given day, approximately 50 truckloads of food are offered to food banks (a truckload averages about 25,000 to 30,000 pounds). Food banks bid on truckloads of food using their shares. Sealed bid first-price auctions occur twice per day, from Monday to Friday.[4] Bidding closes at noon and 4 pm Central Standard Time. All food for each bidding cycle is posted at least two hours beforehand.

Several details of how bidding was designed can be understood as ways of leveling the playing field across food banks. For example, one concern was that under a continuous auction, some food banks (typically the larger ones) could dedicate a staff person to the bidding process and those food banks could wait until the last minute and "snipe." Smaller food banks, which may only have a handful employees on site, could not do this. This inequality of access was averted by sealed bids, with all food posted at least two hours beforehand.

Another instrument used to level the playing field was the option to use fractional bidding. Larger food banks are big enough to use a truckload of a desired food, whereas their smaller counterparts may only be able to effectively distribute say a quarter of a truckload. To alleviate this disadvantage for smaller food banks, they have the opportunity to bid jointly for items. Here two (or more) food banks coordinate and agree to split a truckload offering.

When the Choice System came into being, many food bank directors had never bid online for anything. A concern was that some food banks—again most likely the smaller ones—would find bidding so intimidating that they would largely withdraw from the allocation process. This concern was alleviated by Feeding America giving the option to delegate bidding to an employee of Feeding America, where a food bank could simply outline in broad terms its needs to that person.

The system also allows credit. Credit was implemented because of a concern that the smaller entities might never receive the most desired products, because a truckload of the most desired goods could sell for more than their share balance. Food banks below median Goal Factor, which tend to be the smaller ones, can use short-term credit to increase their balances to at least the estimated cost of a highly desired item. They pay off those debts with at least half their future allocations of shares—meaning the nightly redistribution of shares used on that day (described below)—until the debt is paid off. In this way, they cannot continue to accumulate credit. There is no interest rate on these debts.

The system allows food banks to bid negative prices. Some loads are not very desirable to food banks. Under the old queuing system, a food bank could say "no" to an offered lot. Food banks have a variety of sensible reasons to say "no": the food could be undesirable given its clients' needs, it might not have an available truck,

[4] The group went back and forth on what price would be paid by winning bids. A desire to minimize strategic considerations led to some members (well, the author anyway) arguing for a second-price auction (the highest bidder wins but pays only the amount of the second-highest bid), but the sense among the participants in the process was that the clarity of "you pay what you bid" was more important. As a result, a first-price auction was chosen.

or the value of the food might not be worth the transportation cost. However, if a food bank turned a load down, it was still counted against its position in the queue as if the load had been accepted. This may seem strange: Why penalize a food bank for refusing to take food that it does not want? The underlying reason was based on maintaining donor relations. Donors typically want excess food removed from their warehouses for a variety of reasons: to free up storage space, for tax reasons, and so on. As such, there are pressures on Feeding America to remove food quickly, and that pressure was sometimes felt by the affiliates. Placing this undesired food was a source of tension under the old system. To facilitate the movement of these kinds of goods, the Choice System allows for negative prices. These are called "bonus share": Food banks could bid negative shares for loads (up to a limit of –2,000 shares per load), which means that the food bank received additional shares for agreeing to pay the costs of picking up a load.

Finally, the Choice System includes a mechanism called the Fairness and Equity Committee for overriding the allocation rule in extreme circumstances. Under the old system, Feeding America at times would use its discretion to divert food to some food banks if they realized that they had needs that were not addressed under the Goal Factor formula. As part of the Choice System, a Fairness and Equity Committee (staffed by three food bank directors) would take appeals from food banks for greater allocations of shares based on some unobserved factor, and decide its merits on a case-by-case basis. As one extreme example, Hurricane Katrina hit New Orleans soon after the redesign committee's deliberations. A less-extreme example might be the closure of a major local manufacturing plant.

Supply

A significant issue throughout the deliberations was finding ways to increase the supply of food reaching the poor. One piece of this was a new source of food called Maroon pounds. This is food that an individual food bank already has, perhaps from another source, but for which it may not be the highest-value user. The Choice System allows food banks to place this food on the internal market. These loads are bid on in exactly the same way as other products, but here the shares from the winning bid are transferred to the seller rather than redistributed to all food banks. Negative prices are not available for Maroon Pounds.

Money Supply

Food banks bid with a constructed form of money, and one design aspect that consumed much of the committee's time was an appropriate money supply rule. An objective converged upon relatively early in the process was to ensure that prices remain constant if demand and supply conditions do not change. This objective mattered because it helped food banks know how much to bid: specifically, observing historical prices of a good would give a food bank a good indication of a reasonable price.

This goal was implemented in two ways beyond the initial share allocation. First, consider short-run money supply calibration: Over a typical day, shares are spent and money balances are drawn down. Say that aggregate purchases total 10,000 shares. These 10,000 shares are then recirculated at midnight of that day, and a food bank's slice of this pie is its Goal Factor relative to the sum of all Goal Factors. As such, the flow of resources to food banks also depends on this measure of need. In this way, the money supply additionally remains constant over the short run. Second, over the longer run, the supply of food to the system could change. Suppose that from one year to the next, supply rises by 5 percent. To maintain constant prices, all else equal, the supply of shares is changed in proportion to that increase in total number of pounds of food in the system. (It was deemed too complicated to make this depend on changes in the quality of food.)

The Website

Food banks bid online. The web page lists available offerings: kind of food, its weight, location, and any other conditions. Bidders simply type in their sealed bid. (In order to help participants become comfortable with the online setting, the system was used as a test run for three months before it went live, where the participants would simulate bidding.) Two such screens will be seen each day, one for the offerings at noon and the other for the 4 pm auction. Outcomes are transmitted by email to all bidders immediately at the close of the auction.

Outcomes

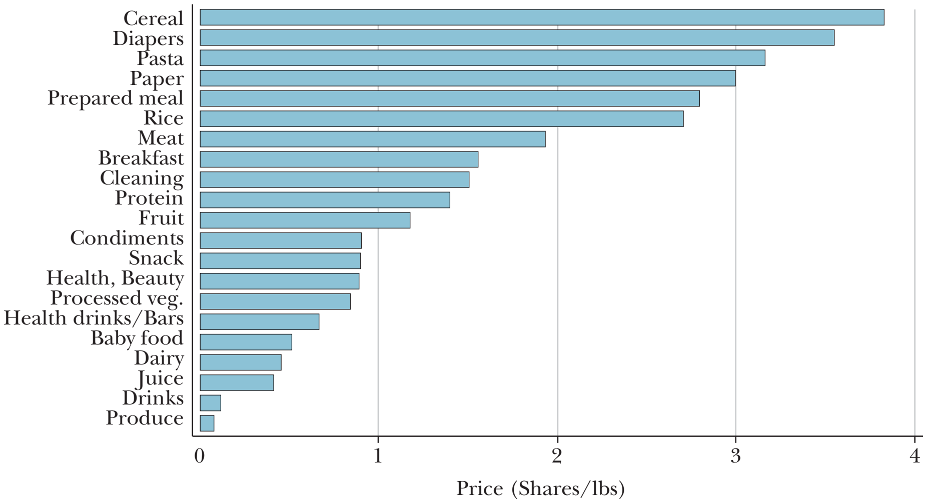
We now turn to how food banks responded to the new allocation mechanism, using data from its introduction on July 1, 2005, to the end of 2011.[5] We begin by considering the most general source of gain from a market: that it allows consumers to express their unknown preferences. We do this by identifying the extent to which outcomes differ from the old way of assigning food. We additionally show that the Choice System resulted in the food poor spending more of their "money" than the food rich, once again redistributing resources to those most in need. Finally, we show how the Choice System system has induced more supply of food through Maroon pounds. Before beginning, it is worth noting that almost any kind of food can be offered to the system—fruit, vegetables, dairy, pasta, rice, meat, and prepared meals. Nonfood items such as health care or beauty products can be offered as well (particularly valuable are paper plates and plastic cutlery, primarily used by soup kitchens). Yet almost half of pounds are either produce or beverages, and as will become clear below, these are the least-desired foods.

[5] The data used here come in two forms. For some exercises, aggregates will be provided both before and after the change to the choice system, from 1999 to 2011. Analysis of what happened after the changeover derive from aggregating data on 64,570 auctions from 2005 to 2011.

Figure 1

The Average Price of a Pound of Food by Food Type, 2005–2011

(the price of the median good is normalized to one)



Note: Figure 1 shows how average prices vary by food type. The numbers have been normalized so that the median good has a pseudo-price of 1.

Reallocation of Demand

To show reallocative benefits of the Choice System, we begin by documenting that food varies wildly in its desirability. Remember, in the older queuing approach, Feeding America treated all pounds of food as equal. With the bidding system, some food banks have chosen to buy mainly large quantities of cheap food, while others buy smaller quantities of more expensive food. Moreover, some food-rich banks never spend all their shares, which benefits the food poor. (A food bank can accumulate shares until they get 200,000 shares, at which point, Feeding America gives them no more because it seems they don't really need them.)

Some foods are valued more than others, which is apparent in how the price of a pound of food varies enormously. Figure 1 shows how average prices vary by food type. The numbers have been normalized so that the median good has a pseudo price of 1. At the cheaper end of the distribution, produce sells for only 7.7 percent of the price of the average good, and beverages trade for 11.6 percent. On the other hand, cereal, diapers, and pasta are the most desired categories, and trade for over three times the price of the average good. To put this concretely, a food bank can buy 49 pounds of produce for the price of a single pound of cereal.

These price ratios are often wildly different than those one would see in a supermarket because they reflect the residual demand of food banks after taking account of all the other food that they have. Trading almost 50 pounds of produce for one pound of cereal does not necessarily tells us that food banks do not like produce, but rather that they already have so much of it from other sources that their marginal valuation of it is close to zero. In this sense, the extreme prices tell us how far the food supplies of food banks are from the mix of foods they desire.

Of course, prices vary for reasons beyond the broad categories used here: for instance, based on quality of product within a category, whether the donation is at a convenient location, whether it is a time of the year when the product is more or less available, and so on. From its inception in June 2005 to December 2011, a food bank on average received three to four pounds of food per share. However, there is enormous variation. For almost 50 percent of auction outcomes, a food bank received 20 pounds per share, and in 25 percent of cases, it received at least 100 pounds of food. About 5 percent of prices are negative. Those goods with negative prices are typically loads of produce or carbonated beverages. At the other extreme, in 10 percent of cases, the buyer got two pounds of food per share or less.

The Sorting of Food Banks

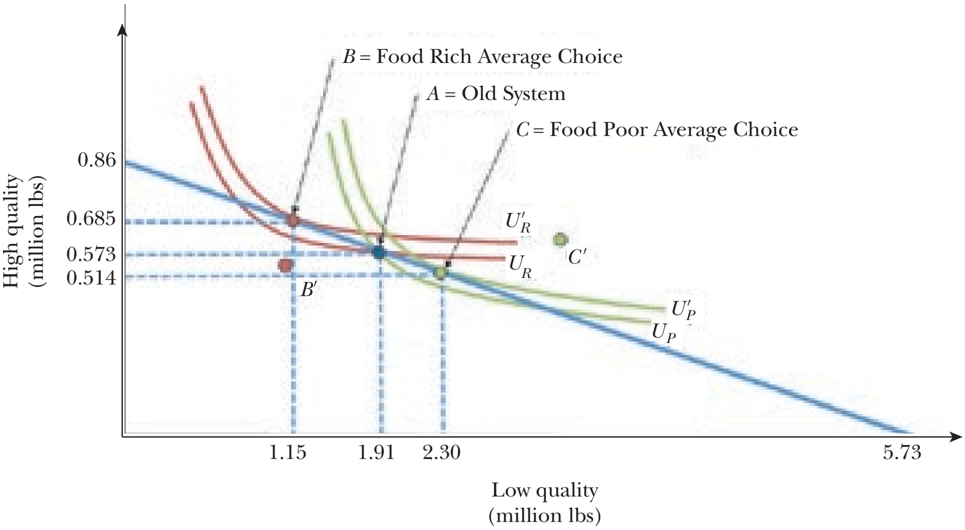
Showing that goods vary in their desirability does not necessarily tell us anything about the value of the new system unless food banks vary in what they choose. Here we show variation on the quality–quantity tradeoff: do food banks spend their money on a small amount of expensive goods or on large quantities of cheaper foods?

To do this, we create a binary categorization of goods. The goods in Figure 1 that are more expensive than the average good are denoted "high quality," and those less expensive than the average good are "low quality." Specifically, all foods below Fruit in Figure 1 are low quality, and the others high quality. Low-quality food accounts for about 65 percent of the pounds of food, but 25 percent of expenditures.[6] We also offer a binary categorization of buyers: we denote any food banks that buy more of the expensive goods than average as "food rich," and a food bank that buys less of the expensive goods as "food poor." In effect, this allows us to compare the average net buyer of expensive goods to the average net buyer of cheap goods. To compute these averages, we weight by the Goal Factor of the food bank, so that we do not overcount smaller food banks compared to larger ones. (This adjustment has the effect of making food bank clients the unit of analysis rather than the food bank per se.) With this binary classification, we can cleanly show sorting on the quality–quantity dimension.

To understand magnitudes, for the period 2005 to 2011, each food banks wins an average of 2,483,000 pounds of food every year, of which 1,910,000 are low quality and 573,000 are high quality. This is given by point *A* in Figure 2. This would be the outcome under the old allocation system. However, the food rich and food poor have different preferences between high-quality and low-quality food; in particular,

[6] This calculation weights the goods in each category by price. For example, if a pound of snacks sells for 20 pounds of produce, it receives 20 times the weight. We price-weight the goods so we can treat this as a two-dimensional problem for expositional purposes with a single budget line.

Figure 2
The Reallocation of Demand



Note: For the period 2005 to 2011, each food banks wins an average of 2,483,000 pounds of food every year, of which 1,910,000 are low quality and 573,000 are high quality. This is given by point *A*. Now let the food banks trade at the equilibrium prices given by the slope of the budget line. The food rich can now attain point *B* (where its indifference curve U'_R is tangent to the budget line), while the food poor can be a net seller of the high-quality goods to reach point *C* (with indifference curve U'_P). Empirically, the food rich don't spend as high a proportion of their shares as do the food poor. But if the food rich reside inside their budget line, the food poor can reach a higher budget line. Points *B′* and *C′* show actual choices of purchased composite low-quality and high-quality goods. The food rich end up at a point interior to the budget line. In turn, this allows the food poor to choose a point beyond the budget line.

the food rich already have a lot of the low-quality staples per client compared to the food-poor food banks, and more may go to waste. Now let the food banks trade at the equilibrium prices given by the slope of the budget line: empirically, food banks trade almost 7 low-quality pounds for a single high-quality pound. The food rich can now attain point *B* (where its indifference curve U'_R is tangent to the budget line), while the food poor can be a net seller of the high-quality goods to reach point *C* (with indifference curve U'_P). Choice offers gains to both parties.

To provide an empirical estimate of points *B* and *C*, we need to consider a time frame. Here we offer two time frames in order to measure what food banks do both over the long run and over a shorter time frame. First consider the average annual choice made by a food bank over the first five years of the system. Over this long time frame, what kind of food does each type of food bank choose to buy, and how different is it from what they were given before? To isolate just the sorting of food banks on the quality dimension, we first assume that savings do not differ between the two kinds of food banks.[7] (We deal with savings below). Then instead of getting 1,910,000 pounds of low-quality food (point *A*), the food rich get 1,150,000: they give up over 700,000 pounds of that food to increase their purchases of high-quality foods by 100,000 pounds (point *B*). By contrast, the typical food-poor food bank consumes 2,300,000 pounds of low-quality food: they buy 400,000 extra pounds of low-quality food by giving up 60,000 pounds of high-quality food (point *C*).

These numbers measure the average choice made by a food bank over the entire five years. This is a very conservative measure of gains from being able to choose, because, for example, the food rich do not choose 1,150,000 pounds of food every year, nor indeed proportionately each month. Instead, they vary their demands based either on what their clientele want, what other food they happen to have, refrigeration capacity, and so on.

Efficiency Gains from Short-Run Choices

To measure the gains from short-run adjustments, we posit a time period over which food banks seek balance in the kind of food that they receive. Consider a time frame of just two months. (For time intervals much shorter than two months, there is likely to be randomness in whether a food bank happens to win or lose an auction, which obviously does not reflect preferences of the food bank.) Here we denote a food bank as "food rich" if during that two months it spends more per client than does the average food bank on expensive food. (Notice that food banks in the "food rich" category over a two-month time horizon are not necessarily the same as the ones in that category on average over the five-year period.)

The two-month time horizon shows much more transitory variation. The banks that are food poor over a two-month horizon receive .62 million pounds of low-quality food. We multiply by 6 to get 3.72 million pounds of low-quality food per year for banks that are food poor over a two-month horizon. This is much higher than the 2.30 million pounds per year that food-poor banks received when the definition of "food poor" is based on a five-year average. Conversely, the banks that are food rich according to the two-month estimates received only 0.81 million pounds of the less-expensive food annually, compared with the 1.15 million pounds food rich banks received when this category is defined by the five-year average. Said another way, the long-run results (discussed earlier and shown as points *B* and *C* in Figure 2) involved the food poor increasing consumption of less-expensive food by 20 percent, and the food rich reducing consumption of this food by 40 percent. Adding short-run variation in demand changes the first number to 94 percent and the second to 58 percent. Hence, much of the value of the Choice System is temporary rebalancing.

Efficiency Gains From Savings

Some food banks never spend their shares. Empirically, the food rich—who already have a lot of food—don't spend as high a proportion of their shares as do the

[7] This calculation weights the goods in each category by price.

food poor. (This is perhaps because they have to pay transportation costs.) As such, the Choice System implies that the food rich reside inside their budget line. But if the food rich reside inside their budget line, the food poor can reach a higher budget line. This is because if the food rich do not spend their shares, average prices fall.

Thus, in Figure 2 we added a couple of points that show actual choices of purchased composite low-quality and high-quality goods. Here rather than predicted consumption of B, which was based on an assumption of full expenditure of shares, the food rich end up at point B', interior to the budget line. In turn, this allows the food poor to choose C', beyond the budget line. These differences are large: compared to the old system (point A), the food poor receive more of both kinds of food (though proportionately more-aimed at low-quality food) and the food rich receive less of both. (As the food rich and food poor have been selected on the kind of food that they buy, not the amount, this result is not hard wired into the analysis.) The food poor receive 66 percent more inexpensive food and 10 percent more expensive food.

In this way, the Choice System has had the effect of redistributing resources to the neediest areas of the country, which some of the food bank directors both noticed and appreciated. John Arnold, the member of the redesign committee and Director of the Western Michigan Food Bank who was highly skeptical of the Choice System, eventually became one of the most ardent users and supporters of the system, and focused his purchases on the less-expensive items.

Efficiency Gains at a More Granular Level

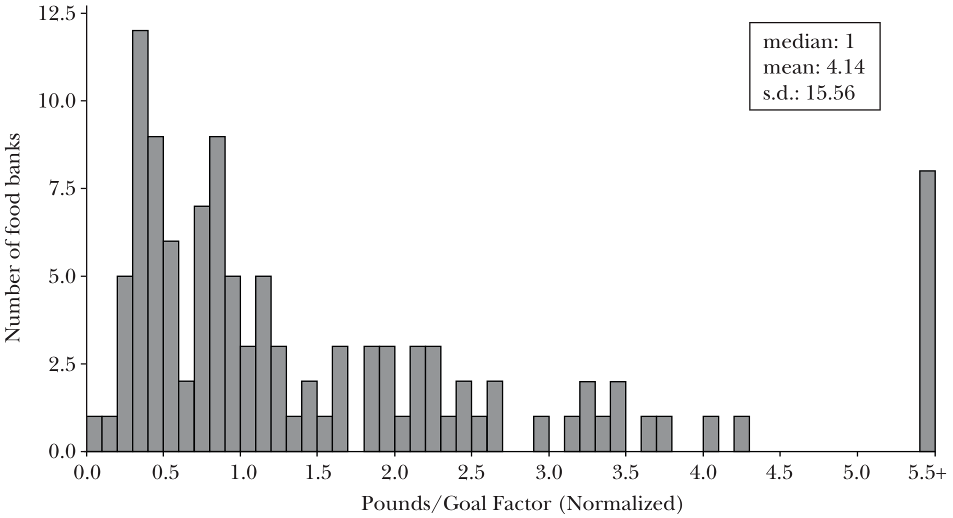
It is solely for expositional purposes that we are treating food banks and goods as binary (food banks being net buyers or sellers of high-quality goods; goods being high- or low-quality). In reality, there is much more dispersion than this. In Figure 3, we present the distribution of food banks according to total pounds of food received per Goal Factor (and remember shares received are proportional to Goal Factor) over the period 2005 to 2011. This distribution reflects both sorting on quality (those who buy more expensive goods would have fewer pounds of food) and permanent saving (as those who do not consume get fewer pounds). The figure is normalized so that Pounds/Goal Factor has a median of 1. Under the old assignment mechanism, this distribution would be bunched around 1. With the Choice System, food bank purchases in Figure 3—over a five-year period—diverge radically from this. For example, 25 percent of food banks get less than half as many pounds as before, while another 25 percent of food banks receive twice as many. Not surprisingly, dispersion over shorter intervals is even greater.

Hence, sorting occurs throughout the distribution of food banks. Taking into account the sorting across this distribution, not just high and low categories, reveals larger efficiency gains.

Maroon Pounds

Along with these efficiency gains, a second potential benefit of the new system is increased supply of food. Supply increased enormously after the introduction of

Figure 3
The Distribution of the Average Number of Pounds of Food per Share, 2005–2011



Note: We present the distribution of food banks according to total pounds of food received per Goal Factor (and remember shares received are proportional to Goal Factor) over the period 2005 to 2011. The figure is normalized so that Pounds/Goal Factor has a median of 1.

the Choice System. In the first year after its introduction, the supply of food rose by over 50 million pounds. As mentioned above, an important source of this gain was Maroon pounds. Maroon pounds add approximately 12 million pounds to supply each year from 2005 to 2012, with a range between 10 million and 18 million. It is also the case that these goods are on average higher quality than the average good in the system, and sell for 50 percent more.

Donor Issues and "Hard to Move" Products

Under the old queuing system, it was difficult to place "hard to move" product, as Feeding America called them. Arms were twisted so that someone would take little-desired items in order to keep donors happy. An innovation of the Choice System is to allow negative prices. In its first two years, 11 percent of loads involved the need for "bonus" shares, yet this has declined considerably to only 5 percent in 2010 and 2011. The 5 percent level is relatively small, and suggests that the need to keep donors happy involves relatively little distortion.[8]

[8] The Choice System does not allow negative prices for produce. This decision was made because produce is so abundant in the system that there was a concern that on some days the average price paid could be negative, which would result in the reallocation of shares at midnight reducing balance from one day to the next, which was seen as politically infeasible.

Before concluding, it is worth noting that these reallocative effects ignore two other welfare gains. First, we have not addressed rebalancing that occurs within price categories. For example, consider a food bank that already has a lot of yogurt. It can rebalance by buying other goods that sell for the same price as yogurt, such as milk or snacks. Gains from such rebalancing are not reflected here. Second, we have said nothing about the geography of the problem. Under the old system, food banks were often offered food far from their location, and would incur significant transportation costs to get it. The market system allows food banks to gain by focusing their purchases on loads of food that are geographically close, and so cut down on transportation costs.

To summarize, Feeding America knew that its previous system of offering the same amount, and kind of food, to food bank clients might not be optimal, but it did not have the hard information to design a better system. Indeed, given the information that Feeding America had available with the queuing system, it is likely that offering everyone the same thing was close to the best option.[9] The Choice System has allowed the participants to match outcomes to their preferences more effectively. Auctions have revealed willingness to pay for different kinds of food (who would have guessed that one pound of cereal was worth almost 50 pounds of produce?), which has allowed food banks to sort more efficiently on the quality–quantity dimension. In this way, the market system has allowed gains not possible with centralized assignment.

Leveling the Playing Field

We alluded earlier to concerns that a market-based system may not offer a level playing field to some food banks, particularly the smaller ones, and pointed out that the Choice System added a series of features to protect the interest of these food banks. Here we evaluate these features.

Credit

Smaller food banks have access to credit. This is extensively used. In the early stages of the Choice System, the use of credit was relatively rare, with only 4 percent of winning bids involving the use of credit shares in the first 18 months. However, over time, food banks have learned to make use of credit, so that from 2008 to 2011, the fraction of winning bids using credit has remained stable at roughly 11 percent. Remember that only about half of all food banks qualify for credit, so that among those food banks that qualify, almost one-quarter of all the winners use credit.

[9] Prendergast (2017) shows that Feeding America could have designed a somewhat better centralized assignment system than the one they used: for example, by offering higher Goal Factor food banks less food but giving them better food. However, given the information available, even a better-designed centralized system such as this does not get close to the outcomes that arise with the Choice System.

Joint Bidding

From 2005 to 2011, joint bids averaged between 1.2 and 2 percent of winning bids. Each joint bid on average has three bidders, so an alternative way to state the number above is that in 4 to 6 percent of cases, the winner is a joint bidder. A feature of joint bidding is that not so many food banks use it, but of those who do, some use it extensively. For example, the five food banks that use joint bidding most extensively use it for half of their winning bids.

Delegation and the Fairness Committee

Feeding America offered food banks the option to delegate bidding to Feeding America, and to appeal to the Fairness and Equity Committee if they felt they had been harmed through the Choice System. No food bank has ever chosen to delegate bidding control to Feeding America except for cases where the director is on vacation for a short period. Even more striking is that food banks have never submitted a request for a special hearing by the Fairness and Equity Committee, and so that committee has never convened.

The combination of credit use, joint bidding, and the absence of any need for Feeding America to intervene either to fix problems or to bid for the food banks strongly suggests that any concern that "small guys" would be disadvantaged has been alleviated.

Conclusion

As seen from afar by an academic economist, the idea that a specialized currency could be used to allocate food more efficiently while simultaneously respecting the relative needs of different areas may seem straightforward. However, despite the conceptual simplicity of the solution, it is worth pointing out that we rarely observe this kind of "Monopoly money" solution being used to allocate resources in real world settings. Why did it work for Feeding America?

Several unusual features of this setting allowed the use of the Choice System, but two stand out. First, dynamic markets with money only work if a participant who cannot find what is wanted today is willing to wait until tomorrow to spend the budget. In the Feeding America setting, the ongoing flow of goods is large—over a million pounds of food every day. As a result, participants who do not find what they want today likely will not have to wait long for a preferred good to come along. Second, the players here are long-lived: food banks are participants in an extended game with no known end point, which it is plausible to approximate as an infinitely repeated game. Again, this setting facilitates food banks foregoing consumption today if desired products are not currently available.

It is probably best to view the experience of Feeding America with the Choice Program not as a victory for markets per se, but rather as an illustration of how a flexible choice-revealing allocation system can be combined with a myriad of small details that include a focus on equity concerns. These tweaks—simple bidding mechanisms, access to credit, negative prices, the opportunity to delegate bidding, a fairness committee, the ability to bid jointly, the daily reallocation of shares, the use of a fully functioning demonstration game, and so on—seem to have made the difference for the acceptability and thus the longevity of this system. As such, it may be of some value in other not-for-profit settings aimed at improving allocative efficiency though consumer choice.

■ *I am grateful to Carole Theus, Harry Davis, Mike Halligan, Melanie Nowacki, and Bill Thomas for much help on this. Many thanks to Paul Kim and Wendy Wong for excellent research assistance. Most of all, however, none of this could have been written without the other members of the America's Second Harvest Allocation Task Force.*

References

Abdulkadiroglu, Atila, Parag A. Pathak, and Alvin E. Roth. 2005. "The New York City High School Match." *American Economic Review* 95(2): 364–67.

Abdulkadiroglu, Atila, Parag A. Pathak, and Alvin E. Roth. 2009. "Strategy-Proofness versus Efficiency in Matching with Indifferences: Redesigning the NYC High School Match." *American Economic Review* 99(5): 1954–78.

Budish, Eric. 2011. "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes." *Journal of Political Economy* 119(6): 1061–103.

Budish, Eric, and Estelle Cantillon. 2012. "The Multi-Unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard." *American Economic Review* 102(5): 2237–71.

Gatch, Loren. 2008. "Local Money in the United States during the Great Depression." In *Essays in Economic and Business History*, vol. 26, edited by Lynne Pierson Doti, 47–61. Economic and Business Historical Society.

Prendergast, Canice. 2017. "The Allocation of Food to Food Banks." http://faculty.chicagobooth.edu/canice.prendergast/research/foodwithmodel.pdf.

Radford, R. A. 1945. "The Economic Organisation of a P.O.W. Camp." *Economica* 12(48): 189–201.

Roth, Alvin E. 1984. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory." *Journal of Political Economy* 92(6): 991–1016.

Roth, Alvin E. 2008. "What Have We Learned from Market Design?" *Economic Journal* 118(527): 285–310.

Roth, Alvin E., and Elliott Peranson. 1999. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review* 89(4): 748–80.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver. 2004. "Kidney Exchange." *Quarterly Journal of Economics* 119(2): 457–88.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 163–184

Brexit: The Economics of International Disintegration

Thomas Sampson

On June 23, 2016, the United Kingdom held a referendum on its membership in the European Union. Although most of Britain's establishment backed remaining in the EU, 52 percent of voters disagreed and handed a surprise victory to the "leave" campaign. Brexit, as the act of Britain exiting the EU has become known, is likely to occur in early 2019.

The period since World War II has been marked by growing economic and cultural globalization and, in Europe, increasing political integration under the auspices of the European Union. Brexit marks a departure from this trend. For the United Kingdom, leaving the EU will mean withdrawing from the EU's supranational political institutions and will lead to the erection of new barriers to the exchange of goods, services, and people with the remaining 27 member states. More broadly, Brexit raises questions about the future stability of the EU and the extent to which further globalization is inevitable.

This article discusses the economic consequences of Brexit and the lessons of Brexit for the future of European and global integration. I start by describing the options for post-Brexit relations between the United Kingdom and the European Union and then review studies of the likely economic effects of Brexit. The main conclusion of this literature is that Brexit will make the United Kingdom poorer than it would otherwise have been because it will lead to new barriers to trade and migration between the UK and the European Union. There is considerable uncertainty

■ *Thomas Sampson is Assistant Professor, Centre for Economic Performance, Department of Economics, London School of Economics, London, United Kingdom. His email address is t.a.sampson@lse.ac.uk.*

[†] For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.163 doi=10.1257/jep.31.4.163

over how large the costs of Brexit will be, with plausible estimates ranging between 1 and 10 percent of UK per capita income. The costs will be lower if Britain stays in the European Single Market following Brexit. Empirical estimates that incorporate the effects of trade barriers on foreign direct investment and productivity find costs 2–3 times larger than estimates obtained from quantitative trade models that hold technologies fixed. Other EU countries are also likely to suffer economically from Brexit, but their estimated losses are much smaller than those faced by the United Kingdom.

Assessing the broader implications of Brexit for the European Union and for globalization requires understanding why the United Kingdom voted to leave. Thus, I next discuss why the referendum was held and who voted for Brexit. Support for Brexit came from a coalition of less-educated, older, less economically successful and more socially conservative voters who oppose immigration and feel left behind by modern life. Leaving the EU is not in the economic interest of most of these left-behind voters. However, there is currently insufficient evidence to determine whether the leave vote was primarily driven by national identity and the desire to "take back control" from the EU, or by voters scapegoating the EU for their economic and social struggles. The former implies a fundamental opposition to deep economic and political integration, even if such opposition brings economic costs, while the later suggests Brexit and other protectionist movements could be addressed by tackling the underlying reasons for voters' discontent.

Options for United Kingdom–European Union Relations after Brexit

On March 29, 2017, the United Kingdom formally notified the European Union of its intention to leave, triggering the start of negotiations on a withdrawal agreement. Article 50 of the Lisbon Treaty allows withdrawal negotiations to last for at most two years. After this period, Britain[1] will automatically cease to be a member of the EU even if there is no agreement, unless member states unanimously decide to extend the negotiations. The withdrawal agreement will cover the UK's outstanding financial liabilities to the EU, the future status of EU citizens living in the UK and British people living in Europe, and the framework for future UK–EU relations, but will not finalize the details of any new relationship (European Council 2017).

While the British government has committed to implementing the referendum outcome, the "leave" vote provided no guidance as to what form Brexit should take. In broad terms, there are three options. First, the United Kingdom could remain part of the European Union's Single Market by joining Norway, Iceland, and Liechtenstein in what is called the European Economic Area (EEA). Second, the UK and EU could sign a free trade agreement to govern their trade and economic relations.

[1] With apologies to the people of Northern Ireland, I will use "Britain" and "United Kingdom" interchangeably throughout this article to refer to the United Kingdom of Great Britain and Northern Ireland.

Third, if no alternative agreement is reached, the UK and EU would trade under the most-favored nation terms available to all World Trade Organization members (for further details on these alternatives, see HM Government 2016; Dhingra and Sampson 2016). Each of these options embodies a different resolution to the trade-off Britain faces between maintaining economic integration with the EU and reasserting national control over powers that are shared between EU members.

Joining the European Economic Area, like Norway, is the option closest to remaining a member of the European Union. EEA members are part of the European Single Market, which means they commit to its four freedoms: free movement of goods, services, capital, and labor. EEA members must adopt all EU legislation regarding the Single Market, which covers areas such as employment law, consumer protection, product standards, and competition policy. EEA members also pay to be part of the Single Market through contributing to the EU budget. In 2011, Norway's contribution of £106 per capita was 83 percent as large as the UK's net per capita payment to the EU (House of Commons 2013).

The Single Market lowers trade costs by reducing both border barriers to trade, which are imposed when goods and services cross borders, and behind-the-border barriers, which arise from international differences in regulation and economic policy. For example, Single Market "passporting rights" give financial firms based in one member state the right to provide services throughout the Single Market, thereby reducing border barriers to trade in financial services. In addition, regulatory harmonization lowers behind-the border barriers by ensuring producers do not have to adapt their goods to satisfy different product standards in different countries.

However, trade barriers between European Economic Area countries and the European Union are higher than within the EU because Norway, Iceland, and Liechtenstein do not belong to the EU's Customs Union, which means they can set their own external tariffs and conduct their own trade negotiations with countries outside the EU. It also means trade between EEA members and the EU is subject to border barriers such as customs procedures, enforcement of rules of origin, and anti-dumping duties.

The impact for Britain of leaving the Single Market on trade barriers would depend upon what, if any, new deal the United Kingdom and the European Union negotiated. Absent a new deal, Britain would trade with the EU under World Trade Organization terms, as the United States and China currently do. Goods trade would be subject to most-favored nation tariffs and both border and behind-the-border nontariff barriers would increase. Multilateral trade liberalization under the World Trade Organization has made substantial progress in reducing import tariffs on nonagricultural goods. The EU's average most-favored nation tariff on goods imports was just 4.4 percent in 2015 (World Bank 2017). However, multilateral negotiations have been less successful in lowering nontariff barriers, particularly for services. Borchert (2016) documents how openness to services trade is higher within the EU than between EU and non-EU countries. For example, non-EU firms do not have passporting rights in financial services and only airlines that are majority owned by EU nationals can operate flights within the EU. Overall, the World Trade Organization option would result in the largest increase in trade barriers between the United Kingdom and the EU.

Free trade agreements differ greatly in their depth, scope, and effects on trade (Hofmann, Osnago, and Ruta 2017), offering a menu of options for the United Kingdom and the European Union to negotiate over. Most recent free trade agreements have focused on lowering nontariff barriers and increasing market access in services. However, the EU's existing trade deals, such as the EU–Canada agreement, do much less than the Single Market to harmonize economic regulation and do not guarantee market access for service providers. Consequently, any free trade agreement would lead to higher trade costs with the EU than if Britain remains in the Single Market.

Instead of negotiating a tailor-made free trade agreement, the United Kingdom could also seek to form a customs union with the European Union, as Turkey has done. This would ensure UK–EU goods trade did not face tariffs or other border barriers, such as rules of origin, but customs union membership alone would do nothing to lower behind-the-border barriers or reduce restrictions on services trade. It would also prevent the UK from negotiating its own trade agreements with non-EU countries.

Outside the Single Market, the United Kingdom would not be bound by European Union economic regulation nor subject to the jurisdiction of European courts and would be free to restrict immigration from the EU. However, any free trade agreement with the EU would require relinquishing domestic control over some economic policies. Consider the case of Switzerland. Of all countries outside the European Economic Area, Switzerland is the most economically integrated with the EU and effectively belongs to the Single Market in goods. But to achieve this level of integration, Switzerland has been obliged to adopt many pieces of EU economic legislation, to contribute to the EU budget, and to accept free movement of labor—even though the Swiss electorate voted in 2014 to restrict immigration from the EU. Despite these concessions, Switzerland and the EU have not reached a comprehensive agreement on trade in services, meaning, for example, that Swiss banks do not have passporting rights.

A new trade deal between Britain and the European Union is unlikely to be concluded before March 2019. For example, the EU–Canada trade agreement started to come into force in 2017, eight years after negotiations began. Consequently, an interim agreement will probably be needed to avoid disruption to UK–EU trade in the period between Britain leaving the EU and any new trade agreement being reached.

At the time of writing, the likely shape of future relations between the United Kingdom and the European Union remains unclear. EU leaders have signalled that, although they hope to maintain close economic relations with the UK, they are not willing to compromise on the indivisibility of the four freedoms of the Single Market (*Financial Times* 2017). Contrary to the "continental partnership" proposed by Pisani-Ferry, Röttgen, Sapir, Tucker, and Wolff (2016), this means that in order to remain part of the Single Market in goods and services, the UK would have to continue allowing free movement of labor with the EU.

Facing this choice, Prime Minister Theresa May announced in January 2017 that the United Kingdom would leave the Single Market and seek a new free trade agreement with the European Union that would "allow for the freest possible trade in goods and services between Britain and the EU's member states." She also announced Britain would leave the EU's Customs Union to enable it to negotiate trade deals with non-EU countries. On June 8, 2017, Prime Minister May held a general election to seek a mandate for this position. Her Conservative Party won the most seats, but unexpectedly lost its majority in Parliament, denying May her mandate. The election result has prompted fierce debate over whether the UK should prioritize remaining economically integrated with the EU or taking control of immigration and economic regulation. However, as yet, it has not led the government to change its position.

The Economic Consequences of Brexit

The United Kingdom is a small open economy with a comparative advantage in services that relies heavily on trade with the European Union. In 2015, the UK's trade openness, measured by the sum of its exports and imports relative to GDP, was 0.57, compared to 0.28 for the United States and 0.86 for Germany (World Bank 2017). The EU accounted for 44 percent of UK exports and 53 percent of its imports. Total UK–EU trade was 3.2 times larger than the UK's trade with the United States, its second-largest trade partner. UK–EU trade is substantially more important to the United Kingdom than to the EU. Exports to the EU account for 12 percent of UK GDP, whereas imports from the EU account for only 3 percent of EU GDP. Services make up 40 percent of the UK's exports to the EU, with "Financial services" and "Other business services," which includes management consulting and legal services, together comprising half the total.[2]

Brexit will lead to a reduction in economic integration between the United Kingdom and its main trading partner. How will this change affect the British and European economies? And how will the consequences of Brexit depend upon which option is chosen for future UK–EU relations?

Forecasting the economic consequences of Brexit is made difficult by the lack of a close historical precedent. Algeria left the European Communities (EC), as the European Union was previously known, upon becoming independent from France in 1962, as did Greenland in 1985 after achieving autonomy within Denmark, but neither of these cases is likely to shed much light on the impact of Brexit. Facing this challenge, researchers have used three approaches to estimate the effects of

[2] Trade data is for 2015 and is from the Office for National Statistics Pink Book (Office for National Statistics 2016a). United Kingdom GDP data is from the Office for National Statistics Blue Book (Office for National Statistics 2016b), and European Union GDP data is from the World Bank (2017).

Brexit: 1) historical case studies of the economic consequences of joining the EU; 2) simulations of Brexit using computational general equilibrium trade models, and 3) reduced-form evidence based on estimates of how EU membership affects trade. Each of these methodologies is subject to a number of limitations, but collectively they offer the best available evidence on how Brexit is likely to affect economic outcomes in the United Kingdom and the European Union.

The results I summarize in this section focus on long-run effects and have a forecast horizon of 10 or more years after Brexit occurs. Less is known about the likely dynamics of the transition process or the extent to which economic uncertainty and anticipation effects will impact the economies of the United Kingdom or the European Union in advance of Brexit. Following the June 2016 referendum, sterling depreciated sharply and by the end of June 2017 was 12 percent lower against the dollar than immediately before the vote. As shown in Figure 1, this has contributed to a rise in inflation from 0.5 percent in June 2016 to 2.6 percent a year later and a decline in real wage growth from 1.5 percent to -0.5 percent over the same period. Output growth in the UK has also slowed, with GDP increasing at an annualized rate of 1.0 percent in the first half of 2017, compared to 1.7 percent in the year leading up to the referendum (Office for National Statistics 2017). These statistics suggest the referendum outcome is already harming the UK economy, though, of course, Britain is yet to leave the EU.

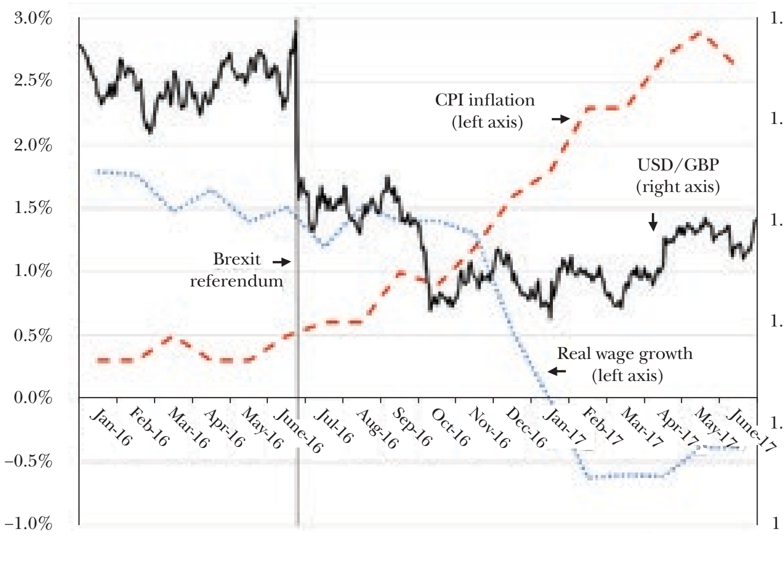
Case Studies of Joining the European Union

Crafts (2016) reviews the historical evidence on how joining the European Communities in 1973 affected the UK economy. He concludes that membership raised GDP per capita in the United Kingdom, particularly through productivity growth resulting from increased product market competition. Falling barriers to trade reduced domestic firms' market power, and firms responded by investing more in productivity improvements. A quantitative analysis of the historical data is undertaken by Campos, Coricelli, and Moretti (2014), who use the synthetic control methodology. Their estimates imply that ten years after joining the EC, UK GDP per capita was 8.6 percent higher than it otherwise would have been. Fully disentangling the treatment effect of accession from other contemporaneous shocks is probably an impossible challenge, and it would be naïve to expect that Brexit will simply have the opposite effect to joining the EC in 1973. But subject to these caveats, historical analysis concludes that the UK obtains substantial economic benefits from being an EU member.

Simulations with General Equilibrium Trade Models

The most widely adopted approach for studying Brexit has been to run simulations using computational general equilibrium trade models. These models use assumptions regarding how Brexit will affect trade costs between the United Kingdom and its trading partners to generate predicted changes in trade, consumption, production, and welfare. Important advantages of this approach are that it accounts for general equilibrium effects, such as trade diversion between the UK

Figure 1
UK Exchange Rate, Inflation, and Wage Growth



Source: Exchange rate from Bloomberg; CPI and real wage growth from Office for National Statistics.
Notes: USD/GBP is end-of-day rate. Inflation is annual change in CPI (series D7G7). Wage growth is annual change in seasonally adjusted Regular Pay (series A2F9).

and non-EU countries, and that it enables researchers to tailor their assumptions regarding how Brexit will affect trade costs to study alternative post-Brexit scenarios.

Modelling changes in nontariff barriers, such as customs procedures, market access restrictions, and regulation, is, of course, an imperfect art. To implement simulations, the assumed impact of Brexit on nontariff barriers must be expressed numerically, typically in terms of ad-valorem equivalent trade costs. However, there is no generally accepted methodology for quantifying counterfactual nontariff barriers, meaning it is important to examine the robustness of simulation results to plausible alternative specifications of changes in trade costs. In addition, no single model will capture all the channels through which trade affects the global economy, making it useful to compare results across studies.[3]

An example of the simulation approach is that of Dhingra et al. (2017), who estimate the effects of Brexit using a quantitative trade model with 31 industries, 35 countries, and trade in intermediate inputs which is based on the multisector version of Eaton and Kortum (2002) developed by Caliendo and Parro (2014).

[3] Kehoe, Pujolas, and Rossbach (2016) review some of the past failings of computational trade models and recommend that future quantitative models account better for heterogeneity within countries and industries.

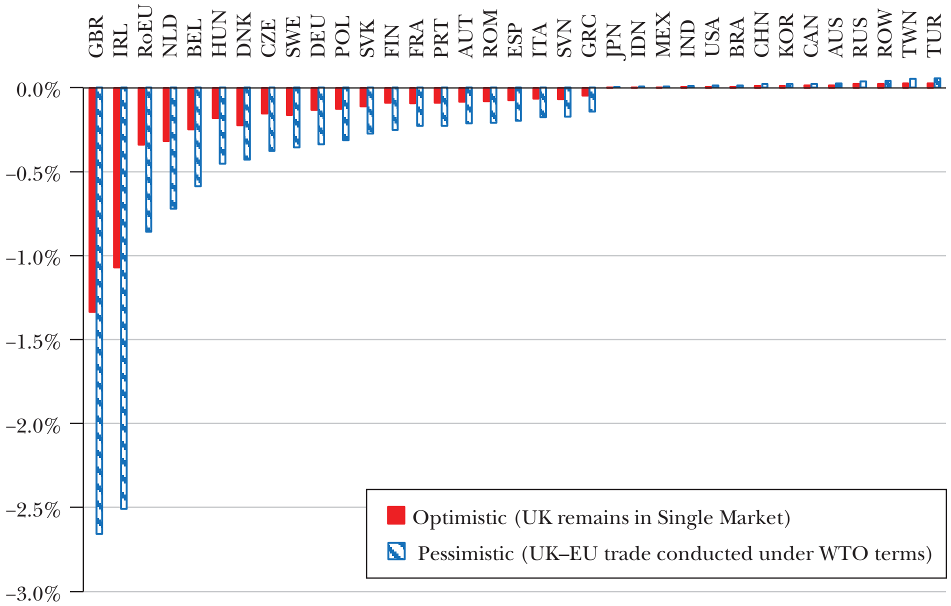
They consider three channels through which Brexit may affect trade costs: tariffs, nontariff barriers, and future declines in intra-EU trade costs in which the United Kingdom participates only if it remains an EU member. Future trade costs changes are included because Méjean and Schwellnus (2009) estimate that intra-EU trade costs have been falling approximately 40 percent faster than trade costs between other OECD countries. Dhingra et al. (2017) model an optimistic scenario in which the UK remains in the Single Market and a pessimistic scenario in which UK–EU trade is conducted under World Trade Organization terms. They also allow for a decline in the UK's net fiscal contribution to the EU budget following Brexit.[4]

In both scenarios, Dhingra et al. (2017) find that the efficiency losses the United Kingdom suffers from higher trade barriers exceed the fiscal savings. Increased trade costs are welfare-reducing because the United Kingdom faces higher import prices and is less able to specialize according to comparative advantage, which reduces production efficiency and output. Higher trade costs can also affect welfare through channels not analyzed by Dhingra et al., such as by reducing product variety and raising mark-ups (Krugman 1979), or by allowing less-efficient firms to survive which decreases aggregate productivity (Melitz 2003). However, all these mechanisms imply that higher trade barriers lead to lower welfare. In the optimistic case (the UK remains in the Single Market), Dhingra et al. (2017) estimate Brexit is equivalent to a permanent 1.3 percent decline in UK consumption per capita, while in the pessimistic case (UK–EU trade is conducted under World Trade Organization terms), the loss doubles to 2.7 percent. Quantitatively, these estimates are dominated by the consequences of higher nontariff barriers and exclusion from future declines in intra-EU trade costs, reflecting the fact that the EU's most-favored nation tariffs are low relative to estimates of nontariff barriers.

Figure 2 shows that European Union countries also suffer from the fall in UK–EU trade. However, with the notable exception of Ireland, the losses are an order of magnitude smaller, because UK–EU trade is relatively less important to the EU than the UK. Brexit is equivalent to a 0.14 percent fall in EU consumption per capita in the optimistic case and a 0.35 percent fall in the pessimistic case. Non-EU countries benefit from Brexit due to trade diversion, but the effects are quantitatively negligible compared to the losses faced by the UK and the EU. Other studies have found qualitatively and quantitatively similar results (for examples, see, Aichele and Felbermayr 2015, who using a modelling framework based on Eaton and Kortum 2002, and Ciuriak et al. 2015, who use a version of the Global Trade Analysis Project model).

[4] In the optimistic case, there are no tariffs between the United Kingdom and European Union, nontariff barriers increase by one-quarter of the estimated reducible nontariff barriers on US–EU trade, intra-EU trade costs fall 20 percent faster than in the rest of the world for ten years after Brexit, and the UK's per capita contribution to the European Union budget is equal to Norway's contribution. In the pessimistic case, the EU's most-favored-nation tariffs are imposed on UK–EU trade, nontariff barriers increase by three-quarters of the reducible nontariff barriers on US–EU trade, intra-EU trade costs continue to fall by 40 percent faster than in the rest of the world for ten years after Brexit, and the UK makes no budget payments to the EU.

Figure 2
Estimated Welfare Effects of Brexit



Source: Dhingra et al. (2017).
Notes: Estimates give the permanent percentage change in income per capita that has the same welfare effect as Brexit. In the optimistic scenario, the UK remains in the Single Market following Brexit. In the pessimistic scenario, UK–EU trade is conducted under WTO terms. See Dhingra et al. (2017) for details. The labels on the *x*-axis are World Bank country codes. RoEU = Rest of EU; ROW = Rest of World.

One limitation of the existing literature lies in how it models financial services. London is Europe's leading financial center, and financial and insurance services accounted for 7.5 percent of UK value-added and 13 percent of exports in 2014 (Office for National Statistics 2016a, b). Oliver Wyman (2016) estimates that around one-quarter of finance industry revenue comes from business related to the European Union. If Britain leaves the Single Market, UK-based finance companies will lose their passporting rights and face higher barriers to accessing European markets. However, the trade models used to study Brexit do not account for the agglomeration forces that shape location decisions in the finance industry. This may lead them to overestimate the Brexit effect if agglomeration externalities insulate the UK's finance industry against higher trade costs, or underestimate the effect if Brexit threatens London's position as Europe's financial hub.

An alternative way to estimate the impact of Brexit on the finance industry is through case studies that analyze how much business the United Kingdom may lose in different subsectors. Using this approach, Djankov (2017) estimates that, if the United Kingdom and European Union trade under World Trade Organization rules, finance industry revenue would fall by between 12 and 18 percent and employment would fall by between 7 and 8 percent. By comparison, Dhingra et al. (2017), in their pessimistic scenario, estimate finance industry output would fall by 6.4 percent. This suggests both approaches lead to similar results, but further research on the finance industry would certainly be valuable.

Reduced-Form Evidence

The reduced-form approach to studying Brexit involves two steps: 1) use the "gravity equation" for bilateral trade, in which trade levels depend upon economic size, geographic distance, and other factors that affect trade costs, to estimate the effect of EU membership on trade, and 2) combine the outcome of step one with an estimate of the elasticity of income per capita to trade to obtain the effect of EU membership on income per capita. The attraction of this approach is that it does not rely on assuming the validity of a specific trade model and allows researchers to exploit richer empirical variation than simply studying changes in output following EU accession. Its main limitation is the difficulty of obtaining causal estimates of the parameters of interest.

Dhingra et al. (2017) implement the reduced-form approach using gravity estimates from Baier, Bergstand, Egger, and McLaughlin (2008) that are identified from variation in trade when countries join the European Union. Baier et al.'s estimates imply that leaving the EU and joining the European Free Trade Association would reduce the UK's trade with EU members by 25 percent.[5] Assuming no trade diversion with non-EU countries and using Feyrer's (2009) estimate that the elasticity of income per capita to trade lies between 0.5 and 0.75, it follows Brexit would reduce UK income per capita by between 6.3 and 9.4 percent.

It is notable the reduced-form approach leads to losses that are several times larger than the estimates from model-based simulations, even though both methods give similar predictions regarding changes in trade.[6] This difference may arise because the reduced form estimates capture channels that are absent from quantitative trade models. The computational models used to study Brexit treat technology as exogenous, implying they will underestimate the costs of Brexit if trade integration raises productivity growth or leads to technology upgrading (as found by Bustos 2011). In Sampson (2016), I show that allowing for trade to affect productivity through knowledge spillovers across firms approximately triples the gains from trade in a version

[5] The Baier et al. (2008) estimates are based on goods trade data for 1960–2000 and assume the trade effect of EU membership is homogeneous across countries. Mulabdic, Osnago, and Ruta (2017) perform a similar exercise using a continuous measure of the coverage of different trade agreements with 1995–2011 data for both goods and services trade and allowing for UK-specific treatment effects. Their estimates suggest Brexit will reduce services trade between the United Kingdom and European by slightly more than goods trade and imply larger reductions in total UK–EU trade following Brexit than Baier et al.'s results (see their table 6).

[6] For example, Dhingra et al.'s (2017) quantitative model implies total British trade declines by 9 percent in the optimistic case and 16 percent in the pessimistic case, while their reduced-form estimates are based on a 12.5 percent decline in UK trade.

of the Melitz (2003) model. In addition, since trade and other forms of economic integration are highly correlated, Feyrer's (2009) estimate of the elasticity of income per capita to trade likely captures not only trade, but also other consequences of closer integration. This observation implies that the reduced form estimates probably incorporate some of the broader effects of Brexit resulting from changes in foreign direct investment, immigration, and international technology diffusion.

Foreign Direct Investment and Immigration

Although most studies of the economics of Brexit focus on trade, another likelihood is that the British economy will suffer from reductions in foreign direct investment and immigration after leaving the European Union. The Single Market has allowed foreign investors to use the United Kingdom as an export platform for serving EU markets. Looking at the automobile industry, Head and Mayer (2015) use a quantitative model of trade and foreign direct investment to estimate that increases in trade costs and intrafirm coordination costs following Brexit will reduce car production in the UK by 12 percent. At the aggregate level, Bruno, Campos, Estrin, and Tian's (2016) estimates using a gravity equation imply that leaving the Single Market will reduce the flow of foreign direct investment into the UK by around 22 percent. Since foreign direct investment has positive effects on domestic investment and productivity, this decline is likely to reduce UK output and living standards (Dhingra, Ottaviano, Sampson, and Van Reenan 2016).

Leaving the Single Market would also allow the United Kingdom to adopt policies to restrict immigration from the European Union. The effects of changes in immigration policy are difficult to forecast, but an application of the reduced-form methodology to immigration by Portes and Forte (2017) concludes lower immigration from the EU could reduce the UK's GDP per capita by between 0.9 and 3.4 percent by 2030. I am unaware of any aggregate-level analysis of how changes in trade, foreign direct investment, and immigration may interact following Brexit. But these interactions could be important, particularly for sectors such as finance that rely on access to highly skilled workers from across the EU.

Economic Arguments in Favor of Brexit

Economic arguments for Brexit have focused on the ideas that leaving the EU's Customs Union would allow the United Kingdom to strike new trade agreements with non-EU countries and that leaving the Single Market would allow the United Kingdom to deregulate its economy (Booth, Howarth, Persson, Ruparel, and Swidlicki 2015). It is unclear whether Brexit will result in the United Kingdom facing lower or higher barriers to trade with non-EU countries in the long run. The advantage of not needing to compromise with 27 other countries to reach new agreements must be weighed against the costs of being a smaller market than the European Union with less bargaining power in negotiations and the risk of losing access to existing free trade agreements between the EU and other countries. Whichever effect dominates, it is highly unlikely new trade deals could fully compensate for lower UK–EU trade. Ebell (2016) estimates membership of the Single Market has approximately twice as large an effect on bilateral goods trade as an average free trade agreement and finds that, unlike the Single Market, the average services free trade agreement in her dataset has no statistically significant trade effects. Because around 60 percent of the UK's trade is with either the EU or countries that have already signed a free trade agreement with the EU, these results imply leaving the Single Market would reduce total UK trade even under optimistic assumptions about the UK's success in negotiating new trade agreements following Brexit.

Claims that the United Kingdom will reap substantial benefits from post-Brexit deregulation are even less convincing. Open Europe (2015) lists 57 regulations based on EU legislation for which economic impact assessments by the UK government find higher costs than benefits. The net annual cost of these regulations is 0.9 percent of UK GDP, but half this cost comes from just two regulations aimed at reducing carbon dioxide emissions and limiting working hours. Support for these policies within the United Kingdom exists independently of EU legislation. More generally, there is no persuasive evidence that UK voters see Brexit as a reason for further deregulation. According to the OECD's Indicators of Product Market Regulation and its Employment Protection Database, the UK's product and labor markets are already among the least regulated in the OECD with similar levels of regulation to the US economy and much lower regulation than most EU countries. This suggests EU membership has not prevented the United Kingdom from tailoring regulation to suit its national preferences.

Drawing Conclusions

Overall, the research literature displays a broad consensus that in the long run Brexit will make the United Kingdom poorer because it will create new barriers to trade, foreign direct investment, and immigration. However, there is substantial uncertainty over how large the effect will be, with plausible estimates of the cost ranging between 1 and 10 percent of the UK's income per capita. European Union countries are also likely to suffer from reduced trade, but in percentage terms their losses are expected to be much smaller. The uncertainty over the size of the Brexit effect has two sources. First, alternative research strategies produce quantitatively different results. Second, the losses will depend upon the terms under which the United Kingdom and EU trade following Brexit. Continued membership of the Single Market is the best option for the British and European economies. But if Britain leaves the Single Market, the research shows that, to minimize the costs of Brexit, UK–EU negotiations should prioritize keeping nontariff barriers low and ensuring market access in services rather than focusing purely on tariffs.

In years to come, the experience of Brexit is likely to stimulate much interesting research. It offers a novel natural experiment that will allow researchers to study the economic effects of raising barriers to trade and to evaluate the results of the estimation methods described above. There should also be opportunities to study the dynamics of adjustment to trade deliberalization, the relative importance of different nontariff barriers, and whether trade, foreign direct investment, and immigration are complements or substitutes, among other questions.

Implications of Brexit for the Future of the European Union and Globalization

Sixty years after the Treaty of Rome first established the European Economic Community, the European Union is struggling with the aftermath of the global financial crisis, geopolitical instability on its eastern and southern borders, and the success of anti-European political parties in many member states. Brexit adds to these challenges. This final section of the paper discusses what Brexit means for the future of the EU and, more broadly, global economic integration. To address these questions, we first need to consider why Britain voted to leave the EU.

The Brexit Referendum

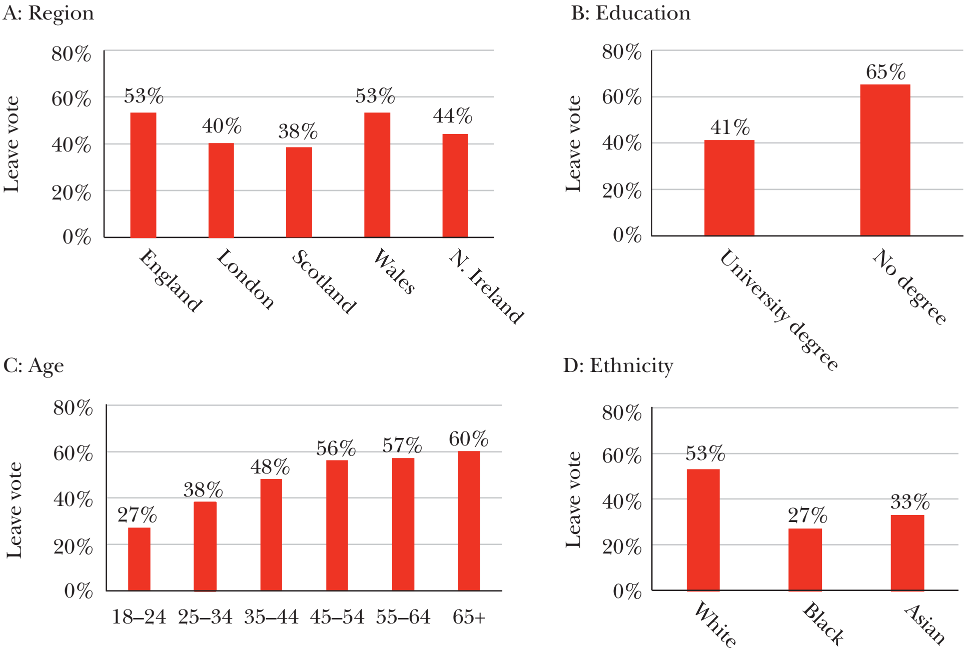
The 2016 referendum was the culmination of a 20-year campaign against Britain's membership of the European Union that started after the Maastricht Treaty transformed the European Communities into the EU and launched the European Single Market in 1993. In Britain, the energy behind this campaign came primarily not from the Conservative or Labour parties, but from single issue parties set up to advocate for Brexit—first the Referendum Party and then the United Kingdom Independence Party (UKIP). These parties argued that sharing political power with the European Union was an unwanted constraint on Britain's sovereignty. Particular bones of contention were the UK's commitments to allow free movement of labor within the EU and to accept the jurisdiction of the European Court of Justice.

The movement to leave the European Union became increasingly influential after Nigel Farage took over as leader of the United Kingdom Independence Party in 2006 and broadened the party's appeal among working class voters. In 2014, UKIP won a plurality of votes in Britain's elections to the European Parliament and captured 24 of the UK's 73 seats. Under pressure both from supporters of UKIP and from within his increasingly euro-skeptic Conservative Party, then-Prime Minister David Cameron pledged to hold a referendum on EU membership if the Conservatives won the 2015 general election. Although Cameron supported remaining in the EU, he hoped his pledge would shore up right-wing support for the Conservatives and gambled that the British public would not vote to leave the EU. After the Conservatives won a surprise majority, Cameron's gamble was put to the test. On June 23, 2016, 17.4 million voted to leave the EU and only 16.1 million to remain. Cameron resigned as Prime Minister the following day, and the Conservative Party chose Theresa May as his replacement.

Who Voted for Brexit?

The referendum split the British electorate on the basis of geography, age, education, and ethnicity. Figure 3 shows data on voting patterns. England and Wales voted to leave, while Scotland and Northern Ireland voted to remain. Within England, support for Brexit was noticeably lower in London, where only 40 percent voted to leave.

Figure 3
"Leave" Vote Shares in Brexit Referendum



Source: Regional data from the Electoral Commission. Demographic data from Lord Ashcroft Polls.
Notes: The geographic breakdown uses actual votes cast in the referendum. All other data on voting patterns is from polling conducted by Lord Ashcroft Polls (2016) on the day of the referendum. See http://lordashcroftpolls.com/2016/06/how-the-united-kingdom-voted-and-why/.

Older and less-educated voters were more likely to vote "leave." Of those aged 18–24, 27 percent voted to leave compared to 60 percent of voters aged over 65. Only 41 percent of voters with a university degree chose leave, whereas 65 percent of those without a degree voted to leave. A majority of white voters wanted to leave, but only 33 percent of Asian voters and 27 percent of black voters chose leave. There was no gender split in the vote, with 52 percent of both men and women voting to leave. Interestingly, although Brexit has never received much backing from liberal or left-wing political leaders, leaving the European Union received support from across the political spectrum. A strong majority of 58 percent of Conservative voters supported leave, but so did 37 percent of Labour voters and 36 percent of Scottish National Party supporters.

Voting to leave the European Union was strongly associated with holding socially conservative political beliefs, opposing cosmopolitanism, and thinking life in Britain is getting worse rather than better. Among people who said feminism is a "force for ill," 74 percent voted to leave, compared to 38 percent of those who thought feminism a "force for good." Similarly, 69 percent of people who thought globalization a force for ill voted to leave, as did 81 percent of people who viewed multiculturalism as a force for ill. Among voters who backed staying in the EU, 73 percent thought "life in Britain today is better than it was 30 years ago," while 58 percent of leave voters thought life was worse.

Econometric studies of voting outcomes by area (Goodwin and Heath 2016a; Becker, Fetzer, and Novy 2017; Colantone and Stanig 2016) and voting intentions at the individual level (Goodwin and Heath 2016b; Colantone and Stanig 2016) provide a richer picture of the demographic and economic variables associated with voting to leave.

First, education and, to a lesser extent, age were the strongest demographic predictors of voting behavior. For example, Becker, Fetzer, and Novy (2017, table 3) show that the share of the population with a university degree or equivalent qualification, on its own, accounts for 62 percent of the variation in the share of the vote received by "leave" across 380 areas and that both the level and growth of the proportion of the population aged 60 and over are associated with a higher leave vote share.

Second, poor economic outcomes at the individual or area level were associated with voting to leave, but economic variables accounted for less of the variation in the leave vote share than educational differences. Controlling for age, gender, and ethnicity, Goodwin and Heath (2016b) find support for leave was 10 percentage points higher among households with income below £20,000 than among households with income above £60,000, but was 30 percentage points higher from individuals whose highest educational qualification is at the General Certificate of Secondary Education level (a qualification usually obtained at age 16) than from those with a university degree. While most studies of the referendum vote have focused on documenting correlations, Colantone and Stanig (2016) use an estimation strategy based on Autor, Dorn, and Hanson (2013) to show that exposure to Chinese import competition led to increased support for Brexit. At the regional level, their estimates imply a one standard deviation increase in Chinese import competition raised the leave vote share by almost two percentage points. By contrast, they estimate a one standard deviation increase in the proportion of the population with a university degree is associated with a five percentage point fall in the leave vote. They also find higher unemployment is associated with greater support for leave, as do Becker, Fetzer, and Novy (2017).

Third, support for leaving the European Union is strongly associated with self-reported opposition to immigration, but not with exposure to immigration. Immigration played a central role in the Brexit campaign, and Goodwin and Heath (2016b) report 88 percent of people who thought the United Kingdom should admit fewer immigrants supported Brexit. However, studies find that a higher share of EU immigrants in the population is actually associated with a reduction in the leave vote share across local areas. There is some evidence that growth in immigration, particularly from the 12 predominantly eastern European countries that joined the EU in 2004 and 2007, is associated with a higher leave vote share, but the effect is small and not always present (Colantone and Stanig 2016; Goodwin and Heath 2016a; Becker, Fetzer, and Novy 2017).[7]

Overall, the picture painted by the voting data is that the Brexit campaign succeeded because it received the support of a coalition of voters who felt left-behind by modern Britain. People may have felt left-behind because of their education, age, economic situation, or because of tensions between their values and the direction of social change, but, broadly speaking, a feeling of social and economic exclusion appears to have translated into support for Brexit.

Why Did Britain Vote for Brexit?

Knowing that the left-behind voted for Brexit does not tell us *why* they voted for Brexit. Hobolt and de Vries (2016) detail three factors that affect support for European integration: economic cost–benefit calculations; values and identity; and the information available to voters. One possible explanation for the referendum outcome can be ruled out immediately. Britain's vote to leave the European Union was not the result of a rational assessment of the economic costs and benefits of Brexit. As highlighted in the previous section, there is a broad consensus in the literature that being part of the EU has benefited the UK economy on aggregate.

Moreover, there is no evidence that changes in either trade or immigration due to EU membership have had large enough distributional consequences to offset the aggregate benefits and leave left-behind voters worse off. There is little direct evidence on the distributional impact of UK–EU trade. Using a quantitative model in which trade affects wage inequality through both inter- and intra-industry changes in the demand for skill, Burstein and Vogel (forthcoming) estimate that moving to autarky would reduce wage inequality in the United Kingdom but would also make both skilled and unskilled workers worse-off. Extrapolating from this result suggests neither high- nor low-skill British workers stand to gain from a reduction in trade with the EU.

In practice, most discussion of the effect of EU membership on inequality in the United Kingdom centers not on trade, but on the wage effects of immigration. Immigration to the United Kingdom from EU countries increased rapidly from the late 1990s onwards; and between 1995 and 2015, the share of EU nationals in the UK's population rose from 1.5 to 5.3 percent (Wadsworth, Dhingra, Ottaviano, and Van Reenen 2016). Studies do not find significant negative effects of immigration on average employment or wages for UK natives, but there is some evidence immigration has reduced wages for lower-paid workers (Dustmann, Frattini, and Preston 2013; Nickell and Saleheen 2015). Wadsworth, Dhingra, Ottaviano, and Van Reenen (2016) report that, based on the level of immigration from the European Union between 2004 and 2015, Dustmann, Frattini, and Preston's estimates

[7] Relatedly, Becker and Fetzer (2016) find evidence of a small post-2004 increase in support for the United Kingdom Independence Party in European Parliamentary elections in areas where the increase in immigration from the ten 2004 accession countries was higher relative to the initial stock of European Union immigrants.

imply a 1.0 percent wage decline for native workers in the bottom decile of the wage distribution. Likewise, Nickell and Saleheen's estimates imply a 0.7 percent decline in wages in semi-skilled and unskilled service sectors. These losses are lower than the estimated gains from trade due to EU membership.

The observation that Brexit will impose economic costs even on many of its supporters establishes an important difference between Brexit and protectionist trade policies, such as anti-dumping duties or restrictions on agricultural imports, which receive support because they shield particular groups of voters from loses caused by economic integration. In this sense, support for Brexit is a distinct phenomenon from opposition to trade with China among manufacturing workers in the United States. The insignificance of economic considerations in explaining the Brexit vote also suggests the negative correlation between education and voting to leave the European Union is not driven by economic interests, but instead by how education is related to voters' values, identities, and information sets. However, it is consistent with evidence that economic self-interest is less important in explaining attitudes towards immigration than cultural attachments and concerns about how immigration affects the nation as a whole (Hainmueller and Hopkins 2014).

So why did left-behind voters back Brexit? Ruling out the economics of European Union membership as a cause leaves two plausible hypotheses for why Britain voted to leave.

Hypothesis 1: Primacy of the Nation-State. Successful democratic government requires the consent and participation of the governed. British people identify as citizens of the United Kingdom, not citizens of the European Union. Consequently, they feel that the United Kingdom should be governed as a sovereign nation-state. EU membership erodes Britain's sovereignty. In particular, it prevents the UK from controlling immigration and forces the UK to implement laws made by the EU. According to this hypothesis, British people voted to leave the EU because they want to take back control of their borders and their country.

Hypothesis 2: Scapegoating of the EU. Many people feel left-behind by modern Britain. The left-behind are older, less educated, more socially conservative, less economically successful and think life in Britain is getting worse not better. Since the global financial crisis, the UK's median wage has declined (Costa and Machin 2017). Influenced by the anti-EU sentiments expressed by Britain's newspapers and eurosceptic politicians, these individuals have come to blame immigration and the EU for many of their woes. According to this hypothesis, voters supported Brexit because they believe EU membership has contributed to their discontent with the status quo.

The nation-state hypothesis explains Brexit as an assertion of national identity, while the scapegoating hypothesis views Brexit as resulting from voters being misinformed about the effects of EU membership. It is likely that both hypotheses played some role in the referendum outcome, but the existing evidence is insufficient to assess their relative contributions. When leave voters are asked to explain their vote, national sovereignty and immigration are the most frequently cited reasons (see, for example, the survey data from Lord Ashcroft Polls discussed above), but these responses are consistent with either hypothesis. They could reflect voters' attachment to the UK as a nation-state, or they may mirror the language used by pro-Brexit newspapers and politicians. However, the implications of the hypotheses differ in important ways. If voters supported Brexit to reclaim sovereignty from the EU, then, provided they are willing to pay the economic price for leaving the Single Market, they will view Brexit as a success. But if misinformation drove support for Brexit, then leaving the EU will do nothing to mitigate voters' discontent. More broadly, the two hypotheses have quite different implications for how policymakers should respond to Brexit and for the future of European and global integration.

Brexit and the Future of International Integration

The nation-state hypothesis is closely related to Rodrik's (2011) idea that the global economy faces a political trilemma. Rodrik argues that nation-states, democratic politics, and deep international economic integration are mutually incompatible, and that countries can choose at most two of the three options. Viewed through this framework, the nation-state hypothesis sees the Brexit vote as a democratic response to the erosion of British sovereignty caused by EU membership. If this perspective is correct, it means the deep integration promoted by the EU is incompatible with national democracy. For Europe to remain democratic, either the people of Europe must develop a collective identity in place of their separate national identities or the supranational powers of the EU must be reduced. Otherwise, the tensions evident in the Brexit vote will recur in other countries and the EU may lose more members.

Two components of the EU's deep integration are obvious candidates for inclusion in any retrenchment: free movement of labor and the supremacy of EU law in regulating the Single Market.[8] The indivisibility of the "four freedoms" of movement of goods, capital, services, and labor within the Single Market is a core principle of the European Union, but, in practice, restrictions on immigration could coexist with free movement of goods, services, and capital, even if this would reduce economic efficiency. Similarly, harmonization of economic regulation throughout the EU may be economically desirable, but if the nation-state hypothesis holds, allowing greater diversity across countries may be a price that has to be paid to ensure the viability of the EU.

The nation-state hypothesis does not directly threaten the sustainability of shallow integration agreements that aim to lower tariffs and border nontariff barriers. This is evident in the British government's response to the Brexit vote. The government has chosen to assume that the nation-state hypothesis explains the referendum outcome, leading it to interpret the vote as a mandate for controlling immigration and withdrawing from the deep regulatory integration of the Single Market. At the same time, the United Kingdom has branded itself as a champion of free trade working towards "the reduction and ultimate elimination of trade barriers wherever they are found" (Fox 2016). Setting aside that leaving the Single Market

[8] Arguably, the single currency also belongs on this list, but since the UK has not adopted the euro, it did not feature prominently in the Brexit debate, and I will not discuss the euro in this article.

contradicts this aim, it is noteworthy the UK government does not view Brexit as part of a broader shift towards protectionism. Consistent with the government's position, Ballard-Rosa, Rickard, and Scheve (2017) present survey evidence showing there is widespread support for free trade and investment in the UK but that supporters and opponents of Brexit have different preferences over immigration and regulation.

The scapegoating hypothesis, on the other hand, assumes that support for Brexit results not from any particular consequence of EU membership, but from voters channelling their discontent with modern life against the European Union. Colantone and Stanig's (2016) finding that exposure to Chinese import competition had a positive effect on support for Brexit is consistent with scapegoating of the EU. The scapegoating hypothesis does not threaten the ideal of the EU as a supranational political project or provide an immediate reason to reconsider the desirability of deep integration. But it does pose a different challenge to the future of international integration. As long as geography continues to be an important determinant of group identity, international institutions will always be more vulnerable to losing popular support than domestic institutions. The scapegoating of outsiders is a recurring phenomenon in world history. Brexit illustrates how this can lead to outcomes that limit integration.

If the scapegoating hypothesis proves correct, policymakers seeking to promote European and global integration have two main options available. One option would be to channel popular protests against another target. Both eurosceptic and pro-EU politicians have proved willing to blame the European Union for problems with domestic origins, but this could change. For example, left-wing politicians could embrace a progressive populism that blames the financial industry, large corporations, and rich individuals for the economic malaise that has followed the global financial crisis.

Alternatively, policymakers in the United Kingdom and elsewhere could focus on tackling the underlying reasons creating discontent among left-behind voters. Addressing economic and social exclusion is a daunting challenge, but enacting policies to support disadvantaged households and regions and broaden access to higher education would be an obvious starting point. O'Rourke (2017) argues the EU should position itself as a port in the storm for anxious electorates and should respond to Brexit by renewing its commitment to protecting Europeans from economic shocks, partly by allowing greater flexibility for governments to implement shock-absorbing policies.

It is too soon to know whether Britain leaving the European Union will prove to be merely a diversion on the path to greater integration, a sign that globalization has reached its limits, or the start of a new era of protectionism. In the year since the Brexit vote, EU leaders have worked to ensure Brexit does not lead to other countries leaving the union and, in the short-run at least, they have succeeded. A dialogue on the longer-term implications of Brexit has also started to develop, demonstrating how the referendum has made new futures imaginable. For example, the European Commission has issued a white paper laying out scenarios for Europe's future that include not only muddling through or committing to closer integration, but also scaling back the EU to just the Single Market or building a multi-speed Europe (European Commission 2017). Understanding and responding to the motivations of voters who oppose the European Union will play an important role in determining which of these futures comes to pass and whether the many benefits of economic and political integration can be preserved.

■ *I would like to thank Hanwei Huang and Kohei Takeda for excellent research assistance. I am also grateful to Dennis Novy, Gianmarco Ottaviano, and the editors for helpful comments and suggestions.*

References

Aichele, Rahel, and Gabriel Felbermayr. 2015. "Costs and Benefits of a United Kingdom Exit from the European Union." GED Study. Bertelsmann Stiftung.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103(6): 2121–68.

Baier, Scott L., Jeffrey H. Bergstrand, Peter Egger, and Patrick A. McLaughlin. 2008. "Do Economic Integration Agreements Actually Work? Issues in Understanding the Causes and Consequences of the Growth of Regionalism." *World Economy* 31(4): 461–97.

Ballard-Rosa, Cameron, Stephanie Rickard, and Kenneth Scheve. 2017. "Liberal Populism: Public Support for Globalization in Post-Brexit United Kingdom." Unpublished paper, London School of Economics.

Becker, Sascha O., and Thiemo Fetzer. 2016. "Does Migration Cause Extreme Voting?" CAGE Working Paper 306.

Becker, Sascha O., Thiemo Fetzer, and Dennis Novy. 2017. "Who Voted for Brexit? A Comprehensive District-Level Analysis." CEPR Discussion Paper 11954.

Booth, Stephen, Christopher Howarth, Mats Persson, Raoul Ruparel, and Pawel Swidlicki. 2015. "What If…? The Consequences, Challenges and Opportunities Facing Britain Outside EU." Report 03/2015. London: Open Europe.

Borchert, Ingo. 2016. "Services Trade in the UK: What is at Stake?" UK Trade Policy Observatory Briefing Paper 6.

Bruno, Randolph, Nauro Campos, Saul Estrin, and Meng Tian. 2016. "Gravitating towards Europe: An Econometric Analysis of the FDI Effects of EU Membership." Technical Appendix to "The Impact of Brexit on Foreign Investment in the UK." http://cep.lse.ac.uk/pubs/download/brexit03_technical_paper.pdf.

Burstein, Ariel, and Jonathan Vogel. Forthcoming. "International Trade, Technology, and the Skill Premium." *Journal of Political Economy.*

Bustos, Paula. 2011. "Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinian Firms." *American Economic Review* 101(1): 304–340.

Caliendo, Lorenzo, and Fernando Parro. 2014. "Estimates of the Trade and Welfare Effects of NAFTA." *Review of Economic Studies* 82(1): 1–44.

Campos, Nauro F., Fabrizio Coricelli, and Luigi Moretti. 2014. "Economic Growth and Political Integration: Estimating the Benefits from Membership in the European Union Using the Synthetic Counterfactuals Method." IZA Discussion Paper 8162.

Ciuriak, Dan, Jingliang Xiao, Natassia Ciuriak, Ali Dadkhah, Dmitry Lysenko, and G. Badri Narayanan. 2015. "The Trade-related Impact of a UK Exit from the EU Single Market." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2620718.

Colantone, Italo, and Piero Stanig. 2016. "Global Competition and Brexit." Available at:

http://www.italocolantone.com/research.html.

Costa, Rui, and Stephen Machin. 2017. "Real Wages and Living Standards in the UK." CEP Election Analysis no. 36.

Crafts, Nicholas. 2016. "The Growth Effects of EU Membership for the UK: A Review of the Evidence." Working Paper 280, University of Warwick. http://www2.warwick.ac.uk/fac/soc/economics/research/centres/cage/manage/publications/280-2016_crafts.pdf.

Dhingra, Swati, Hanwei Huang, Gianmarco I. P. Ottaviano, Joao Paulo Pessoa, Thomas Sampson, and John Van Reenen. 2017. "The Costs and Benefits of Leaving the EU: Trade Effects." CEP Discussion Paper 1478.

Dhingra, Swati, Gianmarco Ottaviano, Thomas Sampson, and John Van Reenen. 2016. "The Impact of Brexit on Foreign Investment in the UK." CEP Brexit Analysis no. 3.

Dhingra, Swati, and Thomas Sampson. 2016. "Life After Brexit: What are the UK's Options Outside the European Union?" CEP Brexit Analysis no. 1.

Djankov, Simeon. 2017. "The City of London After Brexit." Peterson Institute for International Economics Policy Brief 17-9.

Dustmann, Christian, Tommaso Frattini, and Ian P. Preston. 2013. "The Effect of Immigration along the Distribution of Wages." *Review of Economic Studies* 80(1): 145–73.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70(5): 1741–79.

Ebell, Monique. 2016. "Assessing the Impact of Trade Agreements on Trade." *National Institute Economic Review* 238(1): R31–R42.

European Commission. 2017. "White Paper on the Future of Europe: Reflections and Scenarios for the EU27 by 2025." Available at: https://ec.europa.eu/commission/white-paper-future-europe-reflections-and-scenarios-eu27_en.

European Council. 2017. "European Council (Art. 50) Guidelines for Brexit Negotiations." April 29. http://www.consilium.europa.eu/en/press/press-releases/2017/04/29-euco-brexit-guidelines/.

Feyrer, James. 2009. "Trade and Income—Exploiting Time Series in Geography." NBER Working Paper 14910.

Financial Times. 2017. "Angela Merkel Pledges to Block Brexit 'Cherry Picking.'" January, 18. https://www.ft.com/content/724ee76a-dd95-11e6-9d7c-be108f1c1dce.

Fox, Liam. 2016. Liam Fox's Speech to the World Trade Organization, December, 1. https://www.gov.uk/government/speeches/liam-foxs-speech-to-the-world-trade-organisation.

Goodwin, Matthew J., and Oliver Heath. 2016a. "The 2016 Referendum, Brexit and the Left Behind: An Aggregate-level Analysis of the Result." *Political Quarterly* 87(3): 323–32.

Goodwin, Matthew, and Oliver Heath. 2016b. "Brexit Vote Explained: Poverty, Low Skills and Lack of Opportunities." Joseph Rowntree Foundation, https://www.jrf.org.uk/report/brexit-vote-explained-poverty-low-skills-and-lack-opportunities.

Hainmueller, Jens, and Daniel J. Hopkins. 2014. "Public Attitudes toward Immigration." *Annual Review of Political Science* 17: 225–49.

Head, Keith, and Thierry Mayer. 2015. "Brands in Motion: How Frictions Shape Multinational Production." CEPR Discussion Paper DP10797.

HM Government. 2016. "Alternatives to Membership: Possible Models for the United Kingdom Outside the European Union." March. Available at: https://www.gov.uk/government/publications/alternatives-to-membership-possible-models-for-the-united-kingdom-outside-the-european-union.

Hobolt, Sara B., and Catherine E. de Vries. 2016. "Public Support for European Integration." *Annual Review of Political Science* 19: 413–32.

Hofmann, Claudia, Alberta Osnago, and Michele Ruta. 2017. "Horizontal Depth: A New Database on the Content of Preferential Trade Agreements." Policy Research Working Paper 7981, World Bank.

House of Commons. 2013. "Leaving the EU." Commons Briefing papers RP 13/42, July 1.

Kehoe, Timothy J., Pau S. Pujolas, and Jack Rossbach. 2016. "Quantitative Trade Models: Developments and Challenges." NBER Working Paper 22706.

Krugman, Paul R. 1979. "Increasing Returns, Monopolistic Competition, and International Trade." *Journal of International Economics* 9(4): 469–79.

Lord Ashcroft Polls. 2016. "How the United Kingom Voted on Thursday . . . and Why." June 24. http://lordashcroftpolls.com/2016/06/how-the-united-kingdom-voted-and-why/.

May, Theresa. 2017. "The Government's Negotiating Objectives for Exiting the EU: PM Speech." Lancaster House speech, January 17. https://www.gov.uk/government/speeches/the-governments-negotiating-objectives-for-exiting-the-eu-pm-speech.

Méjean, Isabelle, and Cyrille Schwellnus. 2009. "Price Convergence in the European Union: Within Firms or Composition of Firms?" *Journal of International Economics* 78(1): 1–10.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-industry Reallocations and Aggregate

Industry Productivity." *Econometrica* 71(6): 1695–1725.

Mulabdic, Alen, Alberto Osnago, and Michele Ruta. 2017. "Deep Integration and UK–EU Trade Relations." Policy Research Working Paper WPS7947.

Nickell, Stephen, and Jumana Saleheen. 2015. "The Impact of Immigration on Occupational Wages: Evidence from Britain." Staff Working Paper 574, Bank of England. December 18.

Oliver Wyman. 2016. "The Impact of the UK's Exit from the EU on the UK-based Financial Services Sector." Available at: http://www.oliverwyman.com/our-expertise/insights/2016/oct/The-impact-of-Brexit-on-the-UK-based-Financial-Services-sector.html.

Office for National Statistics. 2016a. "UK Balance of Payments, the Pink Book: 2016."

Office for National Statistics. 2016b. "UK National Accounts, the Blue Book: 2016."

Office for National Statistics. 2017. "Preliminary Estimate of GDP Time Series Dataset." July, 26.

Open Europe. 2015. "The Top 100 Costliest EU-derived Regulations in Force in the UK." A table. http://2ihmoy1d3v7630ar9h2rsglp.wpengine.netdna-cdn.com/wp-content/uploads/2015/03/Open_Europe_Top100_costliest_EU_regulations.pdf.

O'Rourke, Kevin H. 2017. "Brexit, Political Shock Absorbers, and the Three Rs." Chap. 10 in *Quo Vadis? Identity, Policy and the Future of the European Union*, edited by Thorsten Beck and Geoffrey Underhill. A VoxEU.org Book. CEPR.

Pisani-Ferry, Jean, Norbert Röttgen, André Sapir, Paul Tucker, and Guntram B. Wolff. 2016. "Europe after Brexit: A Proposal for a Continental Partnership." Bruegel External Publication, Brussels.

Portes, Jonathan, and Giuseppe Forte. 2017. "The Economic Impact of Brexit-induced Reductions in Migration." *Oxford Review of Economic Policy* 33(Supplement 1): S31–S44.

Rodrik, Dani. 2011. "The Globalization Paradox: Democracy and the Future of the World Economy." WW Norton & Company.

Sampson, Thomas. 2016. "Dynamic Selection: An Idea Flows Theory of Entry, Trade and Growth." *Quarterly Journal of Economics* 131(1): 315–80.

Wadsworth, Jonathan, Swati Dhingra, Gianmarco Ottaviano, and John Van Reenen. 2016. "Brexit and the Impact of Immigration on the UK." CEP Brexit Analysis no. 5.

World Bank. 2017. World Development Indicators. http://data.worldbank.org/data-catalog/world-development-indicators.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 185–204

Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa

Tessa Bold, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane

School enrollment has universally increased over the last 25 years in low-income countries. Enrolling in school, however, does not assure that children learn. A large share of children in low-income countries complete their primary education lacking even basic reading, writing, and arithmetic skills (Hungi et al. 2010; PASEC 2015; ASER 2014)—a state of affairs that UNESCO (2013) dubbed the "global learning crisis." For example, after more than three years of compulsory language teaching, four out of five students in Mozambique and Nigeria cannot read simple words of Portuguese and English, respectively. Only one-quarter of Indian students in grade four can manage tasks—such as basic subtraction—that are part of the curriculum for the second grade. Roughly half of the students in Uganda, after three years of mathematics teaching, cannot place numbers between 0 and 999 in order (Bold et al. 2017a; ASER 2013).

A growing body of evidence, from both the teacher value-added literature and the experimental literature in development economics, shows that teacher quality is a key determinant of student learning, although other factors also play an

■ *Tessa Bold is Assistant Professor, Institute for International Economic Studies, Stockholm University, Stockholm, Sweden. Deon Filmer is a Lead Economist and co-Director of the 2018 World Development Report, Gayle Martin is a Program Leader, Ezequiel Molina is an Economist, and Christophe Rockmore and Waly Wane are Senior Economists, all at the World Bank, Washington, DC. Brian Stacy is an Economist, Economic Research Services, US Department of Agriculture, Washington, DC. Jakob Svensson is Professor of Economics, Institute for International Economic Studies, Stockholm University, Stockholm, Sweden. Bold is the corresponding author, and her email address is tessa.bold@iies.su.se.*

[†] For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.185 doi=10.1257/jep.31.4.185

important role. Little is known, however, about what specific dimensions of teacher quality matter, and even less about how teachers perform along these dimensions—facts that are crucial in order to guide both research and policy design. This paper discusses an ongoing research program intended to help fill this void. Using data collected through direct observations, unannounced visits, and tests from primary schools in seven sub-Saharan African countries which together represent close to 40 percent of the region's total population, we answer three questions: How much do teachers teach? What do teachers know? How well do teachers teach?

The answers to these questions should be interpreted against the backdrop of a rapidly expanding, but weakly governed, primary education sector in sub-Saharan Africa. Gross primary enrollment rates in sub-Saharan Africa have increased from around 50 percent in 1970, to 98 percent in 2014, and net enrollment rates (children enrolled divided by total children in the age group) have increased from around 40 percent to almost 80 percent, partly in response to reduced or removed formal fees for primary schooling. Most of the increase in enrollment has taken place in the public sector, which remains the dominant actor in the sector. The increase in primary enrollment has also resulted in a huge increase in the number of teachers, which has risen from 500,000 primary school teachers in 1970 to almost 2.8 million in 2009. The salaries of these teachers make up more than 70 percent of the expenditure in education (UNESCO Institute for Statistics 2011) and approximately 12 percent of total government expenditure in the nations of sub-Saharan Africa.

The provision of education in many low-income countries, including the countries surveyed here, is characterized by a combination of centralized but typically weak state control and often low-capacity, locally governed institutions. At the same time, the institutional incentives for teacher performance are largely missing, with both career progression and financial rewards delinked from performance. Teachers' salaries and promotions are largely determined by seniority and educational qualifications, and are unrelated to effort or performance. In most settings, parents have little influence over how teachers are hired or schools are managed, and the various state and local authorities provide limited technical support or supervision.

In the sections that follow, we draw upon evidence from the Service Delivery Indicators program—an ongoing Africa-wide program with the aim of collecting informative and standardized measures of what primary teachers know, what they do, and what they have to work with. The Service Delivery Indicators program was piloted in Tanzania and Senegal in 2010 (Bold, Gauthier, Svensson, and Wane 2010; Bold, Svensson, Gauthier, Maestad, and Wane 2011).[1] To date, the program has collected data, including from the two pilot countries, from a total of seven countries (eight surveys): Kenya (2012), Mozambique (2014), Nigeria (2013), Senegal (2010), Tanzania (2010, 2014), Togo (2013), and Uganda (2013). Primary schools

[1] The Service Delivery Indicators program grew out of concern about poor learning outcomes observed in various student tests as well as evident shortcomings most clearly (and perhaps most damagingly) manifested at the school level in fast-expanding systems of education.

with at least one fourth-grade class formed the sampling frame.[2] The samples were designed to provide representative estimates for teacher effort, knowledge, and skills in public primary schools, broken down by urban and rural location. For five of the six nonpilot surveys, representative data were also collected for private primary schools. Private schools—both informal and formal—account for around 20 percent of total primary school enrollment in low-income countries (Baum, Lewis, Lusk-Stover, and Patrinos 2014). The surveys collected a broad set of school-, teacher-, and student-specific information, with an approach that relies as much as possible on direct observation—such as visual inspections of fourth-grade classrooms and the school premises, direct physical verification of teacher presence by unannounced visits, and teacher and student tests—rather than on respondent reports.

For the countries covered by the survey, we address the three questions posed above. We then provide some explanation for the results by discussing what the pipeline to a teaching position looks like, what kind of teachers emerge from it, and what incentives these teachers face to teach well when deployed in schools. Finally, we conclude with a brief discussion of the core implications of the findings, both for education systems and education policy reform and for the experimental and quasi-experimental research agenda on ways to improve education quality.

How Much Time Do Teachers Teach?

Being present in the classroom is a *conditio sine qua non* for teachers to exert effort at teaching. To measure the time teachers spend teaching, an extended version of the approach described in this journal by Chaudhury, Hammer, Kremer, Muraldiharan, and Rogers (2006) was employed. In each school, during a first announced visit, up to ten teachers were randomly selected from the teacher roster. At least two teaching days after the initial survey, an unannounced visit was conducted, during which the enumerators were asked to identify whether the selected teachers were in the school, and if so, if they were in class teaching. Both assessments were based on directly observing the teachers and their whereabouts.

Table 1 summarizes the findings (and the online Appendix available with this paper at http://e-jep.org provides country-specific details). Averaging across countries, 44 percent of teachers were absent from class, either because they were absent from school or in the school, but not in the classroom. In three of the eight surveys, more than half of the teachers were absent from the classroom, and only in one country—Nigeria—do we observe average absence below 30 percent. Being absent

[2] In each country, representative surveys of between 150 and 760 schools were implemented using a multistage, cluster-sampling design. In Nigeria, due to security constraints, surveys representative at the state level were implemented in four states (Anambra, Bauchi, Ekiti, and Niger). Across the eight surveys, the Service Delivery Indicators survey collected data on 2,600 schools, over 21,000 teachers, and 24,000 students in sub-Saharan Africa (for details of the sample, see Bold et al. 2017a).

Table 1

Teacher Absence

	All	*Min*	*Max*
Absence from class	44%	23% (Nigeria)	57% (Uganda)
Absence from school	23%	15% (Kenya, Tanzania survey II)	45% (Mozambique)
Number of teachers	**16,543**		
Scheduled teaching time	5h 27m	4h 21m (Mozambique)	7h 13m (Uganda)
Time spent teaching	2h 46m	1h 43m (Mozambique)	3h 10m (Nigeria)
Number of schools	**2,001**		
Orphaned classrooms	33%	24% (Togo)	45% (Uganda)
Number of schools	**1,647**		

Notes: The table reports the absence rate for all teachers in government school, the scheduled teaching time, actual teaching time, and share of orphaned classrooms for all government schools. All individual country statistics are calculated using country-specific sampling weights. The average for all countries, reported under the heading "All" is taken by averaging over the country averages. The names of the countries with the lowest and highest score for each item are given in parentheses. Teachers are marked as absent from school if during the second unannounced visit, they are not found anywhere on the school premises. Otherwise, they are marked as present. Teachers are marked as absent from class if during the second unannounced visit, they are absent from school or present at school but absent from the classroom. Otherwise, they are marked as present. The scheduled teaching time is the length of the school day minus break time. Time spent teaching adjusts the length of the school day by the share of teachers who are present in the classroom, on average, and the time the teacher spends teaching while in the classroom. The orphaned classrooms measure is the ratio of the classrooms with students but no teacher to the number of classrooms with students with or without a teacher (not collected for the pilot countries). For country-specific estimates, see the Online Appendix.

from school is about as common as being present in the school, but absent from class. The rank correlation coefficient between the two measures is less than 0.5 at the country level, making the *school* absence rate at best a partial measure of teacher effort. This is most starkly illustrated in the cases of Kenya and Tanzania, both of which have relatively low school absence rates (15 percent in each case) but relatively high classroom absence rates conditional on being in school (about 38 percent in each case).

When a large share of teachers is not teaching, unsurprisingly, a large share of classrooms will be occupied by only students. Consistent with the absenteeism findings discussed above, we find, averaging across countries, that one-third of the classrooms were "orphaned" classrooms, where students are present but there is no teacher. And in Uganda, almost one-half of the classrooms were orphaned.

Over time, these absenteeism rates appear remarkably stable. In this journal, Chaudhury et al. (2006) estimated a school absence rate of 27 percent in Uganda in 2002–03, which compares to our measure of 28 percent in 2013. Similarly, while absence from school fell by one-third in Tanzania between 2010 and 2014, this was largely offset by an increase in absence from the classroom while being in school; the net result being a small decline in absence from class between the two surveys.

What do these results imply for the amount of instruction time that students receive? To answer this, the surveys first recorded the scheduled time of a teaching day—after break times—according to school records. Averaged across schools and countries, this comes to 5 hours and 27 minutes. We then multiply this number by the proportion of teachers present in class. If ten teachers are supposed to teach 5 hours and 27 minutes per day, yet four teachers are absent from either the school or the classroom at any one time, then the scheduled teaching time is reduced to 3 hours and 16 minutes.

Moreover, even when in the classroom, teachers may not necessarily be teaching. We carried out classroom observation as part of the survey, recording a minute-by-minute snapshot of what the teacher was doing, for a randomly selected fourth-grade mathematics or language class. The percentage of the lesson lost to nonteaching activities varied from 18 percent in Nigeria, the country with the lowest classroom absence rate, to 3 percent in Uganda, the country with the highest classroom absence rate. We then combine the absence-adjusted teaching time with the proportion of classroom time devoted to actual teaching activities to estimate instruction time as experienced by students.

Students are taught, on average, 2 hours and 46 minutes per day, or roughly half of the scheduled time (as shown in Table 1). Estimated instruction time varies from 3 hours and 10 minutes in Nigeria to 1 hour and 43 minutes in Mozambique. About 10 percent of the schools provide more than 5 hours of teaching per day. About the same share provide no teaching (because none of the ten randomly selected teachers was found in the classroom). More than a quarter of schools teach less than two hours, and half the schools teach less than three hours. To put this in perspective, on average across the OECD countries, the compulsory instructional time per school day in primary education is about 4.5 hours (OECD 2015).

Our results on teacher absence and time in the classroom are broadly similar to findings from other studies. In this journal, Chaudhury et al. (2006) present results from a multicountry study spanning Asia, Africa, and Latin America, where enumerators made unannounced visits to public schools to measure teacher presence in schools. Pooling data across countries, they find an average teacher absence rate of 19 percent, which is similar to the 23 percent absence rate we report in Table 1. Bruns and Luque (2014) further document, drawing on data from a large sample of classrooms in seven Latin American and Caribbean countries, that teachers only spend 52–85 percent of class time on academic activities, implying a loss of potential instructional time equivalent to one day of instruction per week. Consistent with the findings we report here, they also show that in every Latin American and Caribbean country studied, teachers in classrooms spend about 10 percent of time completely "off-task." In India, Kremer et al. (2005) report that not only were 25 percent of teachers absent from work, but another 25 percent were in school but not teaching and thus only about half of the teachers were found to be actually engaged in teaching, again a result strikingly close to what we document across the seven countries we surveyed.

What Do Teachers Know?

To measure the subject content knowledge of primary school teachers, and specifically those teaching in the lower primary grades, language and mathematics teachers teaching Grade 4 in the current year (or Grade 3 in the previous year) were assessed. (The idea was to sample the teachers who taught the students we sampled in the current year and the previous year.) On average, five teachers were tested in each school. In contrast to other approaches to assess the knowledge of teachers—for example, having teachers take exams—teachers here were asked to mark (or "grade") mock student tests in language and in mathematics.[3] This method of assessment has two potential advantages. First, it aims to assess teachers in a way that is consistent with their normal activities—namely, marking student work. Second, by not testing teachers in the same way as students are tested, it recognizes teachers as professionals. In the analysis, we assess the language knowledge of those teachers who teach language, and the mathematics knowledge of those teachers who teach mathematics. All questions on the teacher test were based on common items taken from the primary curricula of each country.

We start by assessing language tasks on the teacher test that covered (roughly) the lower primary curriculum (first to third year of primary school)—specifically, spelling and simple grammar exercises. We count a teacher as "mastering" the student curriculum if he or she marked 80 percent or more of the spelling and grammar questions correctly. Two-thirds of teachers make it over this bar, though with wide variation across countries, as shown in Table 2. While over 90 percent of teachers in Kenya and Uganda master the knowledge that their students are supposed to learn, only one-quarter of Nigerian teachers do.

Possessing knowledge equivalent to the fourth-grade curriculum is, of course, not sufficient to teach language in lower primary grades because language teaching is "monolithic." That is to say, teaching a student how to compose even a simple text requires knowledge that goes well beyond what is listed in the curriculum. We therefore deem a language teacher in Grade 4 to have minimum subject content knowledge if the teacher can competently correct children's work in such aspects of literacy as reading comprehension, vocabulary, and formal correctness (grammar, spelling, syntax, and punctuation), all of which are competencies a teacher in lower primary would routinely be required to use. To this end, the language test contained (in addition to the spelling and grammar exercises) items involving sentences with blank spaces where students need to fill in words—so-called "Cloze" passages—to assess vocabulary and reading comprehension, and a letter written to a friend describing the student's school, which the teacher had to mark and correct.

[3] The subject test was designed by experts in international pedagogy and validated against 13 sub-Saharan African primary curricula and national teacher standards (Botswana, Ethiopia, Gambia, Kenya, Madagascar, Mauritius, Namibia, Nigeria, Rwanda, Seychelles, South Africa, Tanzania, and Uganda). See Johnson, Cunningham, and Dowling (2012) for details.

Table 2
Teachers' Content Knowledge: Minimum Thresholds

	All	*Min*	*Max*
Subject knowledge: Language			
Teachers with …			
80% of knowledge equivalent to a 4th grader	66%	26% (Nigeria)	94% (Kenya)
Minimum knowledge for teaching	7%	0% (Mozambique, Nigeria, Tanzania survey I, Togo)	34% (Kenya)
Number of teachers	3,770		
Subject knowledge: Mathematics			
Teachers with …			
Minimum knowledge for teaching	68%	49% (Togo)	93% (Kenya)
Number of teachers	3,957		

Notes: The table reports minimum content knowledge indicators for teachers in grade 4 or who taught grade 3 in the previous year in government schools. Language knowledge is computed for teachers teaching language, and mathematics knowledge is computed for teachers teaching mathematics. All individual country statistics are calculated using country-specific sampling weights. The average for all countries, reported under the heading "All," is taken by averaging over the country averages. Names of the countries with the lowest (Min) and highest (Max) score for each item are given in parentheses. A language teacher is defined as "mastering" the student curriculum if he/she scores at least 80 percent on the tasks covered in the language curriculum up to grade 4. A language teacher is defined as having minimum knowledge for teaching if he/she scores at least 80 percent on the grammar, Cloze test, and correcting a student's composition task of the language assessment. A mathematics teacher is defined as having minimum knowledge for teaching if he/she scores at least 80 percent on the tasks covered in the math curriculum up to grade 4. (So, for mathematics, the two measures—minimum knowledge and 80 percent of knowledge equivalent to a fourth grader, are the same; for language, they are different.) For country-specific estimates, see the Online Appendix.

For the language subject area, we formally define "minimum knowledge for teaching" as marking at least 80 percent of the items on the language test correctly. Only 7 percent of the language teachers meet this minimum, with the level uniformly low across the eight countries: in Kenya, 34 percent of language teachers have minimum knowledge for teaching, and no teachers in Togo, Mozambique, Tanzania (survey 1), or Nigeria meet the threshold (again as shown in Table 2).

Which areas of language teaching are especially problematic? Table 3 offers a breakdown of specific tasks on the language and math tests. Teachers could complete simple language and grammar tasks: the average score on a task that asked teachers to spell simple words ("traffic," for example) was 86 percent, and teachers got about 80 percent of simple grammar exercises correct that asked them to identify the option, out of three, that would complete a sentence such as "[_____] [Who, How much, How many] oranges do you have?" Teachers struggled with those tasks that required at least some knowledge beyond the lower primary curriculum to mark. Less than half of the items in the Cloze passage were marked correctly, which included "student" responses such as "[Where] do I have to go to the market?" (In this case, a correct answer could be either "Why or When.")

Table 3
Teachers' Performance on Specific Item Groups of Knowledge

	All	*Min*	*Max*
Language (*score out of 100*)			
Spelling task[a]	86	86 (Tanzania, survey I)	86 (Tanzania, survey I)
Grammar task	79	58 (Nigeria)	92 (Kenya)
Cloze task	44	27 (Togo)	66 (Kenya)
Correct composition task	26	9 (Mozambique)	50 (Kenya)
Number of teachers, Language	**3,770**		
Math (*percent of teachers*)			
Can add double digits	91%	75% (Togo)	98% (Kenya)
Can subtract double digits	76%	59% (Nigeria)	93% (Tanzania, survey I)
Can multiply double digits	68%	44% (Mozambique)	89% (Senegal)
Can solve simple math story problem	55%	17% (Mozambique)	91% (Senegal)
Understands a Venn diagram[b]	31%	12% (Mozambique)	56% (Kenya)
Can interpret data in a graph[b]	11%	3% (Mozambique)	40% (Kenya)
Can solve algebra	35%	3% (Mozambique)	74% (Kenya)
Can solve difficult math story problem[c]	15%	7% (Senegal)	22% (Tanzania, survey I)
Number of teachers, Math	**3,957**		

Notes: The table presents scores on Language tasks, and the percentage of teachers able to perform various math tasks, for teachers in government schools teaching grade 4 or who taught grade 3 in the previous year. Language knowledge is computed for teachers teaching language, and mathematics knowledge is computed for teachers teaching mathematics. All individual country statistics are calculated using country-specific sampling weights. The average for all countries, reported under the heading "All," is taken by averaging over the country averages. Names of the countries with the lowest (Min) and highest (Max) score for each item are given in parentheses. For country-specific estimates, see the Online Appendix.
[a] Question was asked only in Tanzania (2010).
[b] Percentage of teachers who got both questions related to this task correct.
[c] Question was asked only in Senegal and Tanzania (2010).

Teachers corrected a quarter of the spelling, grammar, syntax, and punctuation mistakes in a child's letter that included segments such as "I went to tell you that my new school is better the oldone I have a lot of thing to tell you about my new school in Dar es Salaam."

In mathematics, a teacher is defined to have minimum subject content knowledge if the teacher can accurately correct children's work in such aspects of numeracy as manipulating numbers and using whole number operations. This requirement amounts to correctly scoring 80 percent or more of the questions on the lower primary portion of the mathematics test. The test thus measures whether the math teacher masters his or her students' curriculum, allowing for 20 percent points margin of error. Around 70 percent of mathematics teachers have minimum knowledge according to this definition (as shown in Table 2), and there is again wide variation across countries, with less than one-half of the mathematics teachers in Togo deemed to have minimum knowledge. Looking at specific tasks in mathematics listed in Table 3, almost one-quarter of the teachers cannot subtract double-digit numbers and one-third of the teachers cannot multiply double-digit numbers.[4]

Of course, we would expect a competent math teacher to have knowledge beyond that of his or her students, and the mathematics test, therefore, also included questions one would only encounter in upper primary school. Many mathematics teachers struggled with these tasks: only a minority of teachers, and in some countries very few, could interpret information in a Venn diagram and/or a graph, as shown in Table 3. As we will see below, this low competence in interpreting data has implications for teachers' ability to monitor their students' progress. Finally, only a few teachers could solve a more advanced math story problem, and one-third could solve a simple algebraic equation.

There are few direct studies outside of Africa about how much teachers know about the subjects they teach, but those available show similarly very low results.[5] Bruns and Luque (2014) report findings from a national evaluation of teachers (and students) by the Ministry of Education in Peru. More than eight of ten sixth-grade teachers scored below level 2 on a 2006 test where level 3 meant mastery of sixth-grade math skills, and performance below level 2 implied the "teachers were unable to establish mathematical relationships and adapt routine and simple mathematical procedures and strategies."

How Well Do Teachers Teach?

Good teaching also requires that teachers know how to translate their subject knowledge into effective pedagogy and then apply this in the classroom. Teachers must also know how to assess student capabilities and react appropriately, for example, by asking questions that require various types of responses and by giving feedback on those responses, commonly referred to as "knowledge of the context of learning" (Johnson 2006; Danielsson 2007; Pianta, La Paro, and Hamre 2007; Coe, Aloisi, Higgins, and Major 2014; Ko and Sammons 2013; Mujis et al. 2014; Vieluf, Kaplan, Klieme, and Bayer 2012). In a recent review, although not focused on Africa specifically, Mujis et al. (2014) identify a set of skills and practices in the classroom that are consistently associated with gains in student learning: 1) structuring lessons, and in particular, introducing topics and learning outcomes

[4] Our two measures of teacher knowledge—knowing the students' curriculum and minimum knowledge for teaching—coincide for mathematics but not for language teaching. The reasoning here is that it is possible, in principle, to teach fourth graders how to divide two numbers without having a deeper knowledge of algebra. As a consequence, the number of teachers considered to "master" their students' curriculum is very similar for language and mathematics, while there is a large difference in the number of teachers considered to have "minimum knowledge" for teaching between the two subjects.

[5] The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) collects average achievement scores of grade 6 teachers (data collected in 1995, 2000, and 2007). However, SACMEQ only reports scale scores for teachers (for example, Makuwa 2011), which makes it possible to do comparisons of teacher test scores over time and across the participating countries but makes it difficult to assess teachers' absolute subject knowledge.

at the start of the lesson and reviewing them at the end; 2) frequently checking for student understanding by asking questions, and allowing time for students to review and practice what they learned, either individually or in groups; 3) varying the cognitive level of questions by mixing lower- and higher-order questions; and 4) providing substantive feedback to students by acknowledging correct answers in a positive fashion and correcting wrong answers. To assess how well teachers teach, therefore, we first measure teachers' pedagogical knowledge; then, we examine how well teachers can assess students and monitor their progress; and finally, we gauge the extent to which teachers apply that knowledge in the classroom based on direct lesson observation.[6]

To measure general pedagogical knowledge, we asked teachers to prepare for a lesson with a specified topic by reading and extracting information from a factual text on that topic (general content knowledge) and to state (in 1–2 sentences) what they would expect their students to learn from the lesson. Both these tasks are consistent with professional tasks normally expected of primary teachers, and we therefore consider a teacher who scores 80 percent or more on this portion of the test to have minimum general pedagogy knowledge.

To measure teachers' ability to assess students' learning and give feedback (which we shorten here to "assessing students"), teachers were asked to prepare questions that required students to recall what was learned (lower order) and questions that asked students to apply the material to new contexts (higher order) on the basis of their reading of the factual text. In a second task, teachers were asked to use a marking scheme to give feedback on strengths and weaknesses in students' writing and to distinguish weak and strong learners. In a third task, teachers were provided with a list of students' grades; they were then asked to turn the raw scores into averages and to comment on the learning progression of individuals and groups of students with the help of a bar chart. We define a teacher as having "minimum knowledge in assessing students" if he or she could answer 80 percent of the items in the three tasks correctly.

As reported in Table 4, Panel A, 11 percent of teachers reached the threshold for minimum general pedagogy knowledge. In four countries, fewer than 5 percent of teachers met the threshold. While teachers could usually read and understand the factual text (average score of 47 percent), they were typically not able to translate this information into teaching, as they struggled to formulate what they wanted children to learn from the lesson based on their reading (average score of 23 percent).

As with general pedagogical knowledge, the results in Panel B show that few teachers demonstrated an ability to assess student learning and respond to that assessment. Not many could formulate questions that checked basic understanding based on what they had read, and fewer still could formulate a question that asked students to apply what they had learned to other contexts (average score of

[6] The observation schedule is based on a modified Stallings (1980) snapshot module.

Table 4
Pedagogical Knowledge and Skills

	All	*Min*	*Max*
Panel A: Pedagogical knowledge			
Minimum general pedagogy knowledge (% of teachers)	11%	1% (Nigeria)	36% (Tanzania)
Factual text comprehension (score out of 100)	47	23 (Mozambique)	78 (Tanzania)
Formulate aims and learning outcomes (score out of 100)	23	11 (Nigeria)	41 (Tanzania)
Number of teachers	4,799		
Panel B: Assessing students			
Minimum knowledge assessing students (% of teachers)	0%[a]	0%	0%
Formulate questions to check understanding (score out of 100)	23	5 (Nigeria)	55 (Kenya)
Formulate questions to apply to other contexts (score out of 100)	7	3 (Nigeria)	15 (Tanzania)
Assessing students' abilities (score out of 100)	19	8 (Nigeria)	39 (Kenya)
Evaluating students' progress (score out of 100)	12	5 (Nigeria, Mozambique)	26 (Kenya)
Number of teachers	4,799		
Panel C: Skills and practices in the classroom (% of teachers)			
Introduce and summarize topic of the lesson	41%	16% (Mozambique)	62% (Kenya)
Lesson appears planned to enumerator	64%	37% (Uganda)	75% (Kenya)
Ask a mix of lower and higher order questions	31%	14% (Mozambique)	44% (Uganda)
Give positive feedback, praise, corrects mistakes	52%	32% (Mozambique)	75% (Uganda)
Engages in all of the above practices	8%	1% (Mozambique)	17% (Kenya)
Number of teachers (classrooms)	1,551		

Notes: Panel A reports on minimum general pedagogical knowledge and scores on specific pedagogical tasks for teachers in government schools in grade 4 or who taught grade 3 in the previous year. A teacher is defined as having minimum knowledge of general pedagogy if the teacher scores at least 80 percent on the tasks that relate to general pedagogy (factual text comprehension and being able to formulate learning outcomes and lesson aims). Panel B reports minimum pedagogical knowledge in assessing students as well as scores on specific pedagogical tasks for teachers in government schools in grade 4 or who taught grade 3 in the previous year. A teacher in any subject is defined as having minimum knowledge for assessing students if they score least 80 percent on the tasks that relate to assessment (comparing students' writing and monitoring progress among a group of students). Panel C presents teacher skills and practices in the classroom in government schools in grade 4. The information is not available for Senegal or for Tanzania survey I. All individual country statistics are calculated using country-specific sampling weights. The average for all countries, reported under the heading "All," is taken by averaging over the country averages. Names of countries with the lowest (Min) and highest (Max) score for each item are given in parentheses. All scores are computed for teachers teaching either subject.

[a] No teacher assessed had minimum knowledge to assess students. For country-specific estimates, see the Online Appendix.

23 percent and 7 percent on these two tasks). The average score on a task that asked teachers to give feedback on strengths and weaknesses in student's writing using a marking scheme was 19 percent—ranging from 8 percent in Nigeria to 39 percent in Kenya. Furthermore, the ability to monitor and comment on the learning

progression of students was low (average score of 12 percent on this task)—ranging from 5 percent in Nigeria to 26 percent in Kenya.

Poor knowledge of general pedagogy was mirrored in behavior in the classroom, as shown in Panel C. Less than half of the teachers explained the topic of the lesson at the start and summarized what was learned at the end, and around 35 percent of lessons seemed unplanned to the observers. During their lessons, many teachers asked questions that required students to recall information or to practice what was learned, but significantly fewer asked questions that required higher-order skills and encouraged students to apply what was learned to different contexts and be creative. Overall, 31 percent of teachers mixed lower- and higher-order questions in their class—ranging from 14 percent of teachers in Mozambique to 44 percent of teachers in Uganda. In response to students' answers, around half the teachers consistently gave positive feedback and corrected mistakes without scolding students, with a low of 32 percent in Mozambique and a high of 75 percent in Uganda.

In summary, general pedagogical knowledge and the ability to assess students' learning and respond to that assessment is poor across the seven countries, with roughly 1 in 10 teachers being classified as having minimum knowledge in general pedagogy and none having minimum knowledge in student assessment. Inside the classroom, many teachers deploy some of the teaching practices identified in the literature as promoting learning, but few (less than one in ten) apply the full set of beneficial skills—structuring, planning, asking questions and giving feedback—in their lessons.

Our approach to assess how teachers perform in the classroom differs from other studies in that it combines observational data from inside the classroom with test results from pedagogical assessments of the teachers. As mentioned earlier, Bruns and Luque (2014) draw on data from a large sample of classrooms in seven Latin American and the Caribbean countries. Although students in their sample are offered a relatively enriched learning environment—in contrast to the typical primary school in sub-Saharan Africa—in the sense that students are almost universally equipped with workbooks and writing materials and textbooks are generally available, a significant share of students are visibly not involved in whatever activity the teacher is leading.

Comparing our findings with data from middle- and high-income countries, some interesting parallels emerge. Although teachers in high-income countries generally display better classroom practices than their counterparts in poorer countries (Araujo, Carneiro, Cruz-Aguayo, and Shady 2016; Bruns, de Gregorio, and Taut 2016), teachers show the same relative strengths and weaknesses across a variety of contexts and observation schedules. That is, they tend to perform relatively well when it comes to classroom management and creating a positive climate for their students, but less well when it comes to instructional support including using questions and discussion techniques as well as assessment in instruction (Bruns, De Gregorio, and Taut 2016; Kane and Staiger 2012; Tyler, Taylor, Kane, and Wooten 2010).

Why Does the System Used to Select, Train, and Remunerate Teachers Not Produce High-Quality Teaching?

Many low-income countries have witnessed a huge expansion in the provision of primary education in the last two decades: we find that twice as many teachers have entered the profession in sub-Saharan Africa in the last ten years than in the decade before. This expansion will likely continue. According to recent population projections, close to half the world population of children will live in Africa by the end of the 21st century (You, Hug, and Anthony 2014). Looking at a not-too-distant future, the number of children in the primary school age group in sub-Saharan Africa is set to rise from 170 million to 220 million in the next 15 years, reaching 280 million by the mid-century. Simply to keep pace with population growth—adjusting for teacher retirement—and to maintain pupil/teacher ratios at a rough benchmark of 40 students per teacher (the average in our sample is 45 students enrolled per teacher and 34 students present in the classroom per teacher), would require the hiring of two million new teachers by 2030 and five million by 2050.[7] Such a rapid expansion of the teaching force provides a real opportunity for updating the pipeline—an opportunity that will be lost if the system for selecting, training, and motivating teachers does not ensure good teachers in schools.

So why does the existing system not produce high-quality teaching, as suggested by the evidence presented above? We argue that there are two reasons: the system used to select and train teachers does not deliver high-quality candidates; and the system used to employ and remunerate teachers does not motivate them to deliver high-quality teaching.[8]

All seven countries we study possess de jure well-established systems of teacher training. To enter teacher training, teachers must have completed at least lower secondary education. In our sample, this is true for 45 percent of the teachers, while the majority of the remainder have either completed upper secondary education (28 percent) or post-secondary, non-tertiary education (19 percent). The length of teacher training courses varies among countries, ranging from two years in the case of Kenya, Tanzania, and Uganda to one year in the case of Senegal. At the end of the program, which mostly confers training at the post-secondary, non-tertiary level, teachers qualify with a teaching certificate, held by 90 percent of teachers in our sample. Ten percent of teachers hold (in addition to their certificate) a bachelor's or master's degree in education.

[7] We arrive at these numbers by linearly extrapolating number of births per year from 2000 to 2050 using data reported in You, Hug, and Anthony (2014) for years 1980, 2015, 2030, and 2050. We assume that under-five mortality in the region will fall from 90 per 1,000 live births in 2015 to 50 per 1,000 live births in 2050. Finally, we use our survey data to estimate the age profile of the current stock of teachers and based on that age profile derive the expected number of teachers that will retire each 10-year period from 2015 forward.

[8] This section draws on Jaimovich (2012), World Bank (2014a, b), Nordstrum (2015), Cross, Molina, Scanlon, and Wilichowski (2017), information provided by the Ministry of Public Services in Senegal, and findings from the data.

On a de facto basis, however, teacher training systems in these countries fall short of international best practice (Bruns and de Luque 2014). First, standards for entry into teacher training are low, as compared to high-performing education systems around the world. Second, teacher training programs tend to be of low quality, delivered by former teachers rather than trained instructors, and ill-suited to the needs of the candidates, who, having gone through their country's primary and secondary education system, often arrive poorly prepared, and are then confronted by curricula that focus on teaching methods and pedagogy theory rather than content knowledge.[9] In addition, while research suggests that pre-service training that focuses on the work teachers face in classrooms produces more effective teachers and higher learning for students (Boyd et al. 2009), little time is devoted to actual classroom practice, which can be as low as six weeks in Kenya, for example. Scheduled teaching time can also be low, both because programs are de facto condensed into a few months (as is the case in Senegal), and because absenteeism among teacher trainers is anecdotally high.

In short, it is easy to see how a vicious circle is created in which today's teachers have gone through an education system that does not prepare them adequately, through a training system with low entry requirements that does not compensate for the flaws in the education system, or through no training at all, to be sent into school where they struggle to teach the next generation of students. While we find a positive relationship between a teacher's education and training and their subject and pedagogy knowledge and classroom skills, even teachers with the highest education levels achieve significantly less than full marks.

Despite these shortcomings, teaching remains an attractive profession in most countries in sub-Saharan Africa. There is typically a surplus of applicants both for teacher training and to fill new teaching slots. For example, in Kenya, the Diploma Teacher Colleges admit 300 out of 8,000 candidates in a year, in Uganda, the acceptance rates into teacher training are 71 percent and in Nigeria they range from 50–90 percent, suggesting that the sector is at least somewhat competitive. Official criteria used to determine who gets hired among the applicants include time since graduation, degree, and sometimes grades received during teacher training. In practice, however, deviations from the official rule appear to be relatively common. For example, one-third of the 18,000 new teaching posts in Kenya in 2010 were misallocated, in the sense that district education officers deviated from the official algorithm to favor certain applicants (Barton, Bold, and Sandefur 2017).

In our sample, the large majority of teachers are employed on permanent and pensionable civil service contracts. These teachers are relatively well-paid. As a ratio

[9] For example, the Nigerian teacher training curriculum devotes more than twice the amount of time to pedagogy (theory) than to mathematics, English, and science—and even the time spent on subjects is mostly devoted to subject-specific learning methods. In the case of Kenya, all qualified teachers are expected to teach mathematics at primary level, but mathematics is not a compulsory subject during their training.

of GDP per capita, for example, teachers in sub-Saharan Africa earn on average more than four times as much as their counterparts in high-income countries (OECD 2011; UNESCO Institute for Statistics 2011). However, there is large variation in remuneration of teachers across Africa. The average monthly teacher salary in 2010 in Senegal was $380 (in current dollars), equivalent to 4.5 times GDP/capita, while the average teacher salary in Tanzania was $115, or twice GDP/capita, in the same year. There is also evidence suggesting that teachers are well-paid relative to other workers with similar educational background. Barton, Bold, and Sandefur (2017), for example, find, exploiting the Kenyan government's algorithm for hiring new teachers in 2010 in a regression discontinuity design, a civil service wage premium of over 100 percent.

Hence, it would appear that the current system of employment and remuneration confers substantial benefits to teachers, but—based on our findings—without ensuring that quality teaching is delivered. There are effectively no systems in sub-Saharan Africa that tie salaries and promotions to the performance of teachers. Consistent with this, we find that salary is most strongly predicted by experience and age, characteristics that, in turn, have little systematic relationship with teacher quality.

More recently, some attempts have been made to redress the balance and adapt the system, especially as new teachers are hired. Overall, 19 percent of teachers in sub-Saharan Africa are now employed on some form of nonpermanent contract. In countries where contract teachers are prevalent (four out of seven in our sample), almost one-third of teachers are employed on short-term contracts and this share swells to 50 percent for teachers with less than ten years of experience. This shift reflects both an age and a cohort effect, as many contract teachers graduate to civil service status over time.

Contract teachers tend to have less education and lower training than regular teachers and tend to earn substantially less, though with wide variation across the continent. There are also differences in the institutional setting of contract teachers in the countries we surveyed. In West Africa, the contract teacher program is primarily used as a way to lower costs, although contract teachers still tend to be relatively well paid—about $250 a month, which as a reference is the average regular teacher salary in Kenya. In essence, contract teachers here are effectively junior teachers employed by the government waiting, or hoping, for full civil service status. In East Africa, at least within Kenya, contract teachers originate from a system where parents clubbed together to pay for extra teachers at the school level. Contract teachers in Kenya earn on average $40 per month, and since their employment is outside the civil service system, their tenure is subject to parental approval, at least in principle.

Despite having less training and experience, we do not find any systematic differences in teacher knowledge or classroom skills between regular and contract teachers across the sample of teachers surveyed here. When it comes to absence, contract teachers are—if anything—absent less often (though this is not true in all countries), with significant differences emerging in both Kenya and Senegal.

Taken together, the system employed to train, hire, and motivate teachers falls short in several dimensions. But with a major increase and turnover in teachers—on average 130,000 new teachers are anticipated to be hired each year in the next 15 years in sub-Saharan Africa—a focus on how to ensure that the next cohort of teachers is better-prepared to teach well, and rewarded for doing so when deployed, can potentially go a long way to improve outcomes.

Discussion

The main finding of this paper is that teachers in sub-Saharan Africa perform poorly in several, likely complementary, dimensions. They teach too little, and they lack the necessary skills and knowledge to teach effectively when they do teach. If "adequate" teaching is characterized as being taught by teachers with at least basic pedagogical knowledge and minimum subject knowledge in language and mathematics for the full scheduled teaching day, then essentially no public primary schools in these countries offer adequate quality education.

In Bold, Filmer, Molina, and Svensson (2017b), we show that these shortcomings, and especially poor teacher knowledge, can account for a large share of the dramatic loss in human capital of students we observe already after four years of school, with the majority of fourth graders failing to master tasks covered in the second-year curriculum and more than one-quarter of such students deemed to have knowledge equivalent to a first grader, or below.

Given the results presented here, it is easy at a general level to list what governments "should" do to improve service performance in the education sector. For example, teacher training programs should seek to attract talented candidates and prepare them to teach the curriculum effectively. After teachers are hired, the need is for effective incentive schemes that ensure high effort and continued upgrading of knowledge and skills.

But it is an unfortunate reality that reforms aimed at systematically raising the quality of the teaching body along these lines should be viewed as more of a longer-run solution. For example, the huge improvement in the delivery of high-quality education in countries such as South Korea and Singapore resulted from systemwide efforts over several decades (Murnane and Ganimian 2014). Millions of children in low-income countries, who lack even basic literacy and numeracy skills even after several years of schooling, cannot afford to wait for systemwide reforms to be identified and implemented. Therefore, while planning for longer-term solutions, it is also important to consider shorter-term improvements.

There are now hundreds of experimental studies about different methods of raising student achievement in low-income countries, many of them from the very countries we surveyed here, looking at a wide range of possible interventions. Table 5 summarizes findings from several recent literature reviews relevant to improving the quality of teaching on the subject, which strike some common themes. For example, one step might focus on complementary resources involved

Table 5
Four Literature Reviews on the Promise of Teacher Incentives in Low- and Middle-Income Countries

Studies	*Sample*	*Findings*
Kremer, Brannen, and Glennerster (2013)	"30 primary school programs in raising test scores subject to randomized evaluation where study authors have made detailed cost information available"	"However, among those in school, test scores are remarkably low and unresponsive to more-of-the-same inputs, such as hiring additional teachers, buying more textbooks, or providing flexible grants. In contrast, pedagogical reforms that match teaching to students' learning levels are highly cost effective at increasing learning, as are reforms that improve accountability and incentives, such as local hiring of teachers on short-term contracts. Technology could potentially improve pedagogy and accountability."
Murnane and Ganimian (2014)	"115 studies in 33 low- and middle-income countries ... based on plausible identification strategies"	"Finally, well-designed incentives increase teacher effort and student achievement from very low levels, but low-skilled teachers need specific guidance to reach minimally acceptable levels of instruction."
Glewwe and Muralidharan (2015)	"118 high quality studies conducted from 1990 to 2014"	"Interventions that focus on improved pedagogy (especially supplemental instruction to children lagging behind grade level competencies) are particularly effective, and so are interventions that improve school governance and teacher accountability."
Evans and Popova (2016)	"six reviews of studies seeking to improve student learning in primary schools in developing countries … 227 of those studies report learning outcomes"	"Pedagogical interventions that match teaching to students' learning ... Individualized, repeated teacher training, associated with a specific method or task ... accountability-boosting interventions. These include two intervention subcategories: (i) teacher performance incentives and (ii) contract teachers."

in classroom teaching, such as teacher guides and lesson plans, which were available in two-thirds of classrooms. Our survey finds that while most students have pencils and notebooks and 80 percent of teachers have a functioning board to write on, this equipment is in place simultaneously in half the classrooms. One in ten classrooms are deemed too dark for students to read without straining their eyes and, on average, two to three students must share each textbook. However, there is by now a clear consensus that student learning, even in settings with limited resources, is remarkably unresponsive to just providing more of the same inputs.

There is stronger evidence, some of it reviewed in Banerjee and Duflo (2006, in this journal), that teacher effort can be raised and that this can lead to substantial improvements in learning, especially in settings with very low student achievement and high teacher absenteeism. The strongest evidence comes from studies providing financial incentives tied either to attendance or student performance (Duflo, Hanna, and Ryan 2012; Muralidharan and Sundararaman 2011), or short-term contracts predicated on the operation of dynamic incentives like contract teacher programs (Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur 2013; Duflo, Dupas, and Kremer 2015). But the experimental evidence also highlights barriers to the implementation of incentive systems, especially in the public sector, due to bureaucratic or political constraints. An important question going forward is therefore to identify ways to make these types of program effective within the government system.

Unfortunately, there are few, if any, well-identified studies on how to effectively improve teacher knowledge and skills and the impact thereof. This evidence gap is important to address, and the continued rapid expansion of new teachers ought to provide ample opportunities to do so. There is some related evidence. For example, a growing number of studies have shown that providing detailed guidance on what teachers should teach and how they should teach it—for instance by reorganizing instruction based on children's actual learning levels—can result in large gains in learning outcomes, especially for low-performing students. We are now also seeing the start of studies, such as Banerjee at al. (2016), that take the insights from individual studies and scale them up for broader application. Automated teaching, through computer-aided learning programs or scripted lesson plans, may also be a promising approach, especially when it comes to basic skills and lower-order skills, areas which are undoubtedly in need of improvement. Scripted lessons, however, may not work as well in improving the more complex aspects of teaching that are important for higher-order learning and with which teachers especially struggle: assessing students and responding through that assessment, asking thought-provoking questions to further understanding and knowledge, and giving appropriate feedback.

Dramatic improvements in teaching are hard. But the kinds of changes that would be useful, both for short-run improvements and longer-run systemic reforms, are becoming reasonably clear.

■ *We are grateful to the many researchers, survey experts, and enumerators who have worked on the country surveys that made this cross-country paper possible. We would like to thank Luis Benveniste, David Evans, and Ritva Reinikka for extensive comments on earlier versions of this paper. We would like to thank Serena Cocciolo for excellent research assistance. We are especially grateful to the World Bank, and in particular the Service Delivery Indicators Trust Fund (funded in large part by the Hewlett Foundation), for supporting the research. We also thank the Riksbankens Jubileumsfond (RJ) for financial support, as well as the editors of this journal for their many helpful suggestions. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.*

References

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131(3): 1415–53.

ASER. 2014. *Annual Status of Education Report (Rural) 2013.* New Delhi: ASER Centre.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." NBER Working Paper 22931.

Banerjee, Abhijit, and Esther Duflo. 2006. "Addressing Absence." *Journal of Economic Perspectives* 20(1): 117–132.

Barton, Nicholas, Tessa Bold, and Justin Sandefur. 2017. "Measuring Rents from Public Employment: Regression Discontinuity Evidence from Kenya." CEPR Discussion Paper DP12105

Baum, Donald R., Laura Lewis, Oni Lusk-Stover, and Harry A. Patrinos. 2014. "What Matters Most for Engaging the Private Sector in Education: A Framework Paper." SABER Working Paper Series 8.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian Stacy, Jakob Svensson, and Waly Wane. 2017a. "What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa." World Bank Policy Research Working Paper 7956.

Bold, Tessa, Deon Filmer, Ezequiel Molina, and Jakob Svensson. 2017. "The Lost Human Capital: Teacher Knowledge and Student Achievement in Africa." Unpublished paper.

Bold, Tessa, Bernard Gauthier, Jakob Svensson, and Waly Wane. 2010. "Delivering Service Indicators in Education and Health in Africa: A Proposal." World Bank Policy Research Working Paper 5327.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2013. "Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education." Working Paper 321, Center for Global Development, March.

Bold, Tessa, Jakob Svensson, Bernard Gauthier, Ottar Maestad, and Waly Wane. 2011. *Service Delivery Indicators: Pilot in Education and Health Care in Africa.* Bergen, Norway: Chr. Michelsen Institute.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2009. "Teacher Preparation and Student Achievement." *Educational Evaluation and Policy Analysis* 31(4): 416–40.

Bruns, Barbara, Soledad De Gregorio, and Sandy Taut. 2016. "Measures of Effective Teaching in Developing Countries." RISE Working Paper 16/009.

Bruns, Barbara, and Javier Luque. 2014. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington, DC: World Bank.

Bruns, Barbara, Alain Mingat, and Ramahatra Rakotomalala. 2003. *Achieving Universal Primary Education by 2015: A Chance for Every Child.* Washington, DC: World Bank.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20(1): 91–116.

Coe, Robert, Cesare Aloisi, Steve Higgins, and Lee Elliot Major. 2014. *What Makes Great Teaching? Review of the Underpinning Research.* London: Sutton Trust.

Cross, Jessica, Ezequiel Molina, Cole Scanlon, and Tracy Wilichowski. 2017. "A Global Perspective on Teacher Policies: Insights from Twenty-Eight Education Systems." Unpublished paper.

Danielsson, Charlotte. 2007. *Enhancing Professional Practice: A Framework for Teaching,* 2nd edition. Association for Supervision and Curriculum Development (ASCD).

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools." *Journal of Public Economics* 123: 92–110.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102(4): 1241–78.

Evans, David K., and Anna Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Research Observer* 31(2): 242–70.

Glewwe, Paul, and Karthik Muralidharan. 2015. *Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications.* RISE Working Paper 15/001.

Hungi, Njora, Demus Makuwa, Kenneth N. Ross, Mioko Saito, Stéphanie Dolata, Frank van Cappelle, Laura Paviot, and Jocelyne Vellien. 2010. "SACMEQ III Project Results: Pupil Achievement Levels in Reading and Mathematics." SACMEQ III Working Document Number 1.

Jaimovich, Analia. 2012. *SABER Teacher Country Report: Uganda 2012.* World Bank.

Johnson, David. 2006. "Investing in Teacher Effectiveness to Improve Educational Quality in Developing Countries: Does In-Service Education for Primary Mathematics Teachers in Sri Lanka Make a Difference to Teaching and Learning?" *Research in Comparative and International Education* 1(1): 73–87.

Johnson, David, Andrew Cunningham, and Rachel Dowling. 2012. "Teaching Standards and Curriculum Review." Unpublished paper.

Kane, Thomas J., and Douglas O. Staiger. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Seattle: Bill and Melinda Gates Foundation.

Ko, James, and Pamela Sammons. 2013. *Effective Teaching: A Review of Research and Evidence*. Reading, U.K.: CfBT Education Trust.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340(6130): 297–300.

Kremer, Michael, Nazmul Chaudhury, F. Halsey Rogers, Karthik Muralidharan, and Jeffrey Hammer. 2005. "Teacher Absence in India: A Snapshot." *Journal of the European Economic Association* 3(2–3): 658–67.

Makuwa, Demus. 2011. "Characteristics of Grade 6 Teachers." SACMEQ III Working Paper 2.

Muijs, Daniel, Leonidas Kyriakides, Greetje van der Werf, Bert Creemers, Helen Timperley, and Lorna Earl. 2014. "State of the Art—Teacher Effectiveness and Professional Learning." *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice* 25(2): 231–56.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119(1): 39–77.

Murnane, Richard J., and Alejandro J. Ganimian. 2014. "Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations." NBER Working Paper 20284.

Nordstrum, Lee E. 2015. *Effective Teaching and Education Policy in Sub-Saharan Africa: A Conceptual Study of Effective Teaching and Review of Educational Policies in 11 Sub-Saharan African Countries*. Washington, DC: United States Agency for International Development.

OECD. 2011. *Education at a Glance 2011: OECD Indicators*. Paris: OECD Publishing.

OECD. 2015. *Education at a Glance 2015: OECD Indicators*. Paris: OECD Publishing.

PASEC. 2015. *PASEC 2014: Education System Performance in Francophone Sub-Saharan Africa*. Dakar, Senegal: PASEC.

Pianta, Robert, Karen La Paro, and Bridget Hamre. 2007. *Classroom Assessment Scoring System (CLASS): Manual*. Baltimore: Paul H. Brookes Publishing Co.

Stallings, Jane. 1980. "Allocated Academic Learning Time Revisited, or Beyond Time on Task." *Educational Researcher* 9(11): 11–16.

Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten. 2010. "Using Student Performance Data to Identify Effective Classroom Practices." *American Economic Review* 100(2): 256–60.

UNESCO. 2013. *The Global Learning Crisis: Why Every Child Deserves a Quality Education*. Paris: UNESCO.

UNESCO Institute for Statistics. 2011. *Financing Education in Sub-Saharan Africa: Meeting the Challenges of Expansion, Equity and Quality*. Montreal: UNESCO Institute for Statistics.

Vieluf, Svenja, David Kaplan, Eckhard Klieme, and Sonja Bayer. 2012. *Teaching Practices and Pedagogical Innovation: Evidence from TALIS*. Paris: OECD Publishing.

World Bank. 2014a. *SABER Teachers Kenya Country Report 2014*. Washington, DC: World Bank.

World Bank. 2014b. *SABER Teachers Mozambique Country Report 2014*. Washington, DC: World Bank.

You, Danzhen, Lucia Hug, and David Anthony. 2014. *Generation 2030/Africa*. New York: UNICEF.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 205–228

Population Control Policies and Fertility Convergence

Tiloka de Silva and Silvana Tenreyro

In the middle of the twentieth century, almost all developing countries experienced a significant increase in life expectancy, which, together with high fertility rates, led to rapid population growth rates. The fear of a population explosion lent impetus to what effectively became a global population-control program. The initiative, propelled in its beginnings by intellectual elites in the United States, Sweden, and some developing countries, most notably India, mobilized international private foundations as well as national governmental and nongovernmental organizations to advocate and enact policies aimed at reducing fertility. By 1976, following the preparation of the World Population Plan of Action at the World Population Conference in Bucharest in 1974, 40 countries, accounting for 58 percent of the world's population and virtually all of the larger developing countries, had explicit policies to reduce fertility rates. Between 1976 and 2013, the number of countries with direct government support for family planning rose to 160. In this essay, we will argue that concerted population control policies implemented in developing countries are likely to have played a central role in the global decline in fertility rates in recent decades and can explain some patterns of that fertility decline that are not well accounted for by other socioeconomic factors.

■ *Tiloka de Silva is a Teaching Fellow in Economics and Silvana Tenreyro is a Professor of Economics, both at the London School of Economics (LSE), London, United Kingdom. Tenreyro is also a Program Leader at the Centre for Macroeconomics (at LSE) and Research Fellow at the Center for Economic Policy Research, London, United Kingdom. Their email addresses are t.s.de-silva@lse.ac.uk and s.tenreyro@lse.ac.uk.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.205 doi=10.1257/jep.31.4.205

To set the stage, we begin by reviewing some trends and patterns in the fertility decline in the last half-century across countries and regions. We argue that although socioeconomic factors do play an important role in the worldwide fertility decline, they are far from sufficient to account for the timing and speed of the decline over the past four decades. For example, the cross-country data in any given year show a negative correlation between per capita income and fertility rates. However, that relationship has shifted downward considerably over time: today the typical woman has, on average, two fewer children than the typical woman living in a country at a similar level of development in 1960.

We then discuss the evolution of global population-control policies in more detail. All population-control programs involved two main elements: promoting an increase in information about and availability of contraceptive methods, and creating public campaigns aimed at establishing a new small-family norm. The evidence suggests that these public campaigns appeared to have been critical in complementing contraceptive provision. While estimating the causal effect of these programs is challenging, we examine the relationship between different measures of family planning program intensity and the declines in fertility over the past decades and find a strong association, after controlling for other potential explanatory variables, such as GDP, schooling, urbanization, and mortality rates.

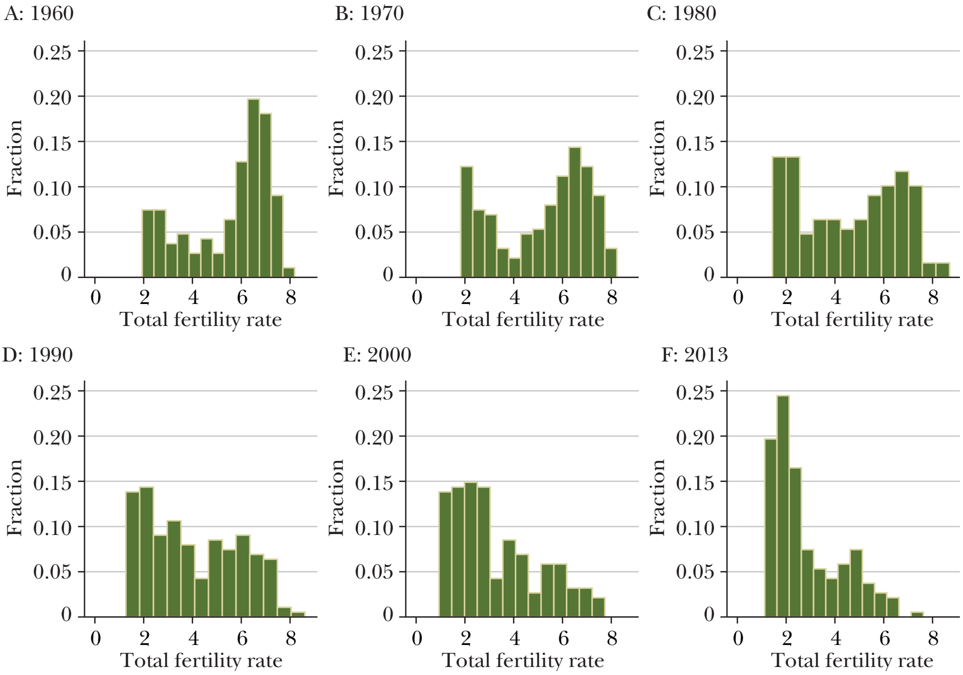
In a final section, we discuss in more detail the role played by these other variables in the decline in fertility and highlight that the drop in fertility rates seems to be occurring and converging across countries with varying levels of urbanization, education, infant mortality, and so on. We conclude that population control policies seem to be the factor that best accounts for this commonality.

Fertility Patterns across Time and Space

The world's total fertility rate declined from over 5.0 children per woman in 1960 to 2.5 children per woman in 2013.[1] This trend is not driven by just a few countries. Figure 1 plots fertility rate histograms for the start of decades since 1960; the bars show the fraction of countries for each fertility interval. (The figure shows 2013 rather than 2010 to report the most recent information.) In 1960, more than half the countries in the world had a fertility rate between 6 and 8, and the median fertility rate was 6.2 children per woman. (When weighted by population, the world's median is 5.8.) In 2013, the largest mass of countries is concentrated around 2, with the median total fertility rate being 2.2.

[1] The total fertility rate is defined as the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates. In this paper, we will use "total fertility rate" interchangeably with "fertility" and "fertility rate."

Figure 1
Fertility Histograms over Time

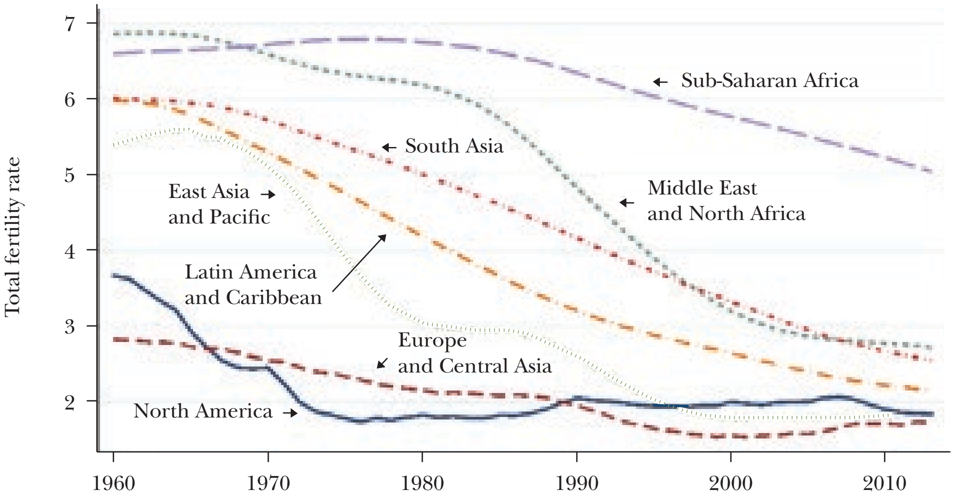


Source: The data comes from the World Bank's World Development Indicators database.
Note: The figure shows fertility histograms at the beginning of each decade. In the final histogram, the year 2013 is used rather than 2010 to report the latest available information.

These large declines in fertility took place in most regions of the world, as shown in Figure 2. Between 1960 and 2013, fertility rates fell from 5.4 to 1.81 in East Asia and the Pacific, from 5.98 to 2.16 in Latin America and the Caribbean, from 6.87 to 2.83 in the Middle East and North Africa, and from 6.02 to 2.56 in South Asia. The fertility decline in sub-Saharan Africa has been slower, but still sizable: since the 1980s, the total fertility rate in this region fell from 6.7 to 5. Within this region, South Africa has already reached a total fertility rate of 2.4 and Mauritius is now at a fertility rate of 1.44. While absolute declines in fertility were not as large in North America or Europe and Central Asia, the percentage declines in both regions have been significant—nearly 50 percent in North America and close to 40 percent in Europe and Central Asia. Interestingly, the fertility rate for North America bottomed out in the 1980s, and in Europe and Central Asia, it bottomed out in the 1990s.

A number of empirical studies have documented a negative relationship between fertility rates and income. While this relationship is indeed negative in the cross-section of countries, the relationship has changed over time, shifting

Figure 2
Fertility Trends across Regions

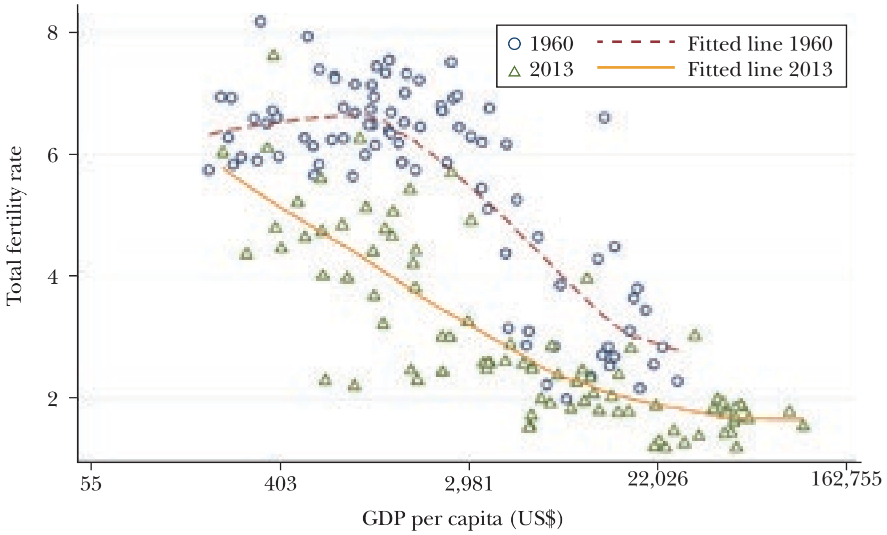


Note: This figure plots the trends in fertility by region, as defined by the World Bank, between 1960 and 2013. The data comes from the World Development Indicators database.

downward and becoming flatter over time. Figure 3 shows the relationship between the total fertility rate and real GDP per capita in 1960 and in 2013. The figure also shows fitted lines for these two years. The downward shift has been, on average, around 2 children per woman, meaning that today a woman has 2 fewer children than a woman living in a country at the same level of development in 1960, which is close in magnitude to the drop in overall world fertility of 2.5 children per woman. The cross-section relationship between fertility and income observed in 1960 would predict a total fertility rate of around 4 at the average per capita GDP for 2013 (recall the actual rate is 2.5).

As Figure 3 illustrates, the issue is not just to explain a decline in global fertility. It is also necessary to explain why the fall in fertility rates witnessed by developing countries in recent decades was so very rapid, compared with the rather slow and secular decline in fertility rates experienced by more mature economies. For example, the fertility decline began as early as the mid-1700s in some European countries and only reached replacement levels in the early twentieth century (Coale 1969). Furthermore, it is necessary to explain why countries with markedly different levels of income, urbanization, education, and other factors are all converging to very similar fertility rates. As we discuss in the next section, the worldwide spread of population-control programs can help to explain these patterns in the fertility data.

Figure 3
Fertility–Income Relation in 1960 and 2013



Source: Authors using data from the World Development Indicators database.
Note: For a sample of 88 countries, the figure shows the scatterplots and fitted line (that is, the lowess smoothed relationship or locally weighted smoothing function) between the total fertility rate and log of per capita GDP (in constant 2005 US$) in 1960 and 2013. The x-axis is log scale.

The Global Family Planning Movement and its Consequences

Global Evolution of Family Planning Programs

After World War II, there was growing concern with the unprecedented levels of population growth.[2] A population-control movement developed, led by, among others, John D. Rockefeller III, whose main preoccupations were the growing imbalance between population and resource growth, and the potential for political instability given that most of the population growth was concentrated in the poorest countries of the world. In 1952, Rockefeller founded the Population Council, aimed at providing research and technical assistance for population programs across the world. That same year, India started the first national population program, and in parallel, the International Planned Parenthood Federation was established.[3] By the late 1950s, the "population question" was receiving the attention of the

[2] This section draws heavily on Robinson and Ross (2007), who provide a compilation of case studies of family planning programs in 22 countries across the world.
[3] The earlier birth-control movement led by Margaret Sanger in the United States (who set up the first birth-control clinic in the United States in 1916) and Elise Ottesen-Jensen in Sweden was another force leading to efforts for fertility reduction.

US government. A report by a Presidential Committee studying the United States Military Assistance Program (Draper 1959) devoted an entire chapter to the issue, ending with a recommendation that the government "assist those countries with which it is cooperating in economic aid programs, on request, in the formulation of their plans designed to deal with the problem of rapid population growth."[4] By this time, private foundations including the Rockefeller and Ford Foundations were providing seed funding for research and planning programs, but it was in the mid-1960s that large-scale funding became available and the population planning movement really took off.

The first large-scale intervention was carried out by the Swedish government, which supported family planning efforts in Sri Lanka (then Ceylon), India, and Pakistan, starting in 1962 (Sinding 2007). Over time, several international organizations, like USAID and the World Bank, joined in providing funds and support for family planning programs around the world. The invention of the modern intrauterine device (IUD) and the oral contraceptive pill around the same time allowed for the possibility of easy-to-use and effective contraceptive methods becoming widely available for public use.

These early family planning efforts showed rapid effects in East Asian countries, including Hong Kong, South Korea, Singapore, and Thailand. Program implementation and success would take longer in other developing countries, partly due to the difficulty of overcoming cultural inhibitions and religious opposition towards birth control, as well as operational problems including inadequate transport infrastructure and insufficient funding. The World Population Conference in 1974 appeared to be a turning point for the global family planning movement. Tables 1 and 2 show how countries around the world have been categorized by their fertility goals and the type of government support for family planning for selected years from 1976 to 2013, according to the UN World Population Policy database.

In 1976, for example, the 40 countries that had explicit policies to limit fertility covered nearly one-third of East Asian countries, a quarter of Latin American and Caribbean countries, and nearly two-thirds of South Asian countries. By contrast, only one-fifth of countries in North Africa, the Middle East, and Sub-Saharan Africa had a fertility reduction policy in 1976. By 1996, 82 countries had a fertility reduction policy in place (by this time, some countries had reached their fertility reduction targets and changed to policies of maintaining fertility rates), including half of the countries in East Asia and Latin America, and more than two-thirds of the countries in Sub-Saharan Africa and South Asia. These countries represent 70 percent of the world's population. In 1976, 95 governments were providing direct support for family planning. (Support for family planning was not always associated with an explicitly stated goal of reducing fertility.) The number of countries with state support for family planning has continued to rise steadily.

[4] For more references that trace the origins of the population control movement primarily to the West, see online Appendix C, available with this paper at http://e-jep.org.

Table 1

Number of Countries with Government Goals for Fertility Policy

Year	*Lower fertility*	*Maintain fertility*	*No intervention*	*Raise fertility*	*Number of Observations*
1976	40	19	78	13	150
1986	54	16	75	19	164
1996	82	19	65	27	193
2005	78	31	47	38	194
2013	84	33	26	54	197

Source: The data is obtained from the UN World Population Policies database.
Note: The table shows the number of countries by type of policy adopted towards fertility. The data begins in 1976. Countries are categorized according to whether they had a policy to lower, maintain, or raise fertility or if they had no intervention to change fertility.

Table 2

Number of Countries by Government Support for Family Planning

Year	*Direct support*	*Indirect support*	*No support*	*Limit/Not permitted*	*Number of Observations*
1976	95	17	28	10	150
1986	117	22	18	7	164
1996	143	18	26	2	193
2005	143	35	15	1	194
2013	160	20	16	1	197

Source: The data is obtained from the UN World Population Policies database.
Note: The table shows the number of countries by the type of support extended by the state for family planning services. The data begins in 1976. Countries are categorized by whether their governments directly supported, indirectly supported, or did not support family planning as well as if the government limited family planning services or did not permit family planning in the country.

Features of Family Planning Programs

The early phases of family planning programs in most developing countries typically sought to provide a range of contraception methods—some combination of oral contraceptives, IUD, condoms, sterilization, and abortion—and information on their use. However, increases in the supply of contraceptives proved insufficient to lower fertility rates to desired levels, particularly in poorer or more traditional societies. This failure led to concerted efforts to change public attitudes and beliefs and establish a new small-family norm through active mass-media campaigns. We discuss these two phases in turn.

The implementation of the family planning programs varied vastly across countries. Differences included the role of public and private provision; the price at which contraception was offered; subsidies to production or sales; the delivery system through which services were provided; the outlets for the mass-media campaigns; and the various supplementary policies that accompanied the core measures (Freedman and Berelson 1976).[5]

Most countries began their family planning programs with a clinic-based approach that took advantage of the existing health infrastructure to provide modern contraceptive methods. Many countries also implemented programs in hospitals to advise women on the use of contraception, often after women had given birth or undergone an abortion. However, this approach had limited success in countries where a large proportion of women gave birth outside of the formal health care system, like India and Iran. Thus, the policy was supplemented by the deployment of trained field workers who made house calls, particularly in rural areas. In some nations, such as Iran and Malaysia, family-planning programs were linked to maternal and child health services at an early stage, which allowed for better integration of the program into the country's health system. Towards the 1990s, with the rebranding of family planning as sexual and reproductive wellbeing, more countries have followed this approach.

Many of the family planning programs established in the 1950s and 1960s, which focused on increasing the supply of contraception, failed to gain much traction. For instance, highly traditional societies and countries with a predominantly Catholic or Muslim population had difficulty gaining wide acceptance for their family planning programs. It became clear that without changing the willingness to use contraceptives and, more importantly, reducing the desired number of children, merely improving access to birth control had limited impact. The importance of changing the desired number of children, in particular, was highlighted by leading demographers at the time such as Enke (1960) and Davis (1967), who argued that a desire to use contraceptives was perfectly compatible with high fertility. Countries thus began to present and to adapt their population-control policies to address these concerns.

For example, early in Indonesia's family planning program, the government published a pamphlet titled "Views of Religions on Family Planning," which documented the general acceptance of family planning by four of Indonesia's five official religions—Islam, Hinduism, and Protestant and Catholic Christianity (Hull 2007). To overcome fears that husbands would resist male doctors or health professionals working with their wives, the family planning program in Bangladesh relied heavily on female health workers visiting women in their homes to educate them about and supply them with contraceptive methods. This modality also ensured a greater diffusion of contraceptive knowledge and methods in rural Bangladesh (Schuler, Hashemi, and Jenkins 1995).

Mass communication was commonly used to shape attitudes toward family planning, often with the aim of changing public views by establishing a small-family norm. During the 1970s, slogans proliferated in different media outlets (television,

[5] For a more detailed summary of the key features of early family planning programs around the world, highlighting the countries that implemented each approach, see the online Appendix Table A1, available with this paper at http://e-jep.org.

radio, and magazines), street posters, brochures, and billboards, all conveying a similar message regarding the benefits of small families. In India, the family planning program's slogan, "Have only two or three children, that's enough," was widely publicized on billboards and the sides of buildings. Other slogans in India were "A small family is a happy family" and "Big family: problems all the way; small family: happiness all the way" (Khanna 2009). Bangladesh publicized the slogans "Boy or girl, two children are enough" and "One child is ideal, two children are enough" (Begum 1993). South Korea ran the slogan "Stop at two, regardless of sex" (Kim and Ross 2007); Hong Kong chose "Two is enough" (Fan 2007), and so on. China took population planning to the extreme in 1979, when it imposed a coercive one-child policy, but the Chinese fertility rate actually started falling significantly in the early 1970s, before the one-child policy was implemented (Zhang 2017). The strong population-control policy enacted in China in 1973 was characterized by mass-media messages such as "Later, longer, fewer" (Tien 1980) and "One is not too few, two, just right, and three, too many" (Liang and Lee 2006). In Singapore, bumper stickers, coasters, calendars, and key chains reinforcing the family planning message were distributed free of charge. In Bangladesh, television aired a drama highlighting the value of family planning (Piotrow and Kincaid 2001). The Indonesian program became particularly noteworthy in its collaboration between the government and community groups in getting the messages of the program across.

In Latin America, the Population Media Centre (a nonprofit organization) collaborates with a social marketing organization in Brazil to ensure the inclusion of social and health themes in soap operas airing on TV Globo, the most popular television network in Brazil. (TV Globo's programming is estimated to currently reach 98 percent of Brazil's population, and 65 percent of all of Spanish-speaking Latin America.) The Population Media Centre studied how programs like "Paginas da Vida" ("Pages of Life") influenced Brazilians: about two-thirds of women interviewed said the telenovela "Paginas da Vida" had helped them take steps to prevent unwanted pregnancy. Brazil's telenovelas have been popular across Latin America since the 1980s; they almost invariably depict the lives of characters from small families who were also very rich and glamorous (Population Media Centre 2016). In Brazil, the main force behind the anti-natalist movement was BEMFAM, an affiliate of the International Planned Parenthood Federation. The military regime of the 1970s and the Catholic Church hierarchy were opposed to birth control, though the local clergy and multiple nongovernmental organizations provided advice and information in favor of contraceptive use. In other Latin American countries, such as Colombia and Chile, family planning had strong support from the government.

Stronger inducements such as monetary or in-kind incentives and disincentives were also used in some countries as means of encouraging families to practice birth control. In Tunisia, for example, government family allowances were limited to the first four children; in Singapore, income tax relief was restricted to the first three children as was maternity leave, the allocation of public apartments, and preferred school places. Incentives for female or male sterilization was a common feature of family planning programs in India, Bangladesh, and Sri Lanka and resulted in a large number of sterilizations taking place during the 1970s. In Bangladesh, field health workers were paid for accompanying an individual to a sterilization procedure, while in Sri Lanka and India both the sterilization provider and patient were given compensation. In Kerala, India, individuals undergoing sterilization received payments in cash and food roughly equivalent to a month's income for a typical person. This type of incentivized compensation scheme, combined with increased regional sterilization targets, led to a drastic increase in sterilization procedures. Critics alleged that many acceptors were coerced by officials who stood to gain from higher numbers, both in monetary and political terms.

In addition to increased provision of information on and access to family planning methods, attempts were made to delay marriage and childbearing or to increase birth spacing as a means of controlling fertility. For example, the legal age of marriage was increased to 18 years for women and 21 years for men in India, and to 17 years for women and 20 years for men in Tunisia. China raised the legal age for marriage in urban areas (to 25 years for women and 28 years for men) and rural areas (23 years for women and 25 years for men). China also imposed a minimum gap of three to four years between births and restricted the number of children to three per couple until it decided to implement the draconian one-child policy in 1979.

More recently, given the sizeable decline in birth rates that has already occurred, fertility control has been put on the back burner. In fact, the current HIV/AIDS epidemic has somewhat overshadowed fertility control, particularly in African countries (Robinson and Ross 2007), while family planning did not even warrant being a sub-goal in the Millennium Development Goals agreed to in 2000. Many countries are now below replacement-level fertility rates. Nonetheless, family planning programs seem to have been incorporated into the broader framework of sexual and reproductive health services and become firmly entrenched in health care systems around the world.

The details of fertility programs differed across countries. But from a broader view, the prevalence and growth of these programs is remarkable. Fertility reduction programs took place under both democratic and autocratic regimes, whether oriented to the political left or right (for example, Chile under both Allende and Pinochet), and in Buddhist, Christian, and Muslim countries alike. In some countries, like Brazil, family planning programs were initiated and almost exclusively run by nonprofit, nongovernmental organizations, while in others, like Singapore or India, the government was fully involved.

A natural question is whether the type of less-coercive intervention carried out by most countries can be effective in helping to rapidly change norms and in overcoming other socioeconomic influences that affect fertility rates. In the context of China, Zhang (in this journal, 2017) observes that the one-child policy can explain only a small change in fertility given that a robust family planning program was already in operation since the early 1970s. He argues that strong family planning programs, such as those observed in most East Asian countries during the 1960s and early 1970s, would be as effective in lowering fertility. In addition, recent experimental (or quasi-experimental) studies also suggest the effectiveness of public persuasion measures in reducing fertility. La Ferrara, Chong, and Duryea (2012) find that Brazilian regions covered by a television network showing soap operas that portray small families experienced a bigger reduction in fertility rates. In Uganda, Bandiera et al. (2014) find that adolescent girls who received information on sex, reproduction, and marriage reported wanting a smaller number of children. Evidence of family planning programs in the United States appears more mixed, though recently, Bailey (2013) has shown that a targeted US family planning program significantly reduced fertility. In the next section, we explore the question using cross-country data on spending and implementation effort of the program, and their relationship with fertility reduction.

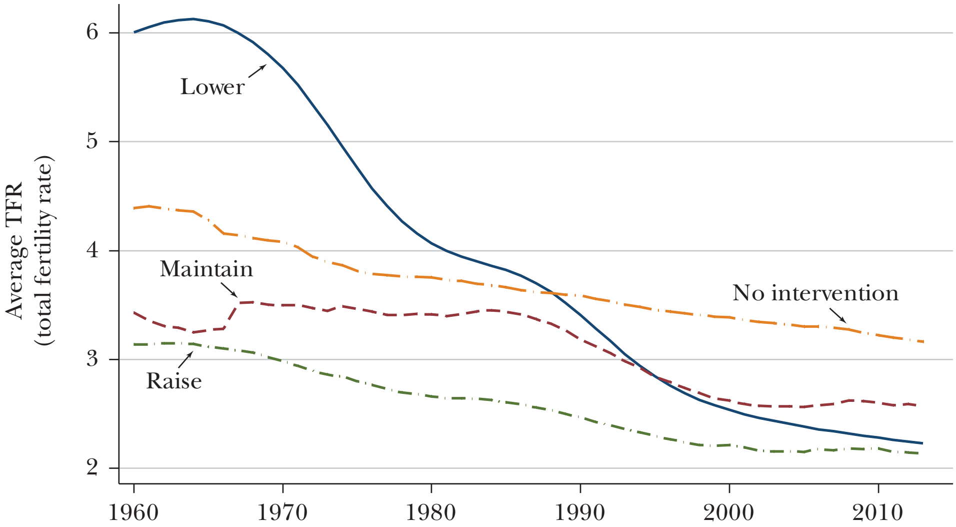
Fertility Policies and the Decline in Fertility Rates

In seeking to assess the quantitative effect of the fertility programs on the basis of cross-country data, there are clearly a number of covariates that could confound the estimation of a causal effect. The task is particularly difficult since different countries opted for a wide and varied range of fertility policies, with the specific choice of measures partly dictated by their feasibility in each country's institutional and cultural setting. Equally important, data availability is also limited. Thus, while estimating the causal effect of these programs is beyond the scope of this essay, our analysis illustrates some descriptive relationships between fertility rates, population policy, and different measures of family planning program intensity, conditioning on covariates of fertility traditionally used in the literature. Taken as a whole, this evidence is strongly consistent with the hypothesis that population control programs have played a major role in the fertility decline.

As a first exercise, we compare the country-level patterns in mean fertility rate by the fertility policy goals stated in 1976, which paints the striking picture shown in Figure 4. The data on fertility policy begins in 1976, but several countries had already adopted fertility reduction policies beforehand. While fertility has fallen in all regions, even in the group of predominantly European countries that wanted to increase fertility, the countries that had identified the need to reduce fertility in 1976 recorded by far the highest average fertility rates before 1976, but the second-lowest average fertility rates by 2013. The countries where there was no intervention had the second-highest average fertility rates in 1976 and became the highest fertility group by 2013.

For the analysis that follows, infant mortality rates, the proportion of urban population, and per capita GDP are obtained from the World Bank's World Development Indicators, while data on the years of schooling of the population aged 25+ are taken from Barro and Lee (2013). Data on the existence of a fertility policy and government support for family planning come from the UN World Population Policies Database. We use three measures of family planning program intensity: funds for family planning per capita; a family planning program effort score; and the percentage of women exposed to family planning messages through mass media. Data on funds for family planning are taken from Nortman and Hofstatter (1978), Nortman (1982), and Ross, Mauldin, and Miller (1993), which, taken together, cover

Figure 4
Evolution of Fertility Rates by Policy in 1976



Source: The data on fertility policy is obtained from the UN World Population Policies Database, and total fertility rates are from the World Bank's World Development Indicators.
Note: The figure illustrates the evolution of weighted average total fertility rate, with countries grouped by the fertility policy observed in 1976. The policy could be to lower, maintain, or raise fertility; there also could be no intervention.

funding for family planning by source for 58 countries over various years starting in 1972 and going up to 1992. Family planning program effort is measured using the Family Planning Program Effort Index published in Ross and Stover (2001). This indicator, based on work by Lapham and Mauldin (1984), measures the strength of a given country's program along four dimensions: policies, services, evaluation, and method access. The score has a potential range of 0–300 points, based on 1–10 points for each of 30 items, and has been calculated for 1972, 1982, 1989, 1994, and 1999, covering 95 countries. Finally, the Demographic and Health Surveys (DHS) from 57 countries in various years provide data on the percentage of women who have been exposed to family planning messages on the radio, television, or newspapers. These three measures altogether aim at capturing the intensity with which population programs were implemented.

As our next exercise to study the relation between population programs and fertility, we use data on funds for family planning. We look at the amount of funds (in real terms) available for family planning, from both government and nongovernment sources over the 1970s, 1980s, and 1990s for each country.

The patterns by region are as follows. Latin American countries appear to have the largest amount of funds for family planning per capita, with total funding exceeding US$2 per capita (in 2005 US dollars) in Costa Rica, El Salvador, and Puerto Rico. The region also has the highest proportion of nonstate funding for

Table 3

Change in Total Fertility Rates (TFRs) and Funding for Family Planning Programs

	Dependent variable is: Change in TFR			
	Absolute change		*% Change*	
	(1)	(2)	(3)	(4)
ln(average funds per capita)	−0.630*** [0.120]	−0.430** [0.181]	−10.47*** [1.487]	−4.974** [2.030]
Change in years of education of adults		−0.13 [0.133]		0.001 [0.002]
Change in urban population as % of total		−0.008 [0.009]		0.001 [0.003]
Change in ln(GDP per capita)		−0.426* [0.227]		−0.382** [0.158]
Change in infant mortality rate		0.006* [0.003]		0.668*** [0.131]
Observations	56	37	56	37
R^2	0.35	0.39	0.418	0.72

Source: Authors. Data on total fertility rate, urban population, per capita GDP, infant mortality rate, and US Consumer Price Index (used to convert the funds to real terms) are from the World Development Indicators. Data on years of schooling are from Barro and Lee (2013). Data on funds for family planning are from Nortman and Hofstatter (1978), Nortman (1982), and Ross, Mauldin, and Miller (1993).
Note: The table reports the results of regressions of the change in total fertility rate between 2013 and 1960 on the logged real value of average per capita funds for family planning for the 1970s, 1980s, and 1990s, controlling for the changes in years of schooling of the population aged 25+, urban population as a percentage of total population, log GDP per capita, and infant mortality rate between 2013 and 1960. Given the small number of observations for infant mortality rate and GDP per capita in 1960, we use the earliest available observation before 1965 to construct the change. All regressions include a constant. Per capita funds for family planning are converted to 2005 US$ before averaging. The values in parentheses are robust standard errors.
*, **, and *** indicate significance at 10, 5, and 1 percent levels, respectively.

family planning, more than double the state-funding in some countries. By contrast, in Asia, the funding available for family planning is predominantly state-led. As a percentage of GDP, total funds for family planning averaged at around 0.05 percent in the 1970s and 0.07 percent in the 1980s, but was as high as 0.47 percent in Bangladesh and 0.46 in Korea in the 1980s.[6]

Table 3 shows the results of a regression of the change in fertility on (logged) average family planning funds per capita over the 1970s, ’80s, and ’90s, with and without controlling for changes in the covariates of fertility traditionally used in the literature, such as GDP per capita, educational attainment, urbanization, and infant

[6] The full table with funds for family planning by country for the 1970s and 1980s is available in the online Appendix Table A2, available with this paper at http://e-jep.org.

mortality. (Each of these covariates will be discussed in more detail in the following section.) Columns 1 and 2 use absolute changes in all fertility (and the other covariates) between 1960 and 2013, and columns 3 and 4 use percentage changes in these variables over the same period.

Despite the small number of observations available once the controls are included, the negative relationship between changes in total fertility rate and funds for family planning remains significant, indicating that the countries with more funding for family planning experienced greater reductions in fertility rates, even after controlling for the changes in income, urbanization, infant mortality, and years of schooling of the adult population. (Controlling for years of schooling of adult women instead of adult population leads to similar results.) Quantitatively, the results indicate that a 1 percent increase in funding per capita is associated with a 5 percent reduction in the total fertility rate.

We do not include changes in female labor force participation rates in this regression because the cross-country data for this variable begins only in 1980. However, we replicate the exercise focusing on changes between 1980 and 2013 for all variables and find that the results hardly change, with no significant correlation between changes in female labor force participation and the fertility decline. We also carry out the exercise separately for government funding and private funding for family planning per capita, and find that government spending has a significant, positive correlation with the fertility decline whereas private spending does not appear to be significant (see the online Appendix for the full set of results).

Our third exercise uses the family planning program effort index published by Ross and Stover (2001) as an alternative measure of program inputs. The regional averages of the index indicate that East Asia and South Asia have, in general, had the strongest family planning programs over time. Latin America, North Africa, and the Middle East seem to have caught up on program effort over the three decades, but the greatest gain appears to have been in Sub-Saharan Africa, which was the latest to adopt family planning programs, in 1989–1999.[7] We use these data to examine the relationship between the observed change in fertility over the 1960–2013 period and the average program effort score over the 1970s, ’80s, and ’90s, again controlling for the other covariates of fertility. Table 4 indicates a strong negative relationship, with larger fertility declines in countries with higher program effort.

Next, we use the Demographic and Health Surveys (DHS) data on percentage of women exposed to family planning messages through mass media to carry out the same exercise as for family planning program funds and program effort score. Table 5 shows these results. The context of this analysis is slightly different from the two previous exercises because the data are based on DHS surveys which were carried out predominantly in sub-Saharan African countries (30 of the countries in the sample used in columns 1 and 3, and 15 of the countries in the sample used

[7] For more details on regional average program effort scores by year, see the online Appendix Table A5, available with this paper at http://e-jep.org.

Table 4
Change in Total Fertility Rates (TFRs) and Family Planning Program Effort

	Dependent variable is: Change in TFR			
	Absolute change		*% Change*	
	(1)	(2)	(3)	(4)
Average family planning program effort score	–0.039*** [0.007]	–0.041*** [0.014]	–0.716*** [0.101]	–0.500*** [0.166]
Change in years of education of adults		–0.124 [0.115]		0.003 [0.003]
Change in urban population as % of total		–0.012 [0.008]		–0.0001 [0.005]
Change in ln(GDP per capita)		0.015 [0.198]		–0.108 [0.192]
Change in infant mortality rate		0.002 [0.003]		0.549*** [0.142]
Observations	107	55	107	55
R^2	0.21	0.41	0.321	0.636

Source: Authors. Data on total fertility rate, urban population, per capita GDP, and infant mortality rate are from the World Development Indicators. Data on years of schooling are from Barro and Lee (2013). Data on family planning program effort are from Ross and Stover (2001).
Note: The table reports the results of regressions of the change in TFR between 2013 and 1960 on the average family planning program effort score over the 1970s, 1980s, and 1990s, controlling for the change in years of schooling of the population aged 25+, urban population as a percentage of total population, log GDP per capita, and infant mortality rate between 2013 and 1960. All regressions include a constant. Given the small number of observations for infant mortality rate and GDP per capita in 1960, we use the earliest available observation before 1965 to construct the change. All regressions include a constant. The values in parentheses are robust standard errors. *, **, and *** indicate significance at 10, 5, and 1 percent levels, respectively.

in columns 2 and 4) starting from the early 1990s. Therefore, these results capture more recent efforts in family planning as seen in sub-Saharan Africa. The regression results show a significant, negative association between the fertility change and exposure to family planning messages after controlling for other covariates. It therefore seems likely that the delay in the implementation of the family planning programs in sub-Saharan Africa explains the delayed decline in fertility in the region. Both in Table 4 and Table 5, the coefficients corresponding to the policy measure change little when adding the controls; this suggests that additional omitted variables are unlikely to make a difference.[8]

[8] As an additional robustness check, in the Appendix we exploit variation in the starting year of state-led family planning programs in 31 countries to further explore the relationship between fertility decline and the establishment of these programs. After controlling for changes in covariates as well as shocks that might have affected fertility in all countries in a given year, we find that the decline in fertility accelerated

Table 5
Change in Total Fertility Rates (TFRs) and Exposure to Family Planning Messages

	Dependent variable is: Change in TFR			
	Absolute change		*% Change*	
	(1)	(2)	(3)	(4)
% of women with exposure to family planning messages on mass media	–0.038*** [0.007]	–0.050*** [0.011]	–0.602*** [0.090]	–0.449** [0.169]
Change in years of education of adults		0.054 [0.154]		0.001 [0.002]
Change in urban population as % of total		–0.035** [0.016]		–0.016 [0.010]
Change in ln(GDP per capita)		–0.529** [0.244]		–0.379* [0.197]
Change in infant mortality rate		0.002 [0.005]		0.551*** [0.175]
Observations	57	30	57	30
R^2	0.301	0.567	0.347	0.631

Source: Authors. Data on total fertility rate, urban population, per capita GDP, and infant mortality rate are from the World Development Indicators. Data on years of schooling are from Barro and Lee (2013). Data on exposure to family planning messages are from Demographic and Health Surveys from various years.
Note: The table reports the results of regressions of the change in total fertility rate between 2013 and 1960 on the percentage of women exposed to family planning messages through mass media for the earliest year (before 2005) for which information is available for that country, controlling for the change between 2013 and 1960 in years of schooling of the population aged 25+, urban population as a percentage of total population, log GDP per capita, and infant mortality rate. All regressions include a constant. Given the small number of observations for infant mortality rate and GDP per capita in 1960, we use the earliest available observation before 1965 to construct the change. The values in parentheses are robust standard errors.
*, **, and *** indicate significance at 10, 5, and 1 percent levels, respectively.

These exercises demonstrate a strong association between the establishment and intensity of family planning programs and the decline in fertility rates, after adjusting for changes in per capita income, urbanization, infant mortality, female labor force participation, and educational attainment. Most sub-Saharan African governments acknowledged rapid population growth as a policy concern much later than developing countries elsewhere. Even after the formulation of population control policies, commitment to family planning lagged behind that of other regions leading most international agencies working in family planning to invest their resources in the more

with their inception. Given the very small sample size, which comprises mainly the early adopters of family planning, we do not place too much weight on these results, but consider it to be further suggestive evidence in favor of the importance of these programs in accelerating the fertility decline. The results of this analysis are available in the online Appendix Table A6, available at http://e-jep.org.

promising areas of Asia and Latin America. The onset of the HIV/AIDS epidemic is also likely to have weakened the emphasis on fertility control due to limited resources being targeted towards addressing the epidemic as well as the emergence of a pronatalist response to the high mortality rates caused by the epidemic (National Research Council Working Group on Factors Affecting Contraceptive Use 1993). While almost all African countries now provide direct or indirect support for family planning, their efforts have only recently caught up with the rest of the world. Perhaps not surprisingly in light of the strong correlations, the countries in sub-Saharan Africa tend to be the ones where fertility rates still remain above the world's average.

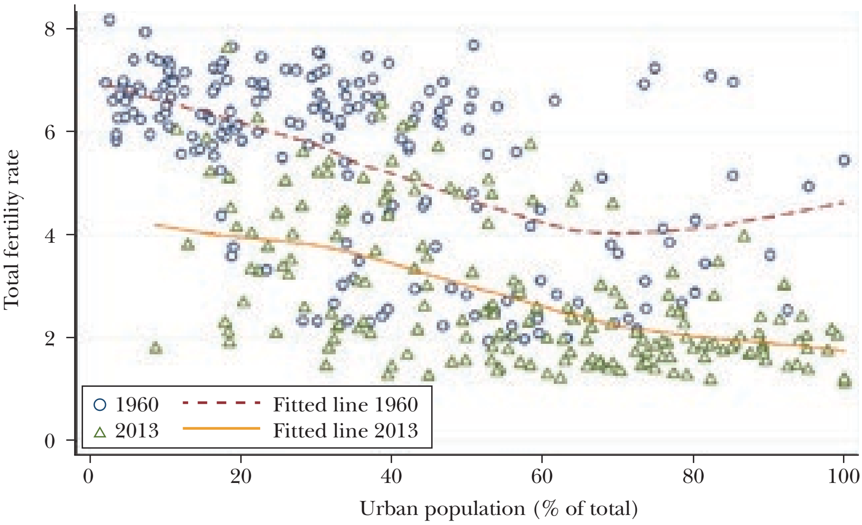
Considering Other Explanations for the Decline in Fertility

A number of other socioeconomic factors have been suggested as possible causes for the decline in fertility: urbanization, greater investment in education per child, rising female labor force participation, and lower infant mortality (Becker 1960; Becker and Barro 1988; Barro and Becker 1989; Manuelli and Sheshadri 2009). The regressions presented in the previous section indicate that population-control policies are strongly associated with the fertility decline, whereas some of the traditional covariates display a much weaker association. Of course, these results are hardly conclusive, as disentangling cause and effect in this area is quite difficult, an issue compounded by the shortage of data and potential measurement error. In this section, we provide further arguments for why these factors, while important, are unlikely to overshadow the role of population-control policies in the fertility decline.

Urbanization has been put forward as an explanation for the decline in fertility, as rural areas have historically had much higher fertility rates than urban ones. Arguably, in rural areas, children can be a significant input in agricultural production. Moreover, despite the fact that parents can earn higher average wages in urban areas, it can cost more to raise children there, as the costs of housing and (typically compulsory) education are higher.[9] The negative relationship between urbanization and fertility is illustrated in Figure 5, which plots the proportion of population living in urban areas against the total fertility rate for all countries in 1960 and in 2013. Although countries with less urbanization have higher fertility, it does not appear that the urbanization process alone can account for the sharp decline in fertility rates observed over the past five decades. Rather, it appears that fertility rates fell rapidly in both urban and rural areas.

[9] Becker (1960) argues that urbanization could explain the decline in fertility. The idea is that farmers have a comparative advantage in producing children and food, though this advantage is smaller for higher "quality" of childrearing. Caldwell (1976)'s net wealth flow theory also supports the view that wealth flows from children to parents in primitive agricultural societies, whereas the direction of flows reverses as society modernizes and costs of raising children go up.

Figure 5
Fertility and Urbanization



Source: Authors using data from the World Development Indicators database.
Note: For a sample of 190 countries, the figure shows the scatter plot and fitted line (smoothed lowess relationship, or locally weighted smoothing function) between fertility and urbanization in 1960 and 2013. Urbanization is measured as the proportion of the population living in urban areas.

Given the strong possibility that the cross-country data on urbanization is mismeasured, we explored this issue in more detail using the Demographic and Health Survey (DHS) data from 57 countries which, through their identification of rural and urban areas, provide separate rural and urban fertility rates. The decline in fertility can be decomposed into a within-area effect, corresponding to the decline in fertility within either rural or urban areas, and a between-area effect (that is, the urbanization effect), corresponding to the decline in fertility rates due to the increase in the share of the population living in (lower-fertility) urban areas rather than (higher-fertility) rural areas.[10] Perhaps surprisingly, the increased urbanization (between-area effect) contributed to only about 14 percent of the fertility decline. Most of the decline in fertility is explained by the within-area effect. Moreover, the contribution of urbanization to the decline in fertility does not vary significantly with a country's fertility or urbanization rates. This result suggests that while urbanization may be a small part of the decline in fertility rates, other forces have been at work driving down fertility in both rural and urban areas around the world.

[10] It should be noted that, because these surveys were carried out in different years and at different intervals in different countries, the period over which the changes are computed is not the same for every country. Details of the data and calculations are available in the online Appendix B available with this paper at http://e-jep.org.

Table 6
Fertility Change by Education in 2010

Schooling in 2010	*Absolute change in total fertility rate, 1960–2013*	*% change in total fertility rate, 1960–2013*	*Total fertility rate in 2010*
Years ≤ 3	−1.35	−19.12	5.87
3 < years ≤ 6	−3.23	−52.26	3.15
6 < years ≤ 9	−4.09	−67.23	2.04
9 < years ≤ 12	−1.67	−43.50	1.73
Years > 12	−1.51	−45.22	1.81

Source: Authors. Data on fertility are from the World Development Indicators database and "years of schooling" comes from Barro and Lee (2013).
Note: The table presents the average absolute and percentage change in total fertility rate between 2013 and 1960 as well as average total fertility rate in 2010 by years-of-schooling groups. Years of schooling is grouped into five categories: years ≤ 3; 3 < years ≤ 6; 6 < years ≤ 9; 9 < years ≤ 12; and years > 12. "Years of schooling" is for the population aged 25+ in 2010 and covers 143 countries.

The decline in fertility is often discussed as being part of a shift away from the quantity of children towards higher quality, as demonstrated by the increase in education levels around the world. There is clearly a strong negative relationship between fertility and education, but it is difficult to establish the direction of causality between fertility and education given that they are both endogenous outcomes of a household's decision-making process. For example, quantity–quality trade-offs are analyzed in Galor and Weil (2000) and Galor and Moav (2002), where technological growth, by raising the return to human capital, can generate a demographic transition (see also Doepke 2004). The link between fertility and education emerges not just because of a trade-off between quantity and quality (or education) of the children, but also because educated parents choose to have fewer children, possibly because they attach more value to quality in that trade-off or they have a comparative advantage in educating children (Moav 2005). Remarkably, fertility has fallen significantly even in countries and rural areas where educational attainment still remains low. For instance, Bangladesh, Morocco, Myanmar, and Nepal all recorded fertility rates below 2.7, with percentage declines of over 60 percent from their 1960 levels, despite their populations having less than five years of schooling on average in 2010. Table 6 presents the average fertility rate in 2010 and fertility change (between 2013 and 1960) for countries grouped by the level of education of the adult population in 2010. While fertility rates are clearly declining in the years of schooling of the population, all but the lowest education group display sizeable percentage declines in fertility. The countries with less than three years of schooling in 2010 are nearly all in sub-Saharan Africa, where the fertility remains very high.

The cross-country correlation between female labor force participation and fertility indicates only a weak relationship, given the high female labor force participation in European and North American countries as well as in sub-Saharan African countries. (Data on female labour force participation rates are obtained from

ILOSTAT.) Furthermore, labor force participation rates did not change much over the past few decades, other than in Latin America and the Caribbean, where the female labor force participation rate rose from 34 percent in 1980 to 54 percent in 2013. For comparison, over the same period, female labor force participation fell slightly in East Asia and the Pacific (from 64 to 61 percent) and South Asia (from 35 to 30 percent), while it rose slightly in the Middle East and North Africa (from 18 to 22 percent) and Sub-Saharan Africa (from 57 to 64 percent).

Changes in infant mortality rates appear to be highly correlated with changes in fertility. There are two, not mutually exclusive, interpretations of this correlation. First, as infant mortality declines, fewer births are needed to ensure that a family's desired number of children survives to adulthood (for example, Kalemli-Ozcan 2002). The second interpretation, which we have emphasized in this paper, is that the decline in mortality rates and the consequent population acceleration in the 1950s and 1960s, triggered the population-control movement; this, in turn, with its emphasis on changing family-size norms and contraception provision, accelerated the fertility fall by reducing the desired number of children and the number of unwanted births.

With regard to the first interpretation (that as infant mortality declines, fewer births are needed), it is apparent that fertility rates did not react quickly to the decline in mortality rates in the mid-20th century; after all, it is precisely the relatively slow change in fertility compared to the relatively rapid growth in life expectancy that caused the remarkable acceleration in population growth in the 1950s and 1960s. As noted in the Report of the President's Committee to Study the US Military Assistance Program (Draper 1959), "high fertility rates are normally part of deeply rooted cultural patterns and natural changes occur only slowly." This was also the view shared by demographers at the time (Enke 1960; Davis 1967). Our regression analysis in the previous section has attempted to gauge the two channels—the direct effect of infant mortality declines, and population-control programs—separately and both appeared relevant. Another way to tease out the role played by population-control programs, as separate from the direct effect of infant mortality, is to study trends in desired or ideal number of children and the share of unwanted pregnancies, which are two main targets of the population-control programs. In principle, according to the first interpretation, lower mortality rates should only affect the number of births, not the ideal number of surviving children.[11] Population-control programs, however, focused on influencing the desired number of children or family size.

The data from the Demographic and Health Surveys provide two measures aimed at capturing fertility preferences: one is the "ideal number of children" and the other is "wanted fertility rate." The ideal number of children is obtained as a response to the question: "If you could go back to the time you did not have any children and could choose exactly the number of children to have in your whole life,

[11] Interestingly, the Barro and Becker (1989) framework predicts that, as mortality rates fall, the ideal (or, in the jargon, "optimal") number of surviving children actually increases, as the cost of raising children decreases. See Doepke (2005), who analyses different variants of the Barro–Becker model yielding this prediction.

Table 7
Changes in Wanted and Unwanted Fertility
(as a percentage of change in total fertility rate)

	Overall	*Urban*	*Rural*
Change in *wanted* fertility rate	75.35%	63.48%	82.26%
Ideal number of children	57.97%	56.08%	51.92%
Other	17.38%	7.41%	30.35%
Change in *unwanted* fertility rate	24.65%	36.52%	17.74%

Note and Source: The table shows the change in wanted fertility rate and unwanted fertility rate (defined as the difference between total and wanted fertility rates) as a percentage of the change in total fertility rate using data from the Demographic and Health Surveys in 52 countries. The change in wanted fertility is further decomposed into the contribution of the change in the ideal number of children and a residual. Note that different countries were surveyed in different years and at different intervals—the earliest available survey is from 1986 while the latest is from 2015.

how many would that be?" The wanted fertility rate is constructed as the fertility rate that would be observed if all "unwanted" births were eliminated; that is, births that raise the number of surviving children over the stated desired number of children (Rutstein and Rojas 2006). We consider the ideal or "desired" number of children as a measure of preference for surviving children—the number of children the woman would choose to have in her whole life. In this context, fertility is directly affected by the desired number of children, but can deviate from it for reasons that are unrelated to preferences, such as infant mortality or the availability of means to control fertility. In particular, the wanted total fertility rate can exceed the desired number of children when women replace children who have died with additional births to reach the desired number of surviving children (Bongaarts 2011). Table 7 uses data from Demographic and Health Surveys in 52 countries to present the average change in wanted fertility rates as a percentage of the change in total fertility rate over the period analyzed (different countries were surveyed in different years and at different time intervals, so the period over which the changes are computed differ across countries). The change in wanted fertility is further decomposed into the contribution of changes in the desired number of children and a second (residual) component that captures other reasons, which might include changes in infant mortality (under the heading "other"). The data indicates that the fall in wanted fertility accounts for a significant share of the fall in fertility, and that a large part of the fall in wanted fertility can be accounted for by the decline in the number of desired children. The pattern is observed in both rural and urban areas. The large role played by the change in the desired or ideal number of children is supportive of the role played by population programs over and above the direct effect of lower mortality rates.

The last row of Table 7 reports the change in unwanted fertility also as a share of the change in total fertility rate. Unwanted fertility is defined as the difference between total fertility rate and wanted fertility. Unwanted fertility has also fallen in both urban and rural areas, pointing to improved ability to control fertility given the wider availability of contraceptives. The decline in unwanted fertility is relatively less important as a share of the change in overall fertility. This, together with the large share accounted for by the decline in the ideal number of children, is consistent with the introduction of additional measures to promote a smaller family size as a result of the sluggish fertility response to wider contraception provision.

Conclusion

The rapid decline in fertility rates in the past five decades cannot be accounted for in a satisfactory way by economic growth, urbanization, education levels, or other socioeconomic variables. The timing and speed of the fertility decline coincides with the growth of a neo-Malthusian global population-control movement that designed and advocated a number of policy measures aimed at lowering fertility rates across the world. The precise measures chosen by different countries varied in nature and scope, depending on the individual country's socioeconomic context. But common to almost all programs was an enhanced provision of contraceptive methods and mass-media campaigns to establish a new small-family norm.

The global convergence in fertility to near replacement fertility rates will eventually ensure a constant world population, although the rise in life expectancy implies that it will take another few decades to reach a constant population level. Projections by the UN Population division suggest that populations in all regions except for Africa will stabilize by 2050. Including Africa, for which the projections are more uncertain, world population is expected to stabilize by 2100 at around 11.2 billion, with total fertility rates converging to 2 in all regions (United Nations Population Division 2015). Concerns over possible imbalances between resources and population will not disappear, but will be mitigated as population growth flattens out. Insofar as the US experience can offer guidance, the diffusion of contraception and the decline of fertility and postponement of childbearing could increase female empowerment in developing countries through higher levels of investment in human capital (Goldin and Katz 2002). To the extent that lower fertility rates are associated with higher investment in human capital, the trends bode well for development and living standards in the world's poorest regions.

■ *For helpful conversations and comments we thank Charlie Bean, Robin Burgess, Francesco Caselli, Laura Castillo, Per Krusell, Omer Moav, Elizabeth Murry, and Gerard Padro-i-Miquel. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.*

References

Bailey, Martha J. 2013. "Fifty Years of Family Planning: New Evidence on the Long-Run Effects of Increasing Access to Contraception." *Brookings Papers on Economic Activity* 1: 341–409.

Bandiera, Oriana, Niklas Buehren, Robin Burgess, Markus Goldstein, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. 2014. "Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa." Suntory and Toyota International Centres for Economics and Related Disciplines (STICERD) Economic Organisation and Public Policy Discussion Paper 50.

Barro, Robert J., and Gary S. Becker. 1989. "Fertility Choice in a Model of Economic Growth." *Econometrica* 57(2): 481–501.

Barro, Robert, and Jong-Wha Lee. 2013. "A New Data Set of Educational Attainment in the World, 1950–2010." *Journal of Development Economics* 104: 184–98. http://www.barrolee.com/ (accessed June 18, 2015).

Becker, Gary S. 1960. "An Economic Analysis of Fertility." In *Demographic and Economic Change in Developed Countries*, edited by George B. Roberts, 209–40. New York: Columbia University Press.

Becker, Gary S., and Robert J. Barro. 1988. "A Reformulation of the Economic Theory of Fertility." *Quarterly Journal of Economics* 103(1): 1–25.

Begum, Hasna. 1993. "Family Planning and Social Position of Women." *Bioethics* 7(2–3): 218–23.

Bongaarts, John. 2011. "Can Family Planning Programs Reduce High Desired Family Size in Sub-Saharan Africa?" *International Perspectives on Sexual and Reproductive Health* 37(4): 209–16.

Caldwell, John C. 1976. "Toward a Restatement of Demographic Transition Theory." *Population and Development Review* 2(3–4): 321–66.

Coale, Ansley J. 1969. "The Decline of Fertility in Europe from the French Revolution to World War II." In *Fertility and Family Planning: A World View*, edited by S. Behrman, L. Corsa, and R. Freedman, 3–24. Ann Arbor: University of Michigan Press.

Davis, Kingsley. 1967. "Population Policy: Will Current Programs Succeed?" *Science* 158(3802): 730–39.

Demographic and Health Survey Program. 2015. The DHS Program STATcompiler. http://www.statcompiler.com (accessed June 18, 2015).

Doepke, Matthias. 2004. "Accounting for Fertility Decline during the Transition to Growth." *Journal of Economic Growth* 9(3): 347–83.

Doepke, Matthias. 2005. "Child Mortality and Fertility Decline: Does the Barro–Becker Model Fit the Facts?" *Journal of Population Economics* 18(2): 337–66.

Draper, William H., Jr. 1959. *Composite Report of the President's Committee to Study the U.S. Military Assistance Program.* Washington, DC: Government Printing Office.

Enke, Stephen. 1960. "The Economics of Government Payments to Limit Population." *Economic Development and Cultural Change* 8(4): 339–48.

Fan, Susan. 2007. "Hong Kong: Evolution of the Family Planning Program." In *The Global Family Planning Revolution: Three Decades of Population Policies and Programs*, edited by Warren C. Robinson and John A. Ross, 193–200. Washington, DC: World Bank.

Freedman, Ronald, and Bernard Berelson. 1976. "The Record of Family Planning Programs." *Studies in Family Planning* 7(1): 1–40.

Galor, Oded, and David N. Weil. 2000. "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond." *American Economic Review* 90(4): 806–28.

Galor, Oded, and Omer Moav. 2002. "Natural Selection and the Origin of Economic Growth." *Quarterly Journal of Economics* 117(4): 1133–91.

Goldin, Claudia, and Lawrence F. Katz. 2002. "The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions." *Journal of Political Economy* 110(4): 730–70.

Hull, Terence H. 2007. "Formative Years of Family Planning in Indonesia." In *The Global Family Planning Revolution: Three Decades of Population Policies and Programs*, edited by Warren C. Robinson and John A. Ross, 235–56. Washington, DC: World Bank.

Kalemli-Ozcan, Sebnem. 2002. "Does the Mortality Decline Promote Economic Growth?" *Journal of Economic Growth* 7(4): 411–39.

Khanna, Sunil K. 2009. "Population Growth and 'Missing' Girls." In *Fetal/Fatal Knowledge: New Reproductive Technologies and Family-Building Strategies in India,* 57–74. Belmont, CA: Cengage Learning.

Kim, Taek Il, and John A. Ross. 2007. "The Korean Breakthrough." In *The Global Family Planning Revolution: Three Decades of Population Policies and Programs*, edited by Warren C. Robinson and John A. Ross, 177–92. Washington, DC: World Bank.

La Ferrara, Eliana, Alberto Chong, and Suzanne

Duryea. 2012. “Soap Operas and Fertility: Evidence from Brazil.” *American Economic Journal: Applied Economics* 4(4): 1–31.

Lapham, Robert J., and W. Parker Mauldin. 1984. “Family Planning Program Effort and Birthrate Decline in Developing Countries.” *International Family Planning Perspectives* 10(4): 109–18.

Liang, Qiusheng, and Che-Fu Lee. 2006. “Fertility and Population Policy: An Overview.” In *Fertility, Family Planning, and Population Policy in China*, edited by Dudley L. Poston, Che-Fu Lee, Chiung-Fang Chang, Sherry L. McKibben, and Carol S. Walther, 7–18. New York: Routledge.

Manuelli, Rodolfo E., and Ananth Seshadri. 2009. “Explaining International Fertility Differences.” *Quarterly Journal of Economics* 124(2): 771–807.

Moav, Omer. 2005. “Cheap Children and the Persistence of Poverty.” *Economic Journal* 115(500): 88–110.

National Research Council Working Group on Factors Affecting Contraceptive Use. 1993. *Factors Affecting Contraceptive Use in Sub-Saharan Africa: Population Dynamics of Sub-Saharan Africa.* Washington, DC: National Academy Press.

Nortman, Dorothy L. 1982. *Population and Family Planning Programs: A Compendium of Data through 1981,* 11th edition, 61–63. New York: Population Council.

Nortman, Dorothy L., and Ellen Hofstatter. 1978. *Population and Family Planning Programs*, 9th edition, pp. 61–63. New York: Population Council.

Piotrow, Phyllis T., and D. Lawrence Kincaid. 2001. “Strategic Communications for International Health Programs.” In *Public Communication Campaigns*, edited by Ronald E. Rice and Charles K. Atkin, 249–68. New York: Sage Publications.

Population Media Center. 2016. “TV Globo Analysis: Brazil.” https://www.populationmedia.org/projects/tv-globo-analysis/ (accessed February 12, 2016).

Robinson, Warren C., and John A. Ross, eds. 2007. *The Global Family Planning Revolution: Three Decades of Population Policies and Programs*, 379–91. Washington, DC: World Bank.

Ross, John A., W. Parker Mauldin, and Vincent C. Miller. 1993. *Family Planning and Population: A Compendium of International Statistics.* New York: Population Council.

Ross, John, and John Stover. 2001. “The Family Planning Program Effort Index: 1999 Cycle.” *International Family Planning Perspectives* 27(3): 119–29.

Rutstein, Shea Oscar, and Guillermo Rojas. 2006. *Guide to DHS Statistics.* Calverton, MD: US Agency for International Development.

Schuler, Sidney R., Syed M. Hashemi, and Ann Hendrix Jenkins. 1995. “Bangladesh’s Family Planning Success Story: A Gender Perspective.” *International Family Planning Perspectives* 21(4): 132–37.

Sinding, Steven W. 2007. “The Global Family Planning Revolution: Overview and Perspective.” In *The Global Family Planning Revolution: Three Decades of Population Policies and Programs*, edited by Warren C. Robinson and John A. Ross, 1–12. Washington, DC: World Bank.

Tien, H. Yuan. 1980. “Wan, Xi, Shao: How China Meets Its Population Problem.” *International Family Planning Perspectives* 6(2): 65–70.

United Nations Population Division. 2013. *World Population Policies Database: 2013 Revision.* http://esa.un.org/PopPolicy/wpp_datasets.aspx (accessed July 20, 2015).

United Nations Population Division. 2015. *World Population Prospects: The 2015 Revision, Key Findings and Advance Tables.* New York: United Nations.

World Bank. 2015. “World Development Indicators.” World Bank. http://data.worldbank.org/data-catalog/world-development-indicators (accessed July 20, 2015).

Zhang, Junsen. 2017. “The Evolution of China’s One-Child Policy and Its Effects on Family Outcomes.” *Journal of Economic Perspectives* 31(1): 141–60.

Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 229–236

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., St. Paul, MN 55105.

Smorgasbord

The IMF has issued a staff report on the topic of "Negative Interest Rate Policies—Initial Experiences and Assessments." "There is some evidence of a decline in loan and bond rates following the implementation of NIRPs [negative interest rate policies]. Banks' profit margins have remained mostly unchanged. And there have not been significant shifts to physical cash. That said, deeper cuts are likely to entail diminishing returns, as interest rates reach their 'true' lower bound (at which point agents shift into cash holdings). And pressure on banks may prove greater;

■ *Timothy Taylor is Managing Editor,* Journal of Economic Perspectives, *based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.229 doi=10.1257/jep.31.4.229

especially in systems with larger shares of indexed loans and where banks compete more directly with bond markets and non-bank credit providers. … On balance, the limits to NIRPs point to the need to rely more on fiscal policy, structural reforms, and financial sector policies to stimulate aggregate demand, safeguard financial stability, and strengthen monetary policy transmission." August 2017, http://www.imf.org/en/Publications/Policy-Papers/Issues/2017/08/03/pp080317-negative-interest-rate-policies-initial-experiences-and-assessments. This report can be read as a complement to the paper by Kenneth Rogoff, "Dealing with Monetary Paralysis at the Zero Bound," in the Summer 2017 issue of this journal.

The *OECD Employment Outlook 2017* devotes a chapter to the topic "Collective Bargaining in a Changing World of Work." "About 80 million workers are members of trade unions in OECD countries, and about 155 million are covered by collective bargaining agreements … On average, 17% of employees are members of trade unions, down from 30% in 1985 … On average across OECD countries, the share of workers covered by a collective agreement has shrunk to 33% in 2015 from 45% in 1985. … Overall, collective bargaining coverage is high and stable only in countries were multi-employer agreements (i.e. at sector or national level) are negotiated and where either the share of firms which are members of an employer association is high or where agreements are extended also to workers working in firms which are not members of a signatory employer association." http://www.oecd.org/els/oecd-employment-outlook-19991266.htm.

Lee Branstetter and Daniel Sichel present "The Case for an American Productivity Revival." "Labor productivity performance in the United States has been dismal for more than a decade. But productivity slowdowns—even lengthy ones—are nothing new in US economic history. This Policy Brief makes the case that the current slowdown will come to an end as a new productivity revival takes hold. Why the optimism? Official price indexes indicate that innovation in the technology sector has slowed to a crawl, but better data indicate rapid progress. Standard measures, focused on physical capital, suggest that business investment is weak, but broader measures of investment that incorporate intellectual and organizational capital report much more robust investment. New technological opportunities in healthcare, robotics, education, and the technology of invention itself provide additional reasons for optimism. This Policy Brief gauges the potential productivity impact of these developments. The evidence points to a likely revival of US labor productivity growth from the 0.5 percent average rate registered since 2010 to a pace of 2 percent or more. A productivity revival of this magnitude would provide a solid foundation for steady increases in wages …" Peterson Institute of International Economics, June 2017, Policy Brief 17-26, https://piie.com/system/files/documents/pb17-26.pdf.

The National Cancer Institute and the World Health Organization offer a nearly 700-page overview of *The Economics of Tobacco and Tobacco Control*. "The global health and economic burden of tobacco use is enormous and is increasingly borne by low- and middle-income countries. Already, around 80% of smokers live in LMICs. While smoking prevalence is falling at the global level, the total number of smokers worldwide is not decreasing, largely due to population growth. … Significant tobacco tax and price increases, comprehensive bans on tobacco industry marketing activities, and prominent pictorial health warning labels are generally the least costly tobacco control interventions, followed by the implementation and enforcement of smoke-free policies and the provision of population-wide tobacco cessation programs. Significant tobacco tax and price increases are the most cost-effective of these interventions." December 2016, https://cancercontrol.cancer.gov/brp/tcrb/monographs/21.

Some Shifting Patterns of International Trade

Chiara Criscuolo and Jonathan Timmis discuss "The Relationship Between Global Value Chains and Productivity." "The bulk of trade is comprised not of final goods or services, but of trade in intermediate parts and components and intermediate services. Among OECD economies, trade in intermediate inputs accounted for 56 percent of total goods trade and 73 percent of services trade over the period 1995–2005. … GVCs [global value chains] present a new means to access international markets: economies need no longer build complete supply chains at home; instead, they can leverage foreign inputs in their production. … GVCs are a well-established vehicle for productivity spillovers to local firms. A substantial part of GVC integration is mediated through FDI [foreign direct investment] … A large literature has investigated FDI spillovers and arrives at a broad consensus in favour of positive productivity spillovers to industries that supply multinationals through backward linkages, with little evidence through other linkages … Knowledge acquisition is an important motive for FDI, which may increase the scope for knowledge diffusion. Firms may relocate some activities, including innovation activities, to obtain access to so-called strategic assets—skilled workers, technological expertise, or the presence of competitors and suppliers—and learn from their experience." *International Productivity Monitor*, Spring 2017, vol. 32, pp. 61–83, http://www.csls.ca/ipm/32/Criscuolo_Timmis.pdf.

The *Global Value Chain Development Report 2017* has the theme of "Measuring and Analyzing the Impact of GVCs on Economic Development." It includes an "Executive Summary" by David Dollar and eight chapters written by other contributors. Here is Dollar's explanation of the "smile curve," in which most of the gains from a global value chain happen at the front-end and tail-end of the chain—not the middle: "The logic of the smile shape is as follows. Research and design activities for critical components of the electrical and optical equipment occur early in the production process … These knowledge activities tend to be high-value-added activities in GVCs [global value chains] and tend to be carried out in more advanced economies. For example, in the 1995 curve Japan and the United States (JPN28 and USA28) are in the upper left corner, reflecting the high-value-added contributions from these two countries' financial services sector. The Chinese industry that manufactures the good, Chinese electrical and optical (CHN14), is at the bottom point of the curve, reflecting assembly activity at low wages. The activities closest to the consumer are marketing, logistics, and after-product servicing. These market knowledge industries are also high value added, as shown by the upward-sloping part of the smile curve on the right. And they tend to be carried out in advanced economies, where the mass consumption products are eventually purchased by households. … [The smile curve] captures anxieties felt by both rich and poor countries in contemplating contemporary trade. Rich-country electorates worry that manufacturing is being hollowed out—that is, that semiskilled production jobs have moved to developing countries or, to the extent that such jobs still remain in advanced economies, have suffered downward pressure on wages. Poor countries worry that they are trapped in low-value-added activities and are locked out of the higher value-added activities in design, key technological inputs, and marketing." Published by the World Bank Group, the Institute of Developing Economies, the Organisation for Economic Co-operation and Development, the Research Center of Global Value Chains headquartered at the University of International Business and Economics, and the World Trade Organization. 2017. https://www.brookings.edu/wp-content/uploads/2017/07/tcgp-17-01-china-gvcs-complete-for-web-0707.pdf.

Michael O'Sullivan and Krithika Subramanian make a case for the likely emergence of a multipolar world trading system in "Getting over Globalization." "We believe that the world is now leaving globalization behind it and moving to a more distinct multipolar setting. … The … scenario is based on the rise of Asia and a stabilization of the Eurozone so that the world economy rests, broadly speaking, on three pillars—the Americas, Europe and Asia (led by China). In detail, we would expect to see the development of new world or regional institutions that surpass the likes of the World Bank, the rise of 'managed democracy' and more regionalized versions of the rule of law—migration becomes more regional and more urban rather than cross-border, regional financial centers develop and banking and finance develop in new ways. At the corporate level, the significant change would be the rise of regional champions, which in many cases would supplant multinationals. … An interesting and intuitive way of seeing how the world has evolved from a unipolar one (i.e. USA) to a more multipolar one is to look at the location of the world's 100 tallest buildings. The construction of skyscrapers (200 meters plus in height) is a nice way of measuring hubris and economic machismo, in our opinion. Between 1930 and 1970, at least 90% of the world's tallest buildings could be found in the USA, with a few exceptions in South America and Europe. In the 1980s and 1990s, the USA continued to dominate the tallest tower league tables, but by the 2000s there was a radical change, with Middle Eastern and Asian skyscrapers rising up. Today about 50% of the world's tallest buildings are in Asia, with another 30% in the Middle East, and a meager 16% in the USA, together with a handful in Europe. In more detail, three-quarters of all skyscraper completions in 2015 were located in Asia (China and Indonesia principally), followed by the UAE and Russia. Panama had more skyscraper completions than the USA." Credit Suisse Research, January 2017, http://publications.credit-suisse.com/tasks/render/file/index.cfm?fileid=BCD82CF0-CF9D-A6CB-BF7ED9C29DD02CB1.

Africa: Urbanization and Electrification

Roland White, Jane Turpie, and Gwyneth Letley consider *Greening Africa's Cities: Enhancing the Relationship between Urbanization, Environmental Assets, and Ecosystem Services*. "Urbanization in Africa began later than in any other global region and, at a level of about approximately 40%, Africa remains the least urbanized region in the world. However … this is rapidly changing: SSA's [sub-Saharan Africa's] cities have grown at an average rate of close to 4.0% per year over the past twenty years, and are projected to grow between 2.5% and 3.5% annually from 2015 to 2055. … From an environmental perspective, this has two important implications. On the one hand, most of Africa's urban space has yet to emerge. Much of the area which will eventually be covered by the built environment has not yet been constructed and populated. Crucial natural assets—and significant biodiversity—thus remain intact in areas to which cities will eventually spread. On the other hand, this is changing quickly: pressures on the natural environment in and around cities are escalating steadily and these assets are increasingly under serious threat." World Bank, May 2017, https://openknowledge.worldbank.org/handle/10986/26730.

Simone Tagliapietra raises the challenge of "Electrifying Africa: How to Make Europe's Contribution Count." "Less than a third of the sub-Saharan population has access to electricity, and around 600,000 premature deaths are caused each year by household air pollution resulting from the use of polluting fuels for cooking and lighting. … Electrification rates in sub-Saharan African countries average 35 percent … The situation is even starker in rural areas, where the average electrification rate in sub-Saharan Africa stands at 16 percent … Furthermore, the number of people living without electricity in sub-Saharan Africa is rising, as ongoing electrification efforts are outpaced by rapid population growth. … In sub-Saharan Africa, average electricity consumption per capita is 201 kilowatt-hours (kWh) per year, compared to 4,200 kWh in South Africa and 1,500 kWh in North African countries. The situation is even worse in rural areas of sub-Saharan Africa with access to electricity, where electricity consumption per capita remains even below 100 kWh per year." Bruegel Policy Contribution, Issue #17, June 2017, http://bruegel.org/wp-content/uploads/2017/06/PC-17-2017-1.pdf.

Applications of Blockchain

Philip Boucher, Susana Nascimento, and Mihalis Kritikos discuss "How Blockchain Technology Could Change Our Lives." "There are many different ways of using blockchains to create new currencies. *Hundreds* of such currencies have been created with different features and aims. The way blockchain-based currency transactions create fast, cheap and secure public records means that they also can be used for many non-financial tasks, such as *casting votes in elections* or *proving that a document existed at a specific time*. Blockchains are particularly well suited to situations where it is necessary to know ownership histories. For example, they could help manage *supply* *chains* better, to offer certainty that diamonds are ethically sourced, that clothes are not made in sweatshops and that champagne comes from Champagne. They could help finally resolve the problem of music and video piracy, while enabling *digital media* to be legitimately bought, sold, inherited and given away second-hand like books, vinyl and video tapes. They also present opportunities in all kinds of *public services* such as health and welfare payments and, at the frontier of blockchain development, are *self-executing contracts* paving the way for *companies that run themselves* without human intervention." European Parliamentary Research Service, February 2017, http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/581948/EPRS_IDA(2017)581948_EN.pdf.

Michael Pisa and Matt Juden evaluate "Blockchain and Economic Development: Hype vs. Reality." "Increasing attention is being paid to the potential of blockchain technology to address long-standing challenges related to economic development. … [W]e examine its potential role in addressing four development challenges: (1) facilitating faster and cheaper international payments, (2) providing a secure digital infrastructure for verifying identity, (3) securing property rights, and (4) making aid disbursement more secure and transparent." Center for Global Development, July 2017, CGD Policy Paper #107, https://www.cgdev.org/sites/default/files/blockchain-and-economic-development-hype-vs-reality_0.pdf.

Interviews with Economists

Douglas Clement offers an "Interview with Hilary Hoynes." "Food stamps started under President Kennedy. His first executive action was to start some pilot programs for food stamps. … Those pilot programs eventually led to passage of the Food Stamp Act in 1964. But it wasn't until 1974, 10 years later, that subsequent legislation compelled all areas to implement food stamps. … This resulted in gradual rollout of food stamps across the almost 3,200 U.S. counties. … The 'rollout design' is one of the tools in our tool bag for doing evaluation. … We couldn't in our data know precisely which families were on food stamps, so it's sort of an indirect estimate. But we know whether food stamps were implemented when these individuals were 2 or 4 or 14 or 20 years old. We essentially analyzed the data within that lens: How old were you when food stamps were rolled out in your county? … The headline finding was about health. We measured metabolic syndrome, which is essentially a range of conditions including high blood pressure, diabetes, heart disease and obesity. … And we found that the more exposure to food stamps that a person had, the lower their risk of metabolic syndrome in adulthood. In particular, the gains were greatest if the food stamps program was implemented before an individual was 3 or 4 years old. That period between in utero exposure—prebirth—to those first three or four years of life, was the age range where having more exposure to food stamps available led to a more dramatic reduction in the incidence of metabolic syndrome in adulthood. …" *The Region*, Federal Reserve Bank of Minneapolis, June 1, 2017. https://minneapolisfed.org/publications/the-region/interview-with-hilary-hoynes.

Jessie Romero plays the interlocutor role in "Interview: Janet Currie." "There is a large environmental justice literature arguing that low-income and minority people are more likely to be exposed to a whole range of pollutants, and that turns out to be remarkably true for almost any pollutant I've looked at. A lot of that has to do with housing segregation; areas that have a lot of pollution are not very desirable to live in so they cost less, and people who don't have a lot of money end up living there. It also seems to be the case, at least some of the time, that low-income people exposed to the same level of pollutants as higher-income people suffer more harm, because higher-income people can take measures to protect themselves." On another topic: "Many people are concerned about overtreatment and excessive [health care] spending, but the problem is more subtle. Bentley, Jessica Van Parys, and I studied heart attack patients admitted to emergency rooms in Florida. … Young, male doctors who trained at a top-20 medical school were the most likely to treat all patients aggressively, regardless of how appropriate the patient seemed to be. In the case of heart attacks, it appears that all patients have better outcomes with more aggressive treatment, so treating only the 'high-appropriateness' patients aggressively harms the 'low-appropriateness' patients. Similarly, many people are concerned that U.S. doctors perform too many C-sections. But actually … it looks like too many women with low-risk pregnancies receive C-sections, while not enough women with high-risk pregnancies receive C-sections. So the goal shouldn't necessarily be to reduce the total number of C-sections but rather to reallocate them from low-risk to high-risk pregnancies." *Econ Focus*, Federal Reserve Bank of Richmond, First Quarter 2017, pp. 23–36, https://www.richmondfed.org/-/media/richmondfedorg/publications/research/econ_focus/2017/q1/pdf/interview.pdf.

Discussion Starters

Arnold Kling asks "How Effective is Economic Theory?" "Economists are not without knowledge. We know that restrictions on trade tend to help narrow interests at the expense of broader prosperity. We know that market prices are important for coordinating specialization and division of labor in a complex economy. We know that the profit incentive promotes the introduction of improved products and processes, and that our high level of well-being results from the cumulative effect of such improvements. We know that government control over prices and production, as in communist countries, leads to inefficiency and corruption. We know that the laws of supply and demand tend to frustrate efforts to make goods more 'affordable' by subsidizing them or to lower 'costs' by fixing prices. But policymakers have goals that go far beyond or run counter to such basic principles. They want to steer the economy using fiscal stimulus. They want to shape complex and important markets, including those of health insurance and home mortgages. It is doubtful that the effectiveness of economic theory is equal to such tasks. … There is a very real possibility that over the next 20 years academic economics will congeal into a discipline, like sociology today, which is definitively shaped by an ideologically driven point of view. … This will be evident in beliefs of economists that are politically consistent but analytically contradictory. For example, it is politically consistent for someone on the left to believe that a rise in the minimum wage would not reduce hiring and also that more immigration would not depress wages. Analytically, however, these are opposite views. … The contemporary state of economic theory reflects a broader crisis in the social sciences and a deepening cleavage between the college campus and the rest of society." *National Affairs*, Summer 2017, http://www.nationalaffairs.com/publications/detail/how-effective-is-economic-theory.

Charles D. Kolstad investigates "What Is Killing the US Coal Industry?" "[O]ver the past 60 years output of coal more than doubled … Despite great expansion in coal production over the past half century, employment has steadily declined … In the first decade of the new millennium, productivity gains—this time in natural gas—generated a fundamental shift in which coal was no longer clearly the cheapest fossil fuel. At the same time, solar and wind have made significant inroads into electricity generation, again providing a competitive threat to coal. Productivity gains, in coal, gas, and other energy sources, have been a primary force of change. This buildup of pressures has finally resulted in the retirement of very old coal-fired generating units that were built before most Americans were born. Ironically, many of these retirements would probably have occurred long ago except for the Clean Air Act's preferential treatment of old coal plants. … What is clear from this discussion is that environmental regulations did not kill coal. Progress is the culprit." Stanford Institute for Economic Policy Research Policy Brief, March 2017, https://siepr.stanford.edu/research/publications/what-killing-us-coal-industry.

Justin Bryan provides an overview of "High-Income Tax Returns for Tax Year 2014." "For Tax Year 2014, there were almost 6.3 million individual income tax returns with an expanded income of $200,000 or more, accounting for 4.2 percent of all returns filed for the year. … Of the 9,692 returns without any worldwide income tax and expanded incomes of $200,000 or more, the most important item in eliminating tax, on 54.7 percent of returns, was the exclusion for interest income on State and local Government bonds … The next three categories that most frequently had the largest primary effect on taxes were: 1) the medical and dental expense deduction (15.6 percent or 1,509 returns); 2) the charitable contributions deduction (8.5 percent or 819 returns); and 3) the foreign-earned income exclusion (6.6 percent or 638 returns). The item that was most frequently the secondary effect in reducing regular tax liability on high expanded-income returns with no worldwide income tax was the deduction for taxes paid (24.4 percent or 2,365 returns)." *Statistics of Income Bulletin*, Internal Revenue Service, Summer 2017, https://www.irs.gov/pub/irs-soi/soi-a-inhint-id1705.pdf.

The *Journal of Economic Perspectives*: Proposal Guidelines

Considerations for Those Proposing Topics and Papers for *JEP*

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

Philosophy and Style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives.* Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at <http://e-*JEP*.org>.

Guidelines for Preparing *JEP* Proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP*. However, given the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant editor, Ann Norman, at <anorman@JEPjournal.org>. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Guidelines for Empirical Papers Submitted to *JEP*

The *JEP* is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1) The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.
2) In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.
3) The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary *JEP* for your e-reader, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $20.00, $30.00, or $40.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

The following Statement of Ownership, Management and Circulation is provided in accordance with the requirements, as contained in 39 U.S.C. 3658. *Journal of Economic Perspectives* is owned, managed and published by the American Economic Association, a nonprofit educational organization, located at 2014 Broadway, Suite 305, Nashville, Davidson County, TN 37203-2425. The Editor is Professor Enrico Moretti. The Managing Editor is Timothy Taylor: *Journal of Economic Perspectives*, 2403 Sidney Street, Suite 260, Pittsburgh, PA 15203. The tax status of the American Economic Association has not changed during the preceding twelve months. During the preceding twelve months, the average number of copies printed for each issue was 5,972; the average total paid and/or requested circulation, 5,002; the average total non-requested distribution, 19; the average number of copies not distributed 951. Corresponding figures for May 2017, the issue published nearest to filing date: total number of copies printed, 5,943; total paid and/or requested circulation, 4,826; number of copies not distributed 1,040; total distribution, 4,903. During the preceding twelve months, the average number of requested and paid electronic copies each issue was 962; the total average requested and paid print and electronic copies, 5,964. Corresponding figures for May 2017, the issue published nearest to filing date: number of requested and paid electronic copies, 955; the total requested and paid print and electronic copies, 5,781. Certified by Barbara Fiser, Director of Finance and Administration.

FSC™
www.fsc.org
MIX
Paper from responsible sources
FSC™ C132124

The Journal of

Economic Perspectives

Fall 2017, Volume 31, Number 4

Symposia

Health Insurance and Choice

From Experiments to Economic Policy

Articles

AMERICAN
ECONOMIC
ASSOCIATION