Handbook of Field Experiments:

Health, Health Care and Health Behavior in Developing Countries

Pascaline Dupas and Edward Miguel

Draft, April 2015

Abstract: Higher levels of health in developing countries could considerably improve wellbeing and possibly promote economic growth. The last decade has seen a surge in field experiments designed to understand the barriers that households and governments face in investing in health and how these barriers can be overcome, and to assess the impacts of subsequent health gains. This chapter first discusses the methodological pitfalls that field experiments in the health sector are particularly susceptible to, then reviews the evidence that rigorous field experiments have generated so far.

1. Methodological Section

Experiments in the health sector have been prominent among the field experiments carried out in development economics over the past two decades, and they have highlighted a number of important methodological issues related to the estimation of externalities, variable measurement, and pre-registration and research transparency. We discuss each of these issues in turn in the three subsections that follow.

Before launching into the detailed discussion, a few observations about the differences between field experiments conducted by development economists and those carried out among clinical trialists and epidemiologists are in order. One key difference relates to the goals of their studies, in relation to the widely used distinction between *efficacy trials* and *efficiency trials*. Efficacy trials are designed to capture the impact of an intervention under the most controlled and ideal circumstances possible, while efficiency trials capture effects under more authentic real-world conditions (Singhal et al 2014).

In reality, many studies lie somewhere in between these two extremes, with both partial study control and some degree of realism. While medical researchers typically carry out both types of studies, and often make a sharp distinction between the two, most recent field experiments in economics have tended to be closer to efficiency trials. Many of these studies have been carried out in close collaboration with government or non-governmental organization (NGO) partners, and have been implemented as "real" programs, rather than experiments that are carried out directly by the researchers themselves, as in many efficacy trials.

This distinction between the types of studies carried out by medical researchers versus development economists working on health topics has a number of important implications that will become apparent in the course of this chapter. One has to do with the quality "standards" and perceptions of the "risk of bias" in a particular design. For medical trialists accustomed to the CONSORT standards or other medical efficacy trial reporting guidelines, studies that do not feature double-blinding, and thus run the "risk" of endogeneous behavioral responses to the medical intervention, are considered less reliable than those studies that employ double-blinding (for a detailed discussion, see Eble et al 2015). While a few of the studies conducted by economists surveyed below do feature double-blinding (most notably Thomas et al 2003, 2006), in nearly all settings blinding participants to their status is either logistically difficult (for instance, if government partners are unwilling to distribute placebo treatments to some of their population) or even impossible.

To illustrate, how would you provide a placebo treatment in a study investigating the impact of the distribution of anti-malarial bednets? Even in settings that might seem promising for placebo treatments, such as the community-level deworming treatments discussed below at several points, blinding participants to their status is basically impossible. Deworming generates side effects (mainly gastrointestinal distress) in roughly 10% of those who take the pills, so community members in a placebo community would quickly deduce that they were in fact not receiving real deworming drugs if there are few or no cases of side effects. Even more importantly, endogeneous behavioral responses are often exactly what we economists (and other social scientists) set out to measure and estimate in our field experiments, and thus are to be embraced rather than rejected as symptomatic of a "low quality" research design that is at "high risk of bias". Finally, economists' interest in many cases in the cost-effectiveness, economic returns, or fiscal implications of particular real-world health interventions once again make efficacy trials of inherently less interest in most cases than more realistic effectiveness trial approaches. In fact, the differences in outcomes between efficacy trials and effectiveness trials are of great interest to social sciences, since understanding why an intervention that "works" in a highly controlled settings might "fail" in a more realistic setting can shed light on the functioning – and limitations – of existing organizations and institutions.

Taken together, it is clear to us that the experimental literature on health interventions in economics (and increasingly in other social sciences such as political science) often has very different objectives than medical, public health and epidemiological research, and thus different methodologies are often called for. Researchers working on health topics in development economists have not simply been able to import existing medical trial methods wholesale, but have instead been quite innovative in developing new approaches to estimating externalities, in measurements, and regarding issues of pre-registration and transparency, as discussed in the three subsections below.

1.1. Experimenting to estimate impacts of health improvements: beware of externalities

Field experiments in development economics focusing on the health sector have been innovative in adapting and creating new approaches to estimating treatment externalities. Treatment externalities go by many different names in different subfields and disciplines – including spillovers, contamination, herd immunity, and indirect effects, among others – and they have been of interest to statisticians and epidemiologists for a long time (for classic treatments, refer to Cox 1958, Rubin 1990, Fine 1993, and Rosenbaum 2007). However, despite their theoretical importance for the health field, these issues have received far less empirical attention in public health and epidemiological research.

As surveyed in a recent synthetic review (Benjamin-Chung et al. 2015), the rapid growth in empirical studies of treatment externalities in epidemiology began after 2000, at roughly the same time that such empirical treatments became more common in economics, although there were a handful of earlier empirical treatments of the issue (for instance, Paul et al 1962 and Cooper and Fitch 1983). This literature has tended to focus on low income settings in Asia, Latin America and Africa in both the public health literature and in economics. As Benjamin-Chung et al (2015) show in their detailed review of the existing literature, both these early public health studies, as well as most recent treatments (for instance, Forleo-Neto et al 1999, Ali et al 2005, 2008, 2013), tend to focus on "herd immunity" effects in vaccine treatment programs, and they estimate treatment spillovers using the "treatment coverage mean", i.e., the proportion of individuals in the area who received treatment, as the key measure of exposure. They then examine whether there is a lower risk of later infection in sites where more people received treatment relative to areas where fewer people were treated. These studies provide consistent support for the existence of positive vaccine spillover benefits among those who did not receive the vaccine themselves.

This is an empirically sensible approach but it has a number of immediate limitations. First, many studies using this approach typically leave the question of how and why particular sites had lower vaccine coverage than other areas unanswered. This is potentially problematic to the extent that coverage rates are affected by omitted variables ("confounders") such as the local disease environment, capacity of local health institutions, and perhaps local attitudes towards particular diseases, all of which could both affect coverage as well as population health outcomes. For instance, it is plausible that areas where populations have less awareness about, or support for, treatment of a disease might both have lower coverage rates and higher subsequent infection (although there are other possible confounders that would lead to bias in the opposite direction). Second, this approach is typically quite imprecise about the geographic area that is relevant for spillovers, and different studies use different definitions of a community or site, thus leading to a lack of comparability across studies. Taken together, this implies that the evidence base within public health regarding the extent of treatment spillovers is not extremely solid, and moreover, the evidence generated so far has tended to focus on a narrow set of treatments, namely, vaccinations.

While it may be surprising that there has not been more empirical research on the empirical estimation of spillover effects within public health and epidemiology (despite the theoretical centrality of these issues in these fields), we speculate that it may be the result of a tendency in most empirical health studies to focus on "standard" RCT empirical approaches that compare treatment to control groups directly, and that tend to regard any sort of externality effect as a source of "contamination" that is to be avoided or minimized, rather than as a key element of our understanding of the overall treatment effect. Economists who have worked on these topics in health have instead been more open to embracing the estimation of externalities, perhaps in part because norms regarding the "right" way to carry out field experiments are less established in economics (given how recently these tools have been adopted in the field), and also given the importance of spillover effects within public economics theory to potentially rationalize public subsidies for health treatments and interventions (Dybvig and Spatt 1983).

The Miguel and Kremer (2004) paper on school-based deworming impacts in Kenya is among the first health studies in development economics to experimentally estimate externality effects. In their main analysis, they exploit the variation in deworming treatment status generated by the experimental assignment of schools to early versus late treatment (in a phase-in, or stepped wedge, research design), and show that this generates extensive variation in local "exposure" to treated schools within 3 km and up to 6 km away from sample schools. Their econometric approach conditions on the total density of local school pupil population within a particular geographic distance, and conditioning on this quantity, the experimental design implies that the number of treated schools is experimentally assigned and should thus be orthogonal to other local observables and unobservables. They cannot reject that the observed characteristics of schools with lots of "exposure" to other local treatment schools are the same as for schools with little such exposure. Thus this analytical approach – which Benjamin-Chung et al 2015 terms "estimation of spillovers conditional on treatment density" – addresses both of the limitations of most existing empirical research on spillovers in public health described above. In particular, the assignment of exposure to treatment spillovers is assigned experimentally (and thus should be largely free of the possible omitted variable bias, or confounding, that could affect most existing vaccine studies in epidemiology), and this approach also makes precise the extent of externalities within precisely defined geographic distances away from a particular site.

Miguel and Kremer (2004) also use another source of variation to estimate spillover effects within treated school communities. Within communities, a subset of the population was not assigned to treatment, namely, older girls for whom the deworming drugs were considered potentially dangerous at the time of the original study (due to potential embryotoxicity), and other students simply did not receive treatment, usually because they did not attend school on the announced day of treatment or did not receive parental consent for treatment. The comparison of subsequent infection rates among those in treatment schools who did not themselves take deworming drugs, compared to those who did take the drugs, is potentially problematic due to non-random selection into treatment, and any such differences lack a reliable counterfactual (since time trends or other secular changes might affect both groups).

However, Miguel and Kremer (2004) are able to exploit the fact that the same treatment inclusion rules were used in subsequent years of the program as later treatment groups were phased into deworming, and they compare infection rates among those who did not receive deworming treatment when it was available in their school, to those in other schools who were not yet offered deworming but who we know did not receive treatment when later offered the opportunity. This approach potentially addresses much of the "selection" into deworming treatment, as long as patterns of selection remain roughly constant across years 1 and 2 of the study. This is a "within-cluster spillover effect", and when focusing on the excluded older girls, Benjamin-Chung et al (2015) term it a "within-cluster spillover effect among ineligibles".

There is evidence for large and statistically significant externality effects on both worm infection rates, and on subsequent school participation rates, using both sources of variation in Miguel and Kremer (2004), namely, the spillover estimates conditional on local treatment density, and the within-cluster spillover effect. These effects are concentrated within school communities, and extend out to at least 3 km away from treatment schools.

In a follow-up study in the same area of Kenya, Ozier (2014) also generates within-cluster spillover effect estimates among ineligible, by focusing on children who were 0-2 years old when the program was launched, and thus were too young to have directly received deworming treatment. However, they were potentially affected by epidemiological spillovers generated by deworming treatment, since treating infected individuals means they are less likely to pass on worm larvae into the environment through fecal matter (the usual route of transmission for intestinal helminth infections). Ozier finds evidence that the youngest children (those under 2) gain substantially a full 10 years after deworming treatment in terms of their cognitive performance and academic test scores, with average gains of roughly half a school year of learning. This finding reinforces the results in Miguel and Kremer (2004) about the potentially large magnitude of positive deworming treatment externalities in an area with high rates of worm infections; infection rates at baseline in this region of western Kenya were over 90%.

An implication of these externality effects is that research on infectious diseases – or other types of health or economic interventions – that does not account for externalities is likely to underestimate total program effects, both by potentially understating differences across the treatment and control groups (if the control group is gaining relative to the counterfactual of no exposure to spillovers), and by missing out on the spillovers entirely, thus doubly undercounting effects. The existence of treatment externalities thus makes cluster randomized designs – that treat most or all individuals in a given area, and consider the entire unit "treatment" in the analysis – more attractive than individually randomized designs in such settings, since treatment spillover benefits are at least partially "internalized" within the cluster (although as Miguel and Kremer 2004 show, some spillovers may extend beyond the cluster and these could be important to consider as well). We discuss this issue in greater detail below, but the presence of sizeable treatment externalities is a possible explanation for why several of the early studies of deworming treatment impacts on growth and cognition – all of which randomized across individuals within the same community or school – tended to find quite small (although typically positive) effects (see Dickson et al 2000), namely, that the control group gained considerably from the intervention, dampening effects. In contrast, both of the large cluster randomized experiments on deworming discussed below (the Miguel and Kremer 2004 study, as well as the Alderman et al 2007 project in Uganda) find both large short-run and long-run impacts of deworming on participant outcomes.

A large number of studies within economics – including both health and non-health studies – have subsequently utilized the basic empirical approach developed in Miguel and Kremer (2004) in order to estimate the magnitude of treatment externalities. Some of these studies modify the estimator to focus on the share of individuals within one's social network that are affected by a treatment, rather than relying on geographic distance per se, as in the original analysis. In the health sector, this includes studies of mental health (Baird et al 2013), water treatment (Ziegelhofer 2012), learning about HIV results (Godlonton and Thornton 2012), community monitoring of health clinic performance (Bjorkman and Svensson 2009), risky sexual behavior (Dupas 2011), child nutrition (Zivin et al 2011, Fitzsimons et al 2012), family planning (Joshi and Schultz 2013), and malaria prevention (Tarozzi et al 2014, Dupas 2014), as well as a study of take-up of the deworming treatments themselves within a social network (Kremer and Miguel 2007), among many other related research studies.

As might be expected given the diversity of health conditions and behaviors that have been examined using these methods, the magnitude and range of externalities vary considerably across cases. However, it is worth mentioning the estimated effects in some of these cases. The case of malaria is particularly important, given how widespread the condition is in many low-income regions (especially in Africa) and its contribution to the total global burden of disease. Both of the malaria studies in economics mentioned above find suggestive evidence that positive externalities are generated when households use insecticide treated bednets, although neither has adequate statistical power to reach definitive conclusions (Tarozzi et al 2014, Dupas 2014). In contrast to deworming, the malaria spillover benefits tends to be localized within a community, and it appears to within 20 to 30 meters from the household using the net (Tarozzi et al 2014). This information on the magnitude and geographic extent of spillovers can be important for both public health planners, as well as for those considering the desirability of large public subsidies for these, or other, health interventions.

As alluded to above, In other recent work economists have moved beyond studying epidemiological spillovers directly (as in the deworming and malaria cases), and have begun to explore spillovers through social networks in terms of technology adoption and behavioral change (as in Kremer and Miguel 2007 and Dupas 2011, for instance), and also the possibility that spillover could occur through channels other than epidemiology or social influence. For instance, one direct way that externalities might occur is through the sharing of medical treatment between those assigned to treatment and those assigned to control; in the case of a treatment such as iron supplementation which has limited side effects, this is something that might be considered quite low risk among participants (see the discussion of Thomas et al 2003, 2006 and Bobonis et al 2006 for studies on iron supplementation in this literature).

Recent research has made methodological progress in understanding how to most efficiently estimate externality effects, and how to address the possibility of nonlinearities in the relationship and complementarities with local treatment decisions. Bhattacharya et al (2013) exploit experimental variation combined with detailed geospatial information to estimate how the local subsidy rates faced by others affect insecticide-treated mosquito nets (ITN) use in Kenya, and show that there are important non-linearities in the subsidy incidence. The issue of possible non-linearities in social effects are raised as a possibility in both Miguel and Kremer (2004) and Kremer and Miguel (2007) but in neither study was there sufficient statistical power to reject linear specifications. Baird et al (2014) discuss the optimal design of experiments to estimate spillover effects in settings where it is possible to randomize the intensity of treatment within clusters, and then randomly assign individual treatment conditional on this intensity. They include calculations of statistical power to detect externality effects given program parameters, which is useful for those prospectively designing experiments with this aim.

In areas beyond health, spillover effects and related general equilibrium effects are increasingly being studied in a wide range of sectors including in the study of cash transfer programs, microfinance programs, and beyond, demonstrating the analytical usefulness of these approaches to economics research as a whole; Angelucci and Di Maro (2015) provide further discussion of such studies across subfields within development economics.

1.2. Experimenting to understand the determinants of health behavior: beware of measurement

Like other field experiments in development economics, experiments focusing on health topics have often relied on original data collection – including individual and household survey data, biomarker data, as well as data from clinics and schools – in the analysis. As they were with research design issues, these studies have also been highly innovative in their development of new data collection methodologies, as well as in clarifying some of the potential biases that could arise from these different types of original data collection. We discuss these different types of data and biases in turn in this subsection.

The simplest and potentially most pervasive form of bias from original data collection would occur if any act of being surveyed itself affected subsequent responses and, even more importantly, behaviors. Zwane et al (2011) quantifies the possible extent of this bias using data from multiple data collection activities in development economics, all of which featured some randomized variation in the frequency with which different groups of households or individuals were surveyed, and show that being surveyed alone can often affect subsequent behavior in health studies, as well as in microfinance projects.

In the context of the health data that was featured in their analysis, the authors show the randomly chosen individuals in rural Kenya communities who were surveyed more frequently regarding their children's health status (here, the diarrhea outcomes and other health dimensions for infants) were significantly more likely to change their behavior in the direction of making more investments in their children's health, specifically, in the use of a point-of-use chlorine disinfectant for household water. These behavioral responses were large in magnitude and statistically significant among the households surveyed at high frequency (biweekly) relative to those surveyed infrequently (every six months), with a near doubling in use of chlorine disinfectant. This response also appears to have led to large reductions in reported diarrhea, and they are large enough to change the estimated effect of an ongoing water investment campaign (namely, spring protection) in the same region. Taken together, the authors suggest that frequent surveys may serve as a reminder to households to engage in particular health practices, similar to the effect that has been documented for explicit reminders through mobile phone and other means (for instance, see Pop-Eleches et al 2011).

This has extremely important implications for health studies. While many economics studies collecting original data utilize relatively infrequent data collection (presumably for reasons of cost), some like those discussed in Zwane et al (2011) do make use of high frequency data collection, and such approaches are actually the "standard" in many public health studies, such as those studying diarrhea outcomes in children (the health data in Zwane et al was modeled on these approaches). Data reliability might be improved to the extent that data can be collected less frequently from a larger sample of individuals, or to the extent that more "passive" forms of data collection, such as from administrative records or "big data" sources (such as mobile phone usage), rather than high frequency enumerator visits. An alternative that is increasingly employed in field experiments in development economics is the creation of a "pure control" group of households or communities who are not contacted by the research team or surveyed until the very end of the study, when outcome measures are collected; for an example of a study that uses this approach, see Muralidharan and Sundararaman (2011). These individuals are thus unlikely to have been affected by the process of data collection, and any such bias on the "regular control" group can also be quantified in this way.

A related but distinct concern with original data collection relying on surveys is the possibility that respondents will provide answers that they think the enumerators want to hear, what is known as social desirability bias, or experimenter demand effects. These are widely discussed in laboratory data collection in experimental economics, and are increasingly recognized as a major concern in field experimental data collection settings.

In many settings where sensitive health information is collected, researchers are increasingly creating "private" situations within the data collection encounter for them to enter in such data in a way that cannot be immediately verified by the enumerator (for instance, see Baird et al 2008). These concerns may be particularly pronounced when it comes to reproductive health and sexual health topics. To address these concerns, scholars have recently been quite creative in employing enumerators who are more likely to elicit truthful responses from respondents, e.g., Robinson and Yeh (2011, 2012) hire former sex workers to survey other sex workers on their sexual practices and decision-making.

An alternative approach that creates privacy for respondents is a survey technique called list randomization. List randomization, also known as the item count or unmatched count technique, allows respondents to report on potentially sensitive behavior without allowing the researcher or surveyors to identify individual responses. In practice, some proportion of survey respondents are randomly selected to receive a short list of statements (e.g., general health choices and outcomes, say) and asked to report how many, but not which, statements are true. The remainder of the survey respondents are presented with the same list of statements and one key additional statement designed to capture sensitive behavior (e.g., regarding sensitive sexual behavior). By subtracting the mean number of true statements in the first group from the mean number of true statements in the second group, researchers can estimate the proportion of the sample that engages in the sensitive behavior. This approach has been widely used to study health behaviors in many contexts (see Droitcour et al 1991, LaBrie and Earleywine 2000, Chong et al. 2013), as well as sensitive behavioral choices in other spheres (Karlan and Zinman 2011).

Even when techniques such as creating a private space for survey respondents, employing more appropriate enumerators, and list randomizations are used, there remain important concerns about the validity of self-reported health behaviors, especially in sensitive areas such as sexual and reproductive health. A growing number of studies have documented a sharp divergence between self-reported sexual behavior and objectively measured infection status. In data collected from over 10,000 adolescents in Western Kenya, Duflo et al. (2014) find that 4.6% of girls and 4.8% of boys who report that they never had sex test positive for Herpes Simplex Virus type 2 (HSV2), a sexually transmitted infection. Gong (2015) uses field experimental data from Kenya and Tanzania in the context of an HIV/AIDS related testing and information campaign, and shows that self-reported sexual behavior becomes less risky for individuals who were informed that they had tested HIV-positive, even while their incidence of STI infections – a more reliable measure of risky sex – increases significantly.

1.3. Research transparency, registration, and pre-analysis plans

There is growing awareness across social science fields that current research methods and practices can sometimes produce misleading bodies of evidence (Miguel et al 2014). For instance, there is growing evidence documenting the prevalence of publication bias in economics, as well specification searching, and widespread inability to replicate empirical findings. While some of these issues have been widely discussed within economics for some time (see Leamer (1983), Dewald, Thursby, and Anderson (1986) and DeLong and Lang (1992), among others), there has been a recent flurry of activity documenting these problems, and also generating ideas for how to make research more transparent and reproducible.

With the vastly greater computing power of recent decades and the ability to run a nearly infinite number of regressions, there is renewed concern that null-hypothesis statistical testing is subject to both conscious and sub-conscious manipulation. At the same time, technological progress has also facilitated various new tools and potential solutions, including by facilitating the online sharing of data, statistical code, and other research materials, as well as the creation of easily accessible online study registries, data repositories, and tools for synthesizing research results across studies. Progress to date in adopting these practices is partial, with some journals and research communities within economics adopting new practices to promote transparency – including study registration, data sharing, and more detailed disclosure standards – and many others failing to do so.

Before discussing potential remedies, it is worth briefly describing three of the major problems that have been identified: publication bias, specification searching, and inability to replicate.

Publication bias arises if statistically significant results – papers that reject the null hypothesis – are more likely to be published than other results. If we do not keep track of the statistical tests that fail to reject the null, then we cannot determine the fraction of hypotheses tested that reject. Since we should expect to reject the null five out of a hundred times even in a population with no true effect, it is clearly important to know how many tests have been run. The term “file drawer problem” was coined decades ago (Rosenthal 1979), and the idea was well known even before that.

New research documents that a large share of analyses across the social sciences that are conducted are never published or even written up, and the likelihood that a finding is shared with the research community falls sharply for findings that are not statistically significant (Franco et al 2014). Franco et al (2014) are uniquely able to look inside “the file drawer” through their access to the universe of studies that passed peer review and were able to utilize a nationally representative social science survey (funded by the NSF). This finding has potentially severe implications for our understanding of the core findings in whole bodies of research. It also implies that the ability to identify which studies were initiated but never completed could be extremely valuable in drawing conclusions in a particular research literature.

Consistent with these findings, new analyses document how widespread publication bias is in economics (Brodeur et al 2013), as well as in related social science fields including political science (Gerber et al 2001, Gerber and Malhotra 2008a), sociology (Gerber and Malhotra 2008b), and psychology (Simmons et al 2011), and in clinical research (Easterbrook et al 1991), as assessed by the “spikes” in p-values observed among published studies just below the traditional significant level of 0.05. These patterns are not likely to occur by chance (Simonsohn et al 2014), and in fact are likely to indicate some combination of selective editor (and referee) decision-making, the file-drawer problem alluded to above, and/or widespread specification searching.

While the growing use of extra robustness checks is designed to limit the extent of specification searching, it is unclear how effective they are in practice. One area of flexibility in analysis that may be particularly important is subgroup analysis. There has been extensive work on this issue within medical research (Schulz and Grimes 2005), where the use of non-prespecified subgroup analysis is frowned upon, and the FDA and NIH specifically disallow evidence based on subgroup analysis. Once again, knowledge of which studies were initiated to study which exact hypotheses would be useful in understanding how the analysis might have changed during the course of the project, possibly due to cherry-picking of statistically significant results.

A number of methods and tools that have emerged in Economics research over the past two decades – and more forcefully over the past 10 years – to address these concerns. These approaches have in common a focus on greater transparency and openness in the research process. They include improved research design – including the experimental designs that are the focus of this handbook – meta-analysis approaches, strengthened disclosure and reporting practices, new norms regarding open data and materials, and most importantly for our purposes, study registration and pre-analysis plans. We will mainly focus on this last issue, although we briefly survey the others.

There have been a number of different responses within Economics to the growing view that pervasive specification searching and publication bias was affecting the credibility of empirical literatures. Arguably the most influential response has been a shift towards a greater focus on prospective research design in applied Economics work (LaLonde 1986, Duflo et al 2007, Angrist and Pischke 2010). Experimental and quasi-experimental research designs arguably place more constraints on researchers relative to earlier empirical approaches, since there are natural ways to present data using these designs that researchers are compelled to present. There is also recent evidence that the adoption of these empirical approaches is beginning to address the concerns about specification search and publication bias mentioned above: Brodeur et al (2013) and Vivalt (2014) both find that the familiar spike in p-values just below the 0.05 level largely disappears in randomized control trial studies, but is evident for studies utilizing non-experimental methods. However, improved research design alone may not solve several other threats to the credibility of empirical Economics literatures, including the possibility that null or “uninteresting” findings never become known within the research community, as shown in the recent Franco et al (2014) article.

Economists are also increasingly adopting meta-analysis techniques originally developed in other fields, including approaches that help account for publication bias (Hedges 1992, Hedges et al 1996, McEwan 2014, Vivalt 2014), in order to clarify what is known in bodies of literature in a more systematic way, to establish priors for the next round of studies, and to better inform policymakers. Other applied econometricians have called for increasing use of multiple testing corrections in order to generate more meaningful inference, and reduce the risk of reaching erroneous conclusions, in settings where many hypotheses are being tested (Anderson 2008).

A leading proposed solution to the problem of publication bias is the registration of empirical studies in public registry. This would ideally be a centralized database of all attempts to conduct research on a certain question, irrespective of the nature of the results, and such that even null (not significant) findings are not lost to the research community. The most high profile attempt at a registry within Economics, and indeed, across the social sciences, is the new AEA Randomized Trial Registry (Katz et al 2013). The registry was launched in May 2013 and was an outgrowth of an earlier registry established in 2009 at the MIT Jameel Poverty Act Lab (J-PAL).

The AEA registry was also explicitly inspired by existing registries for medical trials. Clinical trials began being registered in large numbers in the 1990s, but the proportion registered has increased dramatically since roughly 2005, when more stringent requirements were placed on medical researchers seeking to publish in leading medical journals, as well as by government medical regulatory authorities. While recent research in medicine finds that the registry has not eliminated all under-reporting of null results or other forms of publication bias and specification searching (Laine et al 2007, Mathieu et al 2009), they do over time help constrain inappropriate practices, and at a minimum the existence of a registry allows the research community to quantify the extent of these problems. It also helps scholars locate studies that are delayed in publication, or are never published, helping to fill in gaps in the literature and thus resolving some of the problems of "disappearing" null results identified in Franco et al (2014).

Though it is too soon after the adoption of the AEA's trial registry to measure the impact, the registry is being used by many empirical researchers: in its first year and half, over 300 studies conducted in 59 countries had been registered, and the number continues rising each month. In addition to the AEA registry, several other registries have recently been created across the social sciences, although they have received fewer studies and less attention so far. These include registries created by the International Initiative for Impact Evaluation (3ie) for international development studies (the Registry for International Development Impact Evaluations, RIDIE, launched in September 2013), and the Experiments in Governance and Politics (EGAP) registry, also created in 2013.

Parallel to the trend in the pre-registration of studies, support has grown for including pre-analysis plans (PAP's) that can be posted and time stamped even before data are collected or are otherwise available for analysis in prospective studies (Miguel et al 2014). While there were scattered earlier cases of pre-analysis plans being utilized in the social sciences (most notably Neumark 2001), the numbers of published papers using pre-specified analysis has grown rapidly in the past few years, mirroring the rise of studies on the AEA registry.

Some of these early uses of pre-analysis plans in Economics are in health economics, most notably the influential Oregon Health Insurance experiment studied in Finkelstein et al (2012). This is not unexpected given how widespread pre-registration of studies and analysis plans has become within medical research. However, most economics studies using pre-analysis plans have been within development economics (see Casey et al 2012 and Olken et al 2014 among others). Casey et al (2012) show how the lack of a pre-analysis plan might have provided sufficient latitude for an unscrupulous researcher to report a wide range of different – and erroneous – conclusions using the same data, heightening concern about the possible extent of specification searching and biased reporting even in studies using randomized experimental designs – though, to follow-up on Leamer (1983)'s pun, it is our hope that there are no "con" in randomized controlled trials.

There remain many open questions about whether, when, and how pre-analysis plans could and should be used in Economics research, with open debates about how useful they are in different subfields of the discipline; some of these are the subject of a forthcoming series of papers in the Journal of Economic Perspectives (by Olken, Coffman and Niederle, and others). Yet even among these authors, who are critical about widespread adoption of pre-analysis plans in all cases, there appears to be a growing consensus that, in certain situations – such as large-scale randomized trials that are expensive or difficult to repeat, and/or cases where a government, policymaker, or corporation has a vested interest in the outcome – pre-analysis plans can increase the credibility of reporting and analysis. This certainly appears to be the case for high-profile studies on issues of major economic policy importance in both rich country (Finkelsten et al 2012) and low income country (Casey et al 2012) settings.

There is also the question of how widely these approaches could be used (if at all) for retrospective observational studies. This issue has been extensively discussed in recent years within medical research but there is as yet no consensus in that research community (Dal-Re et al 2014, Epidemiology 2010, The Lancet 2010, Loder et al 2010). A major concern with the pre-registration of non-prospective observational studies using pre-existing data is that there is often no credible way to verify that pre-registration took place before analysis was completed, which is different than the case of prospective studies in which the data has not yet been collected or accessed. This may be an insurmountable barrier to the registration of non-experimental studies in Economics.

A frontier topic in this area is the use of pre-specified algorithms (potentially including machine learning approaches) rather than exact pre-analysis plans, to lay out future analysis in prospective studies. For instance, the exact testing procedure to choose covariates that give the most statistically precise estimate can be laid out in advance, even if those covariates are unknown (and unknowable) before the data has been collected. This approach has not yet been adopted in Economics (to our knowledge), but has begun to see use in medical trials and is being widely studies in the field of bio-statistics (van der Laan et al 2007, Sinisi et al 2007).

The increased institutionalization of new practices – including through the new AEA RCT registry, which has rapidly attracted hundreds of studies, many employing pre-analysis plans, something unheard of in Economics until a few years ago – is evidence that new norms are emerging. The Berkeley Initiative for Transparency in the Social Sciences (BITSS) is another new institution that has emerged in recent years to promote dialogue and build consensus around these new practices, and BITSS also has an active training component for the next generation of Economists (and other social scientists). There are several other recent developments, including the creation of "best practices" guidelines (by BITSS and others), the promotion of pre-analysis plans and data sharing among scholars in several research communities (notably in development economics, through the Abdul Latif Jameel Poverty Action Lab and related organizations), and discuss a set of proposed changes in norms among journals, funders, and most importantly within the scholarly community itself that could make future Economics research more accurate, credible, and reproducible; these are discussed in greater detail in Christensen and Miguel (2014).

2. Experimental estimates of the impact of health on individual productivity

There is a growing literature within Economics that uses experimental variation to estimate the impact of health status on various measures of individual productivity. In this section, we focus on the experimental studies on this topic, and largely ignore the vast observational literature on these issues, a literature that spans economics, public health, epidemiology, and medical trials.

It is useful to divide the emerging experimental health economics literature on this topic into three groups: first, those studies that directly examine the impact of improved health status on current adult labor productivity and other economically relevant outcomes; second, those studies that examine the impact of improved child health and nutrition on current educational and other outcomes; and finally, those studies that estimate longer-term persistent effects of earlier health investments on later productivity measures and other life outcomes. We consider these in turn below.

Impacts of Adult Health and Nutrition on Productivity

An important early experiment that estimated the effect of adult health status on contemporaneous measures of productivity and individual well-being is Thomas et al. (2003, 2006), the Work and Iron Status Evaluation (WISE). The intervention aimed to address iron deficiency anemia (IDA), one of the world's most widespread health and nutritional problems. IDA is well-known to lead contribute to physical weakness and lower aerobic capacity, and thus could plausibly affect individual labor productivity. Thomas et al (2003) is an exceptionally well-designed study featuring a randomized evaluation of iron supplementation (weekly supplements of 120 mg of iron) plus deworming to a large sample of over 17,000 adults in Indonesia, with ages ranging from 30 to 70 years old. Since roughly 30 percent of the sample were infected with intestinal helminths at baseline, the impact of the intervention should be interpreted as the combined effect of iron and deworming.

It is worth noting that WISE features a highly unusual study methodologically within economics – although not medical research – in that it was carried out as a double-blinded experiment, i.e., the control group received placebo pills of identical appearance. This might limit any behavioral responses to the treatment that are due to the fact that beneficiaries know they are receiving treatment. While this sort of design is considered the ideal for medical trials, it is debatable whether it constitutes a similar "gold standard" for social science research studies, where endogenous behavioral responses are often central to the theoretical framework motivating a given study. Double-blinding is possible for a relatively simple intervention, such as the iron supplementation and deworming in Thomas et al (2003, 2006), but it is also often logistically infeasible in more complicated interventions, or those in which participants themselves are called upon to make decision (for instance, in the studies of technology take-up discussed below).

Thomas et al (2003, 2006) follow-up participants in the WISE study for 6 months, and focus on the intention-to-treatment estimates of program impact. They first document that iron status does improve significantly in the treatment group, with particularly large gains among those whose baseline hemoglobin (Hb) level was particularly low (below 12 g/dL, a common cut-off for anemia). The heterogeneity in Hb gains as a function of baseline deficiency motivates an estimation strategy that is similar to a difference-in-difference-in-differences (triple difference) approach: outcomes are compared between the treatment and control groups, over time (post-treatment versus baseline), across groups that had relatively low Hb at baseline (below 12.5 g/dL) versus relatively high Hb. This approach provides more statistical power than the simple treatment versus control difference, since a large share of individuals in the treatment group, namely those with relatively high Hb at baseline, do not experience any gains in Hb as a result of the intervention, and thus would not necessarily be expected to experience any gains in productivity.

They find evidence of sizeable and statistically significant gains in a range of economic and wellbeing outcomes, with effects particularly large for males (although they are generally of the same sign for females, although smaller in magnitude). The probability that individuals are not working falls significantly by between 3 to 5 percentages point for both males and females, there is suggestive evidence that total earnings increase for males, and statistically significant gains in self-employed total earnings and hourly earnings (which is similar to a wage measure) for males, as well. There are also substantial gains in psycho-social outcomes for both genders, with males finding having less difficulty sleeping and having more energy and leisure time, and females feeling less anxious. Given the relatively low cost of iron supplementation and deworming, and authors argue that this investment could have a high economic return. Given these returns, a question remains why individuals are not already doing making these sorts of investments in iron, deworming or improved nutrition more broadly in the absence of the intervention. Taken together, the rigorous research design, large sample size, and rich set of outcome measures make the results of the WISE study some of the most convincing evidence available on the causal impact of improved adult health on contemporaneous economic productivity.

Iron deficiency anemia is a pervasive but rarely fatal health condition. Thirumurthy et al (2008) usefully consider a much more severe disease, HIV/AIDS, and estimate impacts of the introduction of anti-retroviral (ARV) treatment on individual labor productivity measures in a Kenyan site. The introduction of ARV's at the individual level is determined by a "cut-off" value of the individual CD4 count (which captures how compromised the individual immune system is), and thus the study's design exploits experimental variation, although for ethical reasons treatment was not provided in a randomized fashion across individuals. Incorporation into ARV treatment during this period (when ARV treatment was still rare in Kenya) was typically life-saving and thus any labor earnings can plausibly be considered a treatment effect relative to the counterfactual. The authors more conservatively compare earnings after treatment to those immediately before treatment, which is arguably a lower bound on true effects. Due to the relatively high cost of treatment and limited number of individuals incorporated into the sample over the study period, they compare 266 households with at least one HIV positive individual to 503 other households representative of the local population.

Thirumurthy et al (2008) estimate large and statistically significant gains of over 50% in individual labor supply up to six months after the start of ARV treatment, with large impacts on total earnings. There are also important within-household effects, as the labor supply of other individuals in the households, including children, fall after an adult begins treatment, suggesting that adult health status has important externalities for others in the household and may affect the human capital accumulation of the next generation. While perhaps not surprising given that most of the treated individuals would have passed away in the absence of ARV treatment, this study provides further evidence on the important economic consequences of large health shocks.

Impacts of Child Health and Nutrition on Education

A distinct sub-literature estimates effects of child health and nutritional investments on contemporaneous educational outcomes; Glewwe and Miguel (2008) provide a thorough review of both the experimental and non-experimental research in this area. Here we focus on a selection of the experimental studies in this area. This is actually a vast literature that crosses many disciplines, and it is beyond the scope of this survey to cover all relevant studies. We focus mainly recent studies within economics, but also discuss some related contributions from other field.

Many of the earliest randomized studies by nutritionists and other public health researchers focused on the impacts of specific nutrients that were lacking in children's diets. Studies in India and Indonesia by Soemantri, Pollitt, and Kim (1989), Soewwondo, Husaini, and Pollitt (1989), and Seshadri and Gopaldas (1989) found large and statistically significant impacts on cognitive development and school performance of iron supplementation among anemic children, but a study by Pollitt and others (1989) found no such impact in Thailand. See Nokes et al (1998) for a more complete survey of the related iron supplementation literature.

Other early studies focused on parasitic infections, especially intestinal parasites. Kvalsig, Cooppan, and Connolly (1991) examined whipworms and other parasites in South Africa and found that drug treatments had some effect on cognitive and education outcomes, but some impacts were not statistically significant. Nokes and others (1992) evaluated treatment for whipworms in Jamaica and concluded that some cognitive functions improved from the drug treatment, but others, particularly those related to academic performance in schools, appeared not to have changed substantially. Overall, the early experimental literature on the impact of treatment for intestinal parasites on child growth and cognition did not reach strong conclusions, as argued in the Dickson et al (2000) survey and in the more recent Cochrane review on the topic (Taylor-Robinson et al 2012). One possible reason why many of the early experimental deworming studies show limited impacts is that they carried out randomized treatment within school communities, creating the possibility that positive treatment externalities experienced by children in the control group lead to attenuated treatment effects, as discussed earlier and in Miguel and Kremer (2004). Many of these studies also have relatively small sample sizes, such as 210 children in the South African study and 103 in the Jamaican study. Other experimental studies (not reviewed here) include education interventions combined with health interventions, so the impact of the health intervention by itself cannot be credibly assessed.

Other studies have focused on general food supplementation to supply calories and protein. The most well known of these is the INCAP study (Pollitt et al 1993, Martorell et al 1993) initiated in four Guatemalan villages in 1969, two of which were randomly selected to receive a porridge (atole) high in calories and protein while the other two villages received a drink (fresco) with less calories and no protein. Follow-up studies over the next three decades appear to show sizeable effects on later cognitive outcomes from providing the atole to mothers and young children, and we discuss these in greater detail below.

A number of recent randomized experiments have also been carried out by economists on the impact of health interventions on educational outcomes. These studies also typically evaluate actual interventions carried out by real-world non-governmental organizations (NGOs) or governments, and as such the findings of these studies may be of particular interest to policymakers in less developed countries. These are in contrast to several of the studies discussed above, which were often small-scale researcher implemented interventions. Many of these evaluate school-based health or nutrition interventions which some have argued may be among the most cost-effective approaches for delivering health and nutrition services to children in less developed countries (Bundy and Guyatt 1996).

One of the earliest and best-known studies is Miguel and Kremer (2004), which evaluates a randomized program in Kenyan schools of mass treatment for intestinal worms using inexpensive deworming drugs. The study is based on a sample of 75 primary schools with a total enrollment of nearly 30,000 children, a much larger sample size than most other studies in this literature. The sampled schools were drawn from areas where there is a high prevalence of intestinal parasites among children. Worm infections – including hookworm, roundworm, whipworm and schistosomiasis – are among the most widespread diseases in less developed countries: recent studies estimate that 1.3 billion people worldwide are infected with roundworm, 1.3 billion with hookworm, 900 million with whipworm, and 200 million with schistosomiasis. Infection rates are particularly high in Sub-Saharan Africa (Bundy, et al. 1998; WHO 1993). Geohelminths – hookworm, roundworm, and whipworm – are transmitted through poor sanitation and hygiene, while schistosomiasis is acquired by bathing in infected freshwater. School-aged children typically exhibit the greatest prevalence of infection and the highest infection intensity, as well as the highest disease burden, since morbidity is related to infection intensity (Bundy 1988).

The educational impacts of deworming are considered a key issue in assessing whether the poorest countries should accord priority to deworming, but until recently research on these impacts has been inconclusive (see Dickson et al. 2000 for a survey). Indeed, earlier randomized evaluations on worms and education suffer from several important methodological shortcomings that may partially explain their weak results. Earlier studies randomized the provision of deworming treatment within schools to treatment and placebo groups, and then examine the impact of deworming on cognitive outcomes. However, the difference in educational outcomes between the treatment and placebo groups understates the actual impact of deworming if placebo group pupils also experience health gains due to local treatment externalities (due to breaking the disease transmission cycle). The earlier studies also failed to adequately address sample attrition, an important issue to the extent that deworming increases school enrollment.

The study by Miguel and Kremer finds that absenteeism in treatment schools was 25% (7 percentage points) lower than in comparison schools and that deworming increased schooling by 0.14 years per pupil treated (on average). This is a large effect given the low cost of deworming medicine; the study estimates an average cost of only US$3.50 per additional year of school participation. The finding on absenteeism does not reflect increased school attendance on the part of children who attend school only to receive deworming drugs, since drugs were provided at only two pre-announced days per year, and attendance on those two days is not counted in the attendance analysis. There is no statistically significant difference in treatment effects between female and male students.

Somewhat surprisingly, despite the reduction in absence no significant impacts were found on student performance on academic tests. It is unclear what exactly is causing this discrepancy, although one possibility is that the program led to more crowded classrooms and that this may have partially offset positive effects of deworming on learning in the treatment schools.

The schooling data in Miguel and Kremer (2004) are noteworthy from a measurement perspective. School attendance was collected at sample schools by survey enumerators on unannounced days four to five times per year, rather than relying on school registers (which are often thought to be unreliable) or on parent reports in household surveys, as done in most of the previous literature. Efforts were also made to follow children who transferred to other schools in the same Kenyan district. This yields a more detailed and reliable measure of school participation than the data available from most other studies. The Bobonis et al (2006) and Vermeersch and Kremer (2004) papers described below use similar measures of school attendance.

The authors found that child health and school participation – i.e., attendance, where dropouts are considered to have an attendance rate of zero – improved not only for treated students but also for untreated students at treatment schools (roughly a quarter of pupils in treatment schools chose not to receive the deworming medicine) and for students at nearby primary schools located within 3 kilometers. The impacts on neighboring schools appear to be due to reduced disease transmission brought about by the intervention, an epidemiological externality. Econometric identification of the cross-school treatment spillovers on the worm infection rate relies on the randomized design of the project: conditional on the total local density of primary school pupils, there is random exogenous variation in the number of local pupils assigned to deworming treatment through the program. A key finding of the paper is that failure to take these externalities (or spillovers) into account would lead to substantial underestimation of the benefits of the intervention and the cost effectiveness of deworming treatment. Miguel and Kremer (2014) document a coding error in the construction of the variables used to measure treatment externalities at distances between 3-6 km from each school; correcting this issue weakens the statistical significant of externality effects on worm infections at this distance but does not affect other results in the original paper.

Bobonis, Miguel, and Sharma (2006) conducted a randomized evaluation in India of a health program that provided iron supplementation and deworming medicine to pre-school children age 2–6 years in 200 preschools in poor urban areas of Delhi. Even though only 30% of the sampled children were found to have worm infections, 69% of children had moderate to severe anemia according to international standards. After 5 months of treatment, the authors found large weight gains and a reduction of one-fifth in absenteeism, a treatment effect similar to the estimated school participation effect in the Miguel and Kremer (2004) study in Kenyan primary schools. The authors attempted to obtain estimates after 2 years, but high sample attrition and apparently non-random enrollment of new children into the preschools complicated attempts to obtain unbiased longer term impact estimates. It also does not present data on any type of child learning, and thus is limited to examining anthropometric outcomes and school enrollment and attendance. Finally, because all children received a combined treatment of iron supplements and deworming medicine, the India study cannot distinguish between the separate impacts of these two treatments, similar to the Thomas et al (2003, 2006) studies discussed above.

Another early randomized evaluation using a similar research design is Vermeersch and Kremer (2004). Vermeersch and Kremer estimate the impact of a preschool feeding program in 50 Kenyan preschools. The daily feeding, with a protein enriched porridge, led to 30% higher preschool participation rates, and significant cognitive test score gains in schools with relatively experienced preschool teachers, although no significant cognitive gains in schools with less experienced teachers. The authors also document how the program led to large inflows of pupils into the feeding schools, suggesting that households' school choices may be sensitive to such programs. Note that this feeding program is an order of magnitude more expensive than deworming treatment or micronutrient supplement, which will greatly increase the cost-benefit ratio. Vermeersch and Kremer (2004) are also unable to distinguish between school attendance gains resulting from improved child nutrition per se versus a desire to receive food through the daily feeding program, which makes their estimates difficult to interpret relative to previous work (and a similar concern cannot be decisively ruled out in Bobonis et al (2006) with regard to the desire to receive more iron supplementation). A further limitation of Vermeersch and Kremer (2004) is the lack of anthropometric data on sample children, which limits comparability with previous studies in the literature.

Impacts of Child Health and Nutrition on Later Outcomes

A third group of studies estimates long-run impacts of child health interventions on life outcomes, where again we focus on experimental studies in development economics.

We first examine a growing number of studies estimating long-run impacts of deworming. New evidence is rapidly accumulating on the positive long-run educational and socio-economic impacts of child deworming. A key lesson of Miguel and Kremer (2004) is that traditional individual-level randomized designs will miss any spillover benefits of deworming treatment, and this could contaminate estimated treatment effects. Thus cluster randomized designs provide better evidence. Three new working papers with such cluster randomized designs estimate long-run impacts of child deworming up to 10 years after treatment; these effects on long-run life outcomes are arguably of greatest interest to public policymakers.

A main puzzle with the Miguel and Kremer (2004) Kenya deworming study is that increased school participation (primarily attendance, but also reduced dropping out) is not reflected in students' academic test scores or cognitive test scores. The authors present some cost-benefit analyses at the end of the paper that suggest that the intervention is cost-effective, but it is unclear exactly how to interpret these if the intervention does not increase learning of basic skills.

This issue is addressed in the follow-up study, Baird et al (2015), which collects information on a wide range of adult life outcomes. Baird et al. (2015) followed up the Kenya deworming beneficiaries from the Miguel and Kremer (2004) study during 2007-2009 and find large improvements in their labor market outcomes. The paper employs a conceptual framework building on the seminal health human capital model of Grossman (1972), which interprets health care as an investment that increases future endowments of healthy time. Bleakley (2010) further develops this theory, arguing that how the additional time is allocated will depend on how health improvements affect relative productivity in education and in labor. Pitt, Rosenzweig, and Hassan (2012) further note that time allocation will also depend on how the labor market values increased human capital and improved raw labor capacity, and that this in turn may vary with gender. They present evidence consistent with a model in which exogenous health gains in low-income economies tend to reinforce men's comparative advantage in occupations requiring raw labor, while leading women to obtain more education and move into more skill-intensive occupations.

Consistent with Pitt, Rosenzweig, and Hassan (2012), the Kenya deworming program increased education among women and labor supply among men, with accompanying shifts in labor market specialization. Ten years after deworming treatment, women who were eligible as girls are 25% more likely to have attended secondary school, halving the gender gap. They reallocate time from traditional agriculture into cash crops and entrepreneurship. Men who were eligible as boys stay enrolled for more years of primary school, work 17% more hours each week, spend more time in entrepreneurship, are more likely to hold manufacturing jobs, and miss one fewer meal per week. Since deworming treatment is cheap (at less than US$1 per person per year), the authors estimate a very large annualized financial internal rate of return of at least 32.2%. Many studies argue that early childhood health gains in utero or before age three have the largest impacts (for instance, Almond and Currie 2010) and some have argued that health interventions outside a narrow biological window of child development will not have major effects. This evidence suggests that health interventions among school-aged children, which are too late in life to affect cognition or height, can have long-run impacts on labor market outcomes by affecting the amount of time people spend in school or work.

There are several noteworthy methodological features of the Baird et al (2015) article. First, it remains unusual for studies to combine experimental designs and long-run 10 year follow-up longitudinal data, and in this case most individuals in the sample were between 19 to 26 years old at the follow-up. Second, the rate of attrition was quite low in the follow-up Kenya Life Panel Survey (KLPS). KLPS tracked a representative sample of approximately 7,500 respondents who were enrolled in grades 2-7 in the

Kenya deworming schools at baseline. Survey enumerators traveled throughout Kenya and Uganda to interview those who had moved out of local areas. The effective survey tracking rate in KLPS overall is 82.7%, and 84% among those still alive. These are high tracking rates for any age group over a decade, and especially for a mobile group of adolescents and young adults. Tracking rates are nearly identical and not significantly different in the treatment and control groups.

While the primary school children in the Miguel and Kremer (2004) sample were probably too old for deworming to have major impacts on brain development, and there was no evidence of such impacts, Ozier (2014) estimates cognitive gains 10 years later among children who were 0 to 2 years old when the deworming program was launched and who lived in the catchment area of a treatment school. These children were not directly treated themselves but could have benefited from the positive within-community externalities generated by mass school-based deworming. Ozier (2014) estimates average test score gains of 0.3 standard deviation units, which is equivalent to roughly half a year of schooling. This provides further strong evidence for the existence of large, positive, and statistically significant deworming externality benefits within the communities that received mass treatment.

Croke (2014) finds positive long-run educational effects of a program that dewormed a large sample of 1 to 7 year olds in Uganda, with statistically significant average test score gains of 0.2 to 0.4 standard deviation units on literacy and numeracy 7 to 8 years later. These are similar to the effect magnitudes estimated by Ozier (2014) in Kenya. The Ugandan program is one of the few studies to employ a cluster randomized design, and earlier evaluations of the program had found large short-run impacts on child weight (Alderman et al., 2006; Alderman, 2007).

The long-run impacts of the well-known INCAP experiment in Guatemala are described in Hodinott et al (2008), Maluccio et al (2009), and Behrman et al (2009). As mentioned above, INCAP provided substantial nutritional supplementation to two villages while two others served as a control, and the authors find evidence of very large and statistically significant gains in male wages of one third, improved cognitive skills among both men and women, and even positive intergenerational effects on the nutrition of beneficiaries' children up to 35 years after the original project. This is a highly unusual and exceptional data collection effort, and it provides further evidence that childhood health and nutrition gains can have large returns in terms of adult labor productivity.

The pioneering INCAP study also has some limitations. In one sense, it has a sample size of only four villages since the intervention did not vary within villages, and it is unclear if all the existing studies fully account for the intracluster correlation of respondent outcomes in their statistical analyses, thus perhaps leading them to overstate the statistical significance of their findings. Second, strictly speaking, the control group also received an intervention, the fresco drink, albeit one with a relatively small benefit compared with what was received in the treatment group. Third, within each village receipt of the atole or fresco was voluntary, which implies that those who were treated were not a random sample of the population within each village. This means that the most convincing estimation strategy may be an intention to treat analysis, rather than direct estimation of the effect of child health on education. Finally, sample attrition is a major concern in both the 1988-89 follow-up and the most recent surveys, as more than one quarter of the original sample were apparently lost by 1988-89 and roughly 40% were lost by the time of the 35 year follow-up survey.

3. Demand for health

As we just discussed, health is an input -- it matters for how productive one can be. It is also a direct component of well-being (a consumption good, in the terminology used by economists). Both of these are reasons for individuals to invest in their health and that of their children. Health investments include preventive behavior, from getting vaccinated to wearing a seatbelt to avoiding risky sexual contacts, as well as prompt treatment of illness episodes, and diagnosis and management of chronic conditions. In a standard model with no market failure, the demand for any such health input or behavior is a function of its benefits, its costs (both monetary and non-monetary), as well as the horizon over which both benefits and costs are accrued. But households in developing countries are often liquidity constrained, and they often lack information, or the education to process information, on the potential returns to various health investments. There has been a flourish of randomized experiments aimed at understanding the role of these various constraints and their implications for public health policy. In this section we review field experiments on the demand for health, sorting them not by the outcome they are looking at but by the factor they are looking at, e.g. price, non-monetary cost, etc. Because field experiments are often designed to study a number of factors at once, with a multiplicity of randomized treatment arms, our organization by factor implies that we sometimes discuss the results of a given experimental study across multiple sections.

3.1. Pricing Experiments

A number of field experiments have examined the role of prices in the adoption of health products and services. They typically do so by randomizing the price at which a household can access a product, and comparing take-up across price points, thereby tracing out the demand curve and estimating the price elasticity at different price points. The price elasticity is an important parameter because private health investments often have social externalities. Identifying that private demand is low may therefore justify government subsidies. For subsidies to not be wasteful, however, they have to strike a delicate balance: they have to minimize the likelihood that a needy person does not access the health products or services that could benefit her, while also minimizing the likelihood that the subsidy accrues to those for whom the returns to the subsidy are marginal, either because they would have invested in the product privately anyway, or because they are unlikely to make effective use of products they receive at a highly subsidized price. This is a serious concern theoretically since households that are unwilling to pay a high monetary price for a product may also be unwilling to pay the non-monetary costs associated with daily use of the product, or may not actually need the product at all. Indiscriminate subsidies would then undermine the screening or allocative effect of prices. What’s more, subsidies could also reduce the potential for psychological effects associated with paying for a product, such as a “sunk cost” effect in which people, having paid for a product, feel compelled to use it.

Below we review about half a dozen field experiments conducted over the last 15 years that have shed light on these issues. Before we go into their details, we discuss the two main methods used to elicit willingness to pay at different price points: TIOLI and BDM. TIOLI (Take-it-or-leave-it) experiments consist in randomizing the price that an individual face, and observing whether that individual purchases the product at that price or not. The BDM mechanism (named after Becker, DeGroot, and Marschak 1964) is an incentive-compatible elicitation mechanism with real stakes that can be used to elicit individual willingness to pay as follows. People are asked to state the maximum they would be willing to pay for a product, i.e. make a bid, and to put forward their bid amount. Then a price is randomly drawn from a known distribution, and those who had bid at or above the randomly drawn price have to use the money they had put forward to purchase the product at that price (they keep the balance if they were willing to pay more than the price); while those who bid below the price cannot purchase the product. This mechanism is incentive-compatible (that is, it is the dominant strategy for expected utility maximizers) since those who bid less than their true value risk failing to buy the product when the price drawn is low enough that they would in fact prefer to do so. Conversely, bidding above one's true value entails the risk of buying when the price is higher than one would actually be willing to pay. Berry et al. (2012) discuss the merits of each method in detail and compared them in field trials in Ghana. TIOLI is straightforward to implement through door-to-door experiments, voucher distribution or retail-level subsidies. Importantly, TIOLI can be done in a way that allows people time to think through and save for the product – for example by having a fixed TIOLI price in place for a while, or distributing vouchers redeemable for a number of months, as in Dupas (2009). In contrast, BDM can only elicit immediate willingness to pay (unless it is done over credit contracts). But BDM has the advantage of telling us, for each individual in the sample, what their exact willingness to pay is, whereas TIOLI only informs us on the share of the sample willing to pay at any given price point. Thus TIOLI requires much larger sample sizes; and also cannot be used to test for heterogeneity in outcomes based on individual willingness-to-pay without additional experimental features, such as a second, surprise randomization, as in Karlan and Zinman (2009), a method subsequently applied in the health sector by Ashraf et al. (2010) and Cohen and Dupas (2010), two studies we will discuss below.

While BDM has the potential to generate richer data, the quality of this data is unclear, especially in resource-constrained settings where the population whose willingness to pay is elicited has low numeracy skills. Berry et al. (2012) assess the validity of the BDM mechanism in the context of rural Ghana. They compare the demand curve for water filters obtained through BDM with that observed through TIOLI (disabling the time feature of TIOLI, i.e forcing people to decide on their TIOLI offer immediately). They find that even after shutting off the time dimension, BDM systematically under-predicts willingness to pay relative to TIOLI. The magnitude of this under-prediction is not negligible and appears to increase with price. Namely, the demand under BDM is 20 percent lower than under TIOLI at the lowest price considered (USD 1.40, a tenth of the retail price), 34% lower when the price is USD 2.80, and 45% lower when the price is USD 4.20. Berry et al. (2012) remain agnostic as to the reason why BDM is somewhat inaccurate (if we take TIOLI as reflective of the “true” demand), though through additional experimental treatments, they can rule out that the difference between the two mechanisms is driven by either strategic bidding under BDM (i.e., people stating a low willingness to pay in the hope to influence future prices, in particular the likelihood of a NGO subsidy being introduced) or anchoring under TIOLI (the TIOLI price influencing people's willingness to pay). More work is needed to understand when and how BDM can provide accurate estimates of willingness to pay. In the meantime, most pricing experiments have been using TIOLI.

One of the earliest randomized TIOLI experiment for a health product in a poor country took place in 2001 in Western Kenya. Among 50 primary schools enrolled in a free deworming program in 2000, Kremer and Miguel (2007) randomly selected 25 that moved to a *cost-sharing* program: parents now had to contribute a fee in order for their children to receive the deworming pill(s) on deworming day. Parents had to pay the fee at the school in advance of the deworming day, and were informed of this fee XX how, how long before deworming day? Paper doesn't mention that XX. The researchers found that the share of children receiving deworming medication on the day the NGO visited the school for mass deworming was only 18 percent in the cost-sharing schools, compared to 75% in the schools who kept the free program, despite the fee charged per child being just about 20% of the actual program cost on average. Interestingly, parents of sicker pupils were no more likely to pay for deworming drugs, suggestive no screening effect of the cost-sharing program. While these results suggest that demand is highly sensitive to price, understanding why it is the case in this specific context is somewhat difficult. It could be that parents had gotten "used to" the free program and resented the introduction of the cost-sharing fee, and therefore their demand was lower than what it would have been had no free program ever been implemented in the first place. It could also be, and the authors hypothesize it is the case, that the perceived private value of deworming is lower than the fee charged. Subsequent pricing experiments adopted more nuanced designs in order to disentangle these mechanisms from each other.

Cohen and Dupas (2010) use a two-level randomized TIOLI design to estimate (1) the demand curve for a new health product in rural Kenya: long-lasting antimalarial bed nets (LLINs); and (2) the distinct roles of the screening and psychological sunk cost effects that price may have on their usage. LLINs cost $7, and they prevent bites from malaria-carrying mosquitos while sleeping. The experiment, conducted in 2007, randomized the price at which prenatal clinics offered nets to pregnant women. Clinics charged either 0 (free distribution), 15, 30 or 60 US cents (note that the highest price point considered, 60 US cents, still represents a 92% subsidy). This first level of randomization, at the clinic level, involved only 20 observations (20 clinics), something which has implications for inference, as discussed in Cohen and Dupas (2010). The second level of randomization was at the individual level. Namely, a random subset of women who had agreed to purchase the net for 30 or 60 cents were subsequently given a surprise rebate right after they had given their payment to the clinic's cashier. Cohen and Dupas (2010) find that demand is very sensitive to price: the likelihood that pregnant women acquired a net fell from 99% to 39% when price increased from 0 to 60 US cents (with the demand at the intermediate price points of 15 and 30 US cents at 92% and 72%, respectively). This suggests that while there is no discontinuity at zero (it's not the move from free to any positive price that makes the demand drop, but larger price increases), demand is on the whole quite price sensitive, with very low demand rates at prices that are still very heavily subsidized. Interestingly, however, they found that the rate at which pregnant women used the net (measured through home observation visits two months after distribution) was relatively high (60%); and it was completely independent of the price they had paid for the net, whether initially or after the surprise rebate. This suggests that there is neither a screening nor a sunk cost effect of prices in their context. Thus coverage (the share of pregnant women sleeping under a bed net), and hence its potential for public health outcomes, increases very rapidly as the price goes down.

In another TIOLI experiment conducted with a sample of households with school-aged children, also in Kenya, Dupas (2014a) found that demand becomes slightly less price sensitive if subsidies are provided in the form of vouchers that households have three months to redeem at local retail shops: the demand at $0.60 becomes 73%. But overall price remains the primary driver of the demand, with the purchase rate dropping to just around 33% when the price reaches $1.50 (still an 80% subsidy) and to 6% when the price reaches $3.5 (corresponding to a 50% subsidy). Various marketing strategies (e.g., making the morbidity burden or treatment costs salient, targeting mothers, eliciting verbal commitments to invest in the product) failed to change the slope of the demand curve (Dupas, 2010). But here again, the price paid did not matter for usage. In fact, home observation visits showed that usage of bed nets acquired through a subsidized voucher was extremely high, rising from 60% at a three-month follow-up to over 90% after one year, and that was the case across all price groups, including recipients of fully subsidized nets.

The finding that demand is price sensitive has been established by TIOLI experiments for products other than deworming drugs and bed nets. In 2010, Meredith et al. (2013) randomized the subsidy level households faced for rubber shoes to prevent worm infections in Kenya, and in 2008 they randomized the price of soap and vitamins in Uganda, Guatemala, and India. In all context they found that demand was very sensitive to price. In a TIOLI experiment in urban Zambia, Ashraf, Berry and Shapiro (2010), studying the demand for a bottle of water purifying solution (diluted chlorine), also find that demand is sensitive to price, dropping from around 80 percent at the price point of 9 US cents (a 62.5% subsidy) to only about 50 percent at the full market price of 25 US cents. That experiment also used a two-stage randomization design but in this case both randomizations took place at the household level. Namely, not just the surprise rebate but also the initial offer price was randomized across households. With this design, they test for both the screening and sunk cost fallacy effect of prices. As in Cohen and Dupas (2010), they found no evidence of a use-inducing sunk-cost effect, but found some evidence of a screening effect of prices. Specifically, those households who had selected themselves into paying a higher price were more likely to have used the purification solution within 6 weeks of acquiring it, while those who had received a higher subsidy were more likely to still have it on their shelf, possibly because they were keeping it for later or, as per the authors' interpretation, because they were less likely to ever plan to use it for a health purpose.

The studies discussed above suggest that price is often not a good mechanism to target subsidies for prevention tools to those who need them the most. If anything, higher prices seem to create too many errors of exclusion, and to prevent the positive spillovers on disease transmission that justify subsidies in the first place. The evidence regarding treatment products is somewhat different however. Cohen, Dupas and Schaner (2015), in a TIOLI experiment conducted in 2009 in the same area of Kenya as the bed net studies mentioned above, find that price can be (to some extent) used as a targeting mechanism to allocate malaria treatment. Targeting of malaria treatment is very important because of the negative spillovers that overuse of antimalarials generates: it can delay or preclude proper treatment for the true cause of illness, waste scarce resources for malaria control, and may contribute to drug resistance among malaria parasites, making treatment of malaria harder in the long-run. The reason why, within essentially the same population, price can be effective at targeting treatment when it's not effective at targeting prevention is that demand for treatment appears much less price-sensitive (especially among the poor) than demand for prevention. What's more, conditional on experiencing malaria-type symptoms adults are much less likely to be malaria positive than children, but as with most treatments, the price per antimalarial dose for adults (who need to take more pills) is higher than the price for children. Consequently, at a given price per pill, children (the key target for the subsidy) are on a flatter portion of the demand curve. In addition to furthering our understanding of how price can be used to target health products in the developing world, that fourth study makes two other contributions: (1) it highlights the trade-off inherent to subsidies for medications in environments with weak health system governance (which prevents conditioning the subsidy on a formal diagnostic), and (2) it points out that bundling subsidies for medications with subsidies for diagnostic tests has the potential to improve welfare impacts.

While the studies above have mostly focused on the effect of price on contemporaneous demand, some field experiments have been specifically designed to look at the dynamic effects of prices. The questions here are the following: Can a once-off subsidy be enough to trigger learning and generate sustained adoption? Or is there a risk that people are unwilling to pay for a product they once received for free, as a cursory look at the Kremer and Miguel (2007) deworming cost-sharing results could suggest? This could happen if people, when they see a product being introduced for free, come to feel entitled to receive this product for free (i.e., they would "anchor" around the subsidized price). To gauge the relative importance of these effects, Dupas (2014a) looks at the long-run effects of the one-time bed net subsidy vouchers households received in the study mentioned above. Specifically, the research team came back a year after the first pricing experiment had been done, and implemented a second stage in the study in which all households received a second subsidy voucher, but this time they all faced the same price of $2.30 (a 70% subsidy). By observing how the take-up rate of the second, uniformly-priced bed net varies as a function of the price a household faced in the first year, Dupas (2014a) can test whether being exposed to a large or full subsidy in the first year (which, as discussed above, considerably increased adoption at that time) reduces or enhances willingness to pay for the bed net a year later. She finds that it enhances it, suggesting the presence of a positive learning effect which dominates any potential anchoring effect. Interestingly, the learning effect trickles down to others in the community: households facing a positive price in the first year are more likely to purchase a bed net when the density of households around them who received a free or highly subsidized bed net is greater. Though once bed net ownership is widespread, the transmission risk starts to decrease and the returns to private investments decrease, and accordingly those with more subsidized neighbors in year 1 were less likely to invest in year 2. Dupas (2014a) also tests for cross-product effects of price subsidies, namely, whether getting a subsidy for the ITN led households to expect subsidies for another health product, namely, a water purification product. She finds no such effect: willingness to pay for the water product a few months after being exposed to a subsidy for the ITN was not lower among households who had received a full or very high subsidy.

Karlan et al. (2014) adopted a similar design to study the relative importance of learning vs. anchoring effects in free distribution programs for a wider set of products, as well as the importance of who implements the free distribution. They conducted their experiment in northern Uganda. In a first round of door-to-door visits, they offer households one of three products either for free or for sale at the prevailing market price. The three products were chosen to differ in their scope for learning, and included a pain reliever widely known at baseline (thus with no scope for learning); a deworming drug that was moderately well known and which has side effects so the learning effect, as in Kremer and Miguel (2007), was expected to be negative; and a new, largely unknown treatment for childhood diarrhea for which the authors expected positive learning. The second door-to-door visits took place two to three months later and were conducted by a different set of sales agents. Households were offered the same product in round 2 as they had been offered in round 1, except for some randomly selected households who were offered a fourth product (water purification solution) to test the presence of cross-product effects. As in Dupas (2014a), the authors find no evidence of cross-product effect, but they argue that the patterns of demand they observe in round 2 is consistent with anchoring playing an important role: when there is no scope for learning (the pain killer case), they find that demand is lower after free distribution than after a sale. This is the case irrespective of whether the free distribution was done by an NGO or by a for-profit firm advertising the free distribution as a standard marketing tool (a "free trial" to help people learn a product is worth their money). The results are somewhat weak however, with no significant differences in the treatment effects across products. One potential concern with the design of this study was that the authors did not collect information on the demand for the three products outside their experimental door-to-door visit, even though those products were available for a similar if not lower price outside the experiment (at local shops). This highlights a challenge in demand studies – measuring impacts often require measuring demand both within and outside the experiment. A good example of that is Cohen et al. (2015) discussed above, which measured access to ACTs through not only the drug shops involved in the study but also local health facilities, something critical in their context given the scope for crowding out. This can be difficult if recall bias is a concern, especially when the time lag between baseline and endline is long. An alternative is the design used in Dupas (2014a), who offered a product (a long-lasting ITN) unavailable on the market at the time of the experiment. Having perfect control over the supply means that observing the demand in the experiment provides Dupas (2014a) with a complete picture of the demand in both years. The drawback of that design is that when the product is not available on the market, the option value of experimenting with the product in round 1 is lower, as households were not aware they would have a second chance to obtain the product from the experimenters, thus the take-up in round 1 may have been an underestimate of the take-up that would have prevailed in a market environment.

On the whole, the pricing experiments discussed above suggest that investment in preventive health is very sensitive to price, especially when preventive health products are first introduced; and that prices are not a very effective allocation mechanism in the sense that they fail to target those who need the products the most (Dupas, 2014). Glennerster and Kremer (2011) present a framework highlighting that this price sensitivity may come from liquidity constraints, lack of information, non-monetary costs or behavioral biases such as present bias and limited attention. A number of experiments have generated randomly varying access to liquidity, convenience or information, sometimes interacted with random variation in prices, to estimate the relative role of these potential factors. The evidence to date suggest that information is necessary but not sufficient, and in particular does not appear to substitute for higher subsidies, while reducing non-monetary costs and increasing liquidity often matter a lot. We review this evidence below, before discussing their implications for the scope of behavioral factors at play.

3.2. *Liquidity experiments: cash drops and credit experiments*

In their pricing experiment, Meredith et al. (2013) gave households a randomly determined amount of cash (in the form of payouts for incentivized risk preferences elicitations) at the same time they distributed discount vouchers. The market price for the shoes was about 85 Ksh ($1.13) at the time, and discount vouchers varied from a low (20 Ksh) to a high (80 Ksh) discount. The cash drop varied from 0 to 200 Ksh, with a mean of 35. The researchers use the variation in cash drop amount to estimate the effect of liquidity on purchase. They find a large and significant effect of the cash drop on demand: on average, every additional 100 Ksh in randomized cash payout increases the probability of voucher redemption by 22 percentage points. Since the cash drop was very small relative to lifetime income, at about 4% of weekly income on average, its effect on demand reflects a cash-on-hand effect rather than an income effect. This result suggests that present bias is unlikely to explain much of the underinvestment in prevention at even moderate prices. Indeed, if present bias were at play, then we would not expect individuals to use the cash-on-hand received from the experimenter to invest in products with a delayed payoff. Importantly, since households had to travel to a local store and redeem a voucher in order to obtain the product, the cash drop effect is very unlikely to be driven by an experimenter demand effect, whereby study households would feel compelled to use the money obtained from the experimenter to purchase the product endorsed by the experimenter.

One study looks at the health impacts of providing cash transfers to adolescent girls in Malawi. Baird et al. (2012) randomly assigned 175 enumeration areas (EAs) to three groups: girls in 46 EAs received conditional cash transfers (CCTs) if they achieved 80 percent school attendance; those in 27 EAs received unconditional cash transfers (UCTs); 88 EAs served as a comparison group and did not receive transfers. The cash transfers significantly lowered the prevalence rates of HIV and the herpes simplex 2 virus (HSV2). For example, 1.2 percent of girls enrolled in school at baseline who received transfers (CCT or UCT) tested positive for HIV at 18 months relative to 3.0 percent of girls in the comparison group. Self-reported sexual behavior was also lower among girls who received transfers; 3 percent of girls who received transfers reported having sex at least once per week, compared to 7 percent in the comparison group. These results suggest that financially empowering school-aged girls can have substantial effects on their sexual and reproductive health. Interestingly, the authors find that the amount of the money transferred did not itself matter, nor the share of the transfer directly transferred to girls vs. their parents.

Another way to experiment with the liquidity constraint faced by households is to allow them to purchase health products on credit. While researchers that have exploited the random introduction of microcredit have found no impact on health expenditure (ADD REF TO AEJ APPLIED SPECIAL ISSUE), this may due to the coarseness of their data on health investments, and/or to the fact that most microfinance institutions focus on business loans rather than consumption loans, and that flypaper effects are common (Fafchamps et al. 2014). The first studies to directly study demand for health products at full price when credit constraints are relaxed are Devoto et al. (2012) and Tarozzi et al. (2014). Devoto et al. (2012) identified low-income households not connected to the water grid in the city of Tangiers in northern Morocco, and randomized which households were told about a credit program to purchase a water connection. They see an impressive take-up rate of 69 percent, despite the fact that the cost of the connection (which varies with distance to the water mains) averages around $1,000, an amount that they would have to repay over five years. Tarozzi et al. (2014) randomized access to ITNs on credit across villages in the state of Orissa, India. They find that 52% of households offered full-price ITNs on credit purchased at least one ITN (and all of them fully repaid the loan). In contrast, in a follow-up cash sales study, they find that only around 11% of households purchase at least one ITN absent any credit.

3.3. *Information Experiments*

Even when liquidity constraints are alleviated, adoption of high-return health products or behaviors is often not 100%. A potential explanation for this could be that individuals lack of information on the health costs or benefits of different products or behaviors. In this section we review information experiments showing that (exhaustive) information is necessary but often not sufficient.

In their anti-worm rubber-sole shoes pricing experiment in Kenya, Meredith et al. (2013) find that health workshops did not affect total demand nor the price gradient in demand. In their complementary evidence from India, Guatemala and Uganda, they find an effect of the health information in only one of 6 country-product combinations they experimented with – namely, a health script delivered at the time households could obtain subsidized soap in India flattened the effect of price on demand. The authors discuss that this one result may be driven by an experimenter demand effect, however, since in this specific case the purchase decision was contemporaneous to the cash drop.

Ashraf, Jack and Kamenica (2013) also interact subsidy level and information provision, but the information provided concerns the relative merits of a product over another, rather than absolute information about health. Specifically, in a door-to-door marketing campaign in urban Zambia, sampled households were offered the option to buy one of two water purification products, a product well known in the area and available at retail stores (called "Clorin") and a similar product from another brand, which people had never seen before and which was not available at any local stores. The price of the familiar product was fixed at 800 Kwacha, the standard retail price. The price of the unfamiliar product was randomized across households, and varied from 0 to 1200 Kwacha. In addition to the price randomization, the information given about the unfamiliar product was randomly varied. Half of households were provided no information. The other half were told the unfamiliar product is similar and as effective as the familiar product. Ashraf, Jack and Kamenica (2013) find no overall impact of the information treatment on the demand for the new product, or on total demand. However, the demand curve for the unfamiliar product becomes steeper when information is provided, and it pivots exactly around the price at which the familiar product is available. Demand for the familiar product (as a function of the price of the unfamiliar product) pivots the opposite way, and total demand for water purification products doesn't significantly increase in the presence of information. This apparent complementarity between information and subsidies can be interpreted as follows: in the absence of any information, people tend to take the price of the unfamiliar product as a signal of its quality, so they are not completely turned off by high prices, while they are somewhat turned off by low prices. When information is provided, the signaling content of the price diminishes. As a result, people are less likely to be turned off by low prices, and more likely to be turned off by high prices (in particular, there is now no reason why they'd pay more for the unfamiliar product than the price of the familiar product since the information reveals that the two products are the same). The effect of the information is thus to encourage more people to switch from the familiar product to the unfamiliar product at low prices, and to deter more people to do the switch at high prices.

While the studies above look at the impact of information on the willingness to pay for specific products, a number of experiments have studied the impact of information on health behavior. In Kenya, as part of the pricing experiment for deworming medication discussed above, Kremer and Miguel (2007) randomly varied whether schoolchildren received information on how to avoid intestinal worm infections. The information was provided in the classroom by a mixture of trained teachers and NGO staff, and focused on preventative behaviors such as washing hands, wearing shoes, and avoiding infected fresh water. One year later, data on pupil cleanliness and shoe wearing (as observed by the research team) as well as self-reported data on exposure to fresh water showed no effect of the education campaign.

Also in Kenya, Duflo, Dupas and Kremer (2014) and Dupas (2011b) examined the impact of providing different types of HIV/AIDS information to primary school students. In a randomly selected subset of 328 schools, teachers were trained on how to implement the national HIV/AIDS curriculum, which focuses on abstinence as the only prevention method available for adolescents. Duflo, Dupas and Kremer (2014) find that the training greatly increased the likelihood that teachers teach about HIV in the classroom, and two years after the training students whose teachers had been trained had greater knowledge about the disease. The intervention did not reduce childbearing rates among girls, however, suggesting that it did not decrease the likelihood that girls engage in unprotected sex. It also did not reduce the risk of STI as measured after 6-7 years. Within the 328 schools, Dupas (2011) randomly selected a separate 71 schools to receive an information session that discussed the role of cross-generational sex in the spread of HIV, and the relative risk of HIV infection by gender and partner's age. In many African countries, HIV prevalence increases with age among men. Therefore sex with older partners, which in many cases occurs in relationships with so-called sugar daddies, substantially increases the risk of HIV infection for adolescent girls. This information was provided by a trained facilitator, introducing herself as working for a local NGO, to upper grade students in the selected 71 schools. This "relative risk" information intervention, which provided adolescents with information on how to reduce their exposure to HIV conditional on being sexually active rather than only exhorting them to abstain, led to a 28 percent decrease in teen pregnancy among school-going adolescent girls, and was driven by a reduction in cross-generational sex (with male partners five or more years older). Together, the results of these two experiments suggest that providing specific information is more effective at changing behavior than general exhortation.

In Malawi, Godlonton et al. (2014) estimate the impact of an information campaign about the relationship between circumcision and HIV status. They study the impact on men who are not circumcised at baseline, as well as those who are. For the former group, they find that the information increased correct knowledge about relative risk, reduced risky sexual activity, and increased condom use. Specifically, uncircumcised men in the treatment group were 25 percent less likely to have sex each month, and 58 percent more likely to use a condom. For the latter group, which learned they were better protected, there was no evidence of riskier sexual behavior. While uncircumcised men reported an increased willingness to have their male descendants circumcised, overall take-up of adult male circumcision was low. Researchers also found that the circumcision information campaign, though it increased correct understanding about how male circumcision can partially protect males against HIV transmission, also increased the incorrect understanding among participants that male circumcision protects females against infection as well (which it does not.) These results suggests that information alone is not enough to increase the demand for male circumcision, and that one has to be careful in the way information is delivered to mitigate against incorrect learning.

Also in Malawi, Chinkhumba et al. (2014) conducted a randomized experiment to study the impact of information and price on the demand for medical male circumcision. 1,634 men were given vouchers for a subsidized circumcision at a nearby clinic; the researchers randomly assigned different values to the vouchers, with subsidies ranging from 8% to 100% (i.e. free) of the full price. The study randomly selected half of the men to receive comprehensive information about the biological relationship between male circumcision and HIV risk. Results were collected through both self-reports and clinic records. Information increased the number of circumcisions by 66 percent (1.4 percentage points), but overall the rate of circumcision was extremely low: no one offered the full price was circumcised, and only 3.1% of those offered a free circumcision elected to take up the procedure.

Rather than experimenting with the *content* of the information provided, two recent studies have experimented with the *delivery methods* for HIV information. Indeed, recent advances in communication technology mean that information does not need to be delivered in-person by either a teacher or an outside facilitator. In Colombia, Chong et al. (2013) looked at the impact of an online sexual health education course provided through schools. Researchers partnered with Profamilia, a large Colombian NGO, to randomly provide a course to one third of 138 ninth-grade classrooms from 69 public schools in 21 cities. One third of the classrooms were randomly assigned to the comparison group, which did not receive the program, while the remaining one third of classrooms did not participate in the course but were located in the same schools as the classrooms that did receive it. The course increased overall sexual health knowledge by 0.38 standard deviations, and increased positive attitudes towards condom use. There was no impact on self-reported sexual behaviors, but there was a reduction of 5.2 percentage points (83 percent) in self-reported sexually transmitted infections among females who were already sexually active before the program, suggesting that some students adopted safer sex practices. The reliance of the study on self-reported sexual behavior is somewhat problematic however.

In Uganda, Jamison et al. (2013) tested the impact of increasing access to information about sexual and reproductive health for the general population (not just schoolchildren) via a text messaging service about risky sexual behavior. Among 60 villages, marketing teams encouraged individuals in a random subset to use a new mobile phone-based information system through which users could send questions and receive responses on sexual and reproductive health. Usage among these villages was fairly high at 40 percent, but the service had no impact on villagers' sexual or reproductive health knowledge. The intervention led to an overall higher incidence of risky sexual behavior and self-reported promiscuity, particularly among men, while women reported increased abstinence. Qualitative information sheds some light on the potential causes of this mixed impact; men and women reported that married women who learned about the risks associated with having an unfaithful partner insisted their husbands be faithful and get STI tested. According to these reports, some husbands did not comply, leading women to deny them sex and men to seek it from other partners instead. Overall, individuals in treatment villages perceived their sexual behavior to be riskier, which could indicate an actual increase in risky behaviors, or could indicate that the information service increased accurate assessment of health risks. Unfortunately the researchers here again do not have objective measures of the risk level, such as biomarkers of sexually transmitted infections or pregnancy, to tease out these two potential explanations.

All the studies above concern generic information. In contrast, Prina and Royer (2014) study the impact of providing tailored (individual-specific) information – namely, the impact of providing parents with body weight report cards for their schoolchildren. The report cards included information on a child's height and weight as well as the weight classification (i.e., underweight, healthy weight, overweight or obese). This increased parental knowledge and shifted parental attitudes about children's weight, but did not lead to meaningful changes in parental behaviors or children's body mass index, even when the body weight information was accompanied by information on the health risk of obesity. The authors provide evidence that social norms matter: if the report card included information on the distribution of weights in the classroom, then the larger the fraction of overweight children in the child's class, the less likely a parent was to report that her overweight child weighed too much. As the authors note, this implies that as obesity rates increase, programs aimed at reducing obesity may become less and less successful as reference points for appropriate body weights may rise.

A final question regarding the role of information on health behavior concerns who the information should be targeted at. Ashraf and Field (2014) explore this question in the context for the demand for family planning in Zambia. Women in the study received vouchers that granted appointments with a family planning nurse at the local government clinic. Information explaining all methods of family planning, including "concealable" methods such as injectables, was provided along with the vouchers. Women were randomized into two treatment groups. In the "individual" arm of the study, women were given these vouchers alone. In the "couples" arm, women were given these vouchers in the presence of their husbands. In all other respects, the experimental protocol in the individual and couples arms was identical. The results are fascinating: take-up of the voucher was significantly lower when information about family planning services was provided in the presence of husbands. Women who received the voucher in the presence of their husbands were 9 percentage points (18 percent) less likely to use the voucher to obtain an appointment at a family planning clinic. This gap was larger (12 percentage points) among couples with divergent fertility preferences (in particular, where the husband reported wanting more children than the wife).

3.4. Non-monetary Costs Experiments

For some products or services, demand remains low even at very low or even zero prices. For example, in the poor district of Udaipur in India, despite the fact that immunization services were offered free in public health facilities, Banerjee et al. (2011) estimate that only 2% of children aged between 1 and 2 had received the recommended basic package of immunizations. As discussed above, in Malawi, only 3% of uncircumcised adult males who received a voucher for a free circumcision at the local clinic underwent the surgery (Chinkhumba et al. 2014). This could be because of non-monetary costs associated with take-up, such as time costs (e.g. walking 60 minutes to reach the facility where the free service is available), hassle costs (having to fill complicated paperwork in order to receive a subsidy), and cultural barriers (for males thinking of getting circumcised as a means of reducing vulnerability to HIV). Some of these non-monetary costs can be experimentally varied (in particular, distance and convenience). Some (e.g cultural costs) cannot be experimented on, but their relative importance can sometimes be backed out by experimenting with financial incentives – the idea is that if a small financial incentive is successful at increasing adoption, then cultural barriers cannot be too important.

Back to the example from Udaipur, India, Banerjee et al. (2010) run an experiment to test the hypothesis that the reliability of the supply of free services may be at fault. The premise for this hypothesis is the observation made by Banerjee, Deaton and Duflo (2004), and discussed earlier in the chapter, that public facilities in charge of providing free immunization are characterized by very high absenteeism: spot checks conducted over a year suggested that 45 percent of the health staff were absent from their health posts (typically leading to the health post being closed) on any given workday. Because there was no predictable pattern to this absenteeism, obtaining all five shots included in the basic immunization package could require twice as many attempts at visiting the public health facility. In the experiment, some villages served as controls and other were randomly selected to receive a reliable, well-advertised "immunization camp". The researchers found that the adding a reliable camp boosted full immunization rates from 6% to almost 18%. Impressive as the tripling of take-up in the immunization camps experiment may be, at only 18 percent fully immunized the take-up remained among the lowest in the world. Could that be due to cultural barriers? In the experiment, a third group of villages was randomly selected to receive, in addition to the immunization camps, incentives for parents – parents were given a kilogram of lentils per immunization, and a set of plates for a child fully immunized. The incentive treatment increased immunization rates from 18 percent to 39 percent – which suggest that cultural barriers may not be at play since they can be overcome with a fairly small handout. So what is the main barrier? We discuss the potential interpretations of this and other incentives experiment in the next section.

Thornton (2008) conducted a field experiment in rural Malawi that randomized the distance that individuals had to travel in order to obtain results of an HIV test, as well as whether they received a financial incentive to seek their results. This field experiment took place in 2005, at a time that preceded the introduction of rapid HIV tests, thus people had to do two visits to the testing center in order to learn their status – a first visit to get their blood drawn and a second visit a few weeks later to fetch their result. At the time Thornton conducted her experiment, the prevailing conventional wisdom in HIV prevention circles was that demand for knowing one's status was very low due to the high psychic costs of learning one was HIV positive: since there was no access to antiretroviral therapy (ARV) in Malawi at the time, learning one's positivity was akin to a death sentence. If psychic costs were indeed high, then providing a financial incentive and reducing the time costs of fetching one's results to a minimal level would likely have only a small effect on the demand for test results. Thornton (2008) found the exact opposite: providing a financial incentive increased the likelihood that individuals sought their HIV test results from 35% to 78%. Reducing the distance that one had to travel to get results also increased the share of individuals seeking their results. Absent any incentives, those living within 1.5 kilometers from the center where results could be picked up were 6.4 percentage points less likely to seek their HIV results than those living more than 1.5 kilometers away. These large impacts thus teach us two things: that distance matters, i.e. time costs are not to be neglected; and that psychic costs were, on the other hand, much less important than believed at the time.

In their study of the demand for contraceptives in Zambia, the vouchers that Ashraf and Field (2014) distributed granted appointments for ordinary family planning services (provided routinely at government clinics), but with a guarantee that the wait would be less than one hour, and that the modern contraceptive method of their choice would be available. Take up of the voucher was high (47 percent), indicating as in Banerjee et al. (2011) that unreliable supply and its associated substantial time costs may be important barriers to take-up of health services.

The fact that non-monetary costs matter can sometimes be used to improve targeting. Recall the results from the pricing experiments discussed in section 3.1, which on the whole suggested that in environments where people face serious liquidity constraints, as in most of the developing world, price is not a particularly good screening mechanism – it fails at allocating scarce products to those who have higher returns to these products -- as many people with a high valuation for a product may not be able to afford it. On the other hand, under free distribution, the product may be wasted on people with a low valuation for the product, and for products with a high share of low-valuation people, that may be very costly. Imagine that a provider delivers a year's supply of water-treatment product to a household, and the household members learn within a few days that they hate the taste of chlorinated water and stop using the product. In such cases, where households need to learn their own valuation, imposing some non-monetary cost that households have to pay to access the free product may be efficient. This is what the literature refers to as an "ordeal mechanism". The provider may, for example, require that those who want a year's supply go to a store to redeem coupons every month for 12 months. A field experiment conducted in Western Kenya in 2007-2008 suggests that such a micro-ordeal can help target free products only to those who will use them. Dupas et al. (2013) provided households with the opportunity to obtain enough free samples of chlorine solution for the treatment of drinking water for a whole year, but varied the effort required to obtain the samples. Households randomly allocated to treatment arm 1 received a free supply of chlorine delivered directly at their home, while households randomly allocated to treatment arm 2 were given 12 coupons which could be redeemed for chlorine at a local shop over the course of a year. The researchers compare chlorine usage across arms and find no difference in rates of usage, during the year that followed the distribution, between those who were required to redeem coupons, compared to those who were given chlorine directly. They estimate that under reasonable assumptions regarding distribution costs, the results imply a significant increase in cost-effectiveness, with no negative impact on usage, of imposing a non-monetary price on the acquisition of a health good such as chlorine, which is not valued equally by all households. Rozelle et al (2014) perform a similar experiment looking at the demand and usage for eyeglasses for schoolchildren in China.

3.5. Incentive Experiments

We briefly mentioned earlier two experiments that had incentives as one of their arms: the Udaipur immunization study (Banerjee et al., 2011) and the HIV testing study (Thornton 2008). In both cases, small financial incentives were provided to encourage take-up of a specific health behavior: immunizing children, and learning one's HIV status. In both cases, the small financial incentives had a large impact on take-up. Banerjee et al. (2011) interpret this as evidence of present bias -- the natural tendency to delay an action that is slightly costly today even if it has high payoffs in the future, a tendency which can be overcome if the incentive is sufficient to transform the cost into a positive *immediate* benefit.

While these two studies consider incentives rewarding a specific behavior (a specific input in the health production function), more recent studies have looked at the impact of output-based incentives, namely, incentives paid based on achieving a certain health outcome. de Walque et al. (2012) looked at the impact of offering varying amounts of a cash incentive to remain STI-free among adults aged 18-30 years old in Tanzania. They randomly assigned 2,409 individuals to one of three groups: 1) a "high-value" conditional cash transfer (CCT) group that received $20 for testing negative for curable STIs; 2) a "low-value" CCT group that received $10 for testing negative for curable STIs; and 3) a comparison group that received no transfer. STI tests were conducted for all groups every four months for one year. Over the course of the first year, the number of people who tested positive for STI infection decreased among people who received the high-value CCT, but no reduction was found for the group that received the low-value CCT. One downside of this study is that it was not powered to detect effect on HIV infection.

In a similar study in Malawi, Kohler and Thornton (2012) randomly gave cash transfers of random amounts to 1,307 participants, which ranged from no cash to approximately US$16, conditional on maintaining one's HIV status for one year. Researchers conducted interviews with participants throughout the year to collect data on sexual behavior. The promise of financial incentives of any amount had no effect on subsequent self-reported sexual behavior or HIV status. However, receiving cash after the final round of HIV testing had significant effects on respondents' self-reported behavior; men were 9 percentage points more likely to engage in riskier sex, and women were 6.7 percentage points less likely to do so. As in the case of the incentivized immunization experiment discussed above, these results suggest that money given in the present may have stronger effects on behavior than rewards in the future – but sometimes for worse.

3.6. *Marketing Experiments*

As briefly mentioned above, in an ITN pricing experiment, Dupas (2009) evaluated the effects of two marketing interventions based on behavioral models derived from psychology: varying the framing of the perceived benefits; and having individuals verbally commit to purchase the product. At the time they received their first voucher, households were exposed to a randomly assigned marketing message. The "health framing" group emphasized the morbidity and mortality due to malaria which could be avoided by using the net. The "financial framing" group emphasized the financial gains households would realize (from averting medical costs and loss of daily income) if they could prevent malaria. A third group received no marketing message. Finally, a randomly selected half of all the households were asked to verbally commit to buy the ITN, and state who would sleep under it once they had bought it. Neither of the two framing options (health or financial) had any impact on LLIN take up, and women to do not appear to have a different price elasticity than men. Likewise, the verbal commitment treatment had no impact on actual investment behavior, despite a 92 percent initial agreement to purchase the LLIN. Kremer and Miguel (2007) similarly had found no impact of a verbal commitment intervention embedded within their deworming take-up experiment in Kenya.

4. Health Care Delivery

Most of the experiments described above on the demand side concern preventive health demand. This is because demand for acute care is quite high – as discussed previously in Dupas (2011a) and Glennerster and Kremer (2011). While access to quality health services is an important determinant of health outcomes (FIND REF, MAYBE COCHRANE REVIEW IN PUBLIC HEALTH?), a growing body of evidence documents important gaps in both access and quality in the delivery of health services in developing countries, especially for the poor (World Bank 2004; see Das and Hammer 2014 for a review). Major issues identified to date concern absenteeism among public health providers (Chaudhury et al. 2006, Banerjee, Deaton and Duflo 2004, Banerjee et al. 2008); limited knowledge, as well as an important "know-do" gap among health professionals (see Das, Hammer and Leonard 2008 for a review); limited availability of diagnostic testing, leading to high rates of inappropriate treatment (Banerjee, Deaton and Duflo 2004; Cohen, Dupas and Schaner 2015); and drug quality (Bennett and Yin, 2014; Nayyar et al., 2012). Another common concern is that of corruption among health providers, though quantitative evidence on this is limited and the evidence to date is much less stern than anticipated (Dizon-Ross, Dupas and Robinson 2014). The majority of experimental economic research on these issues has focused on testing the effectiveness of interventions aimed at improving quality by changing how providers are monitored and/or incentivized. Only a handful have attempted to tackle the problems of diagnostic and drug quality.

The healthcare market in most countries in the developing world is comprised of government facilities with trained professionals, adjacent to a myriad of loosely regulated informal providers, from quack doctors to drug shops staffed by individuals with no formal pharmaceutical training but who from whom medical advice is regularly sought. Incentives for healthcare providers at public facilities to come to work or to perform well while at work are generally very weak (World Bank 2004; Das and Hammer 2014). In contrast, private providers face (some) market incentives to perform, but their ability to do so can be limited by poor or complete absence of medical training – and given the traditional information asymmetry in the health sector, patients' ability to walk away from low quality informal providers may also be limited. In this set-up, there are two ways to improve the quality of the health services that the majority of the poor have access to. The first is to better improve incentives for trained providers in the public sector. The second is to improve the quality of informal providers through training.

4.1. Monitoring

Interventions of the first kind (improving incentives for trained providers) that have been tested to date can be grouped into three primary types: input-based incentives (e.g. nurses are paid a bonus or avert a fine if their absenteeism level is low enough), output-based incentives (e.g. nurses are paid a bonus if health outcomes in their community are high enough) and decentralization (giving monitoring power to local communities). We discuss each in turn below. First, however, we note that the same types of interventions have been tried in the education sector, and a number of findings are likely to carry through across sectors. For brevity, we do not mention the education experiments here however, and refer the reader to Chapter X (section XX) for those instead. We also note that there have been many provider performance experiments the public health field.

Top-down, Input-based incentives

Banerjee et al. (2008) evaluated an incentives program for Assistant Nurse Midwives (ANM) at Primary Health Subcenters in Udaipur District, in the Indian state of Rajasthan. The program was implemented collaboratively by a non-profit organization and the state and local health administrations, with the goal of improving ANM's attendance at rural subcenters. Indeed earlier research had established that due to pervasive absenteeism among ANMs, health centers were closed 56 percent of the time during regular business hours (Banerjee, Deaton and Duflo, 2004). The program tested consisted in monitoring ANM attendance and "punishing" poor attendance: ANMs absent for more than 50 percent of the time on monitored days would have their pay reduced proportional to the number of absences recorded that month, and ANMs absent for more than 50 percent of the time on monitored days for a second month would be suspended from government service. The program was implemented in 49 randomly selected subcenters. In those centers, the ANM was required to stamp a register secured to the wall of the subcenter three times a day: once at 9am, once between 11am and 1pm, and once at 3pm – using a tamper-proof time/date-stamping machine. Researchers measured the impact of the program on ANM performance through random unannounced visits to the 49 "treatment" subcenters and 51 control subcenters. The results are mixed. In the short-run, the incentive scheme was highly successful, doubling attendance, from around 30 to 60%. The program was not popular with nurses, however, who heavily complained to the local health administration about the pay deductions. The share of “missed stamps” due to either an (intentionally) broken time clock or excused absence considerably increased over time, and 16 months after program inception, the absence rates were comparable between treatment and comparison centers. What the researchers take away from these mixed results is that, on the one hand, nurses are responsive to properly administered incentives, but on the other hand, incentive systems can be very difficult to properly administer, due to the perennial question of “who monitors the monitor?”.

A similar experiment took place a few years later with staff of primary health centers in Karnataka (Dhaliwal and Hanna, 2013). Instead of an NGO, the program was designed and implemented by the National Rural Health Mission (NRHM) of Karnataka, the lead department for the delivery of health services in the state. Instead of stamps, the monitoring system relied on fingerprints taken at the beginning and end of each day. Instead of pay deductions, the penalty was loss of paid vacation days, but the penalty was rarely imposed. The researchers found that the monitoring system increased attendance among medical staff by 18%, but not among doctors. They also find a large, 26% decrease in the incidence of low birth weight, confirming that provider attendance is a critical input in the health production function. The mechanism through which birth weight was affected was not through an increase in prenatal care attendance but an increase the likelihood that prenatal clients received iron folic acid tablets.

Top-down, Output-based incentives

While provider attendance may be the first step, it may not always lead to an increase in health outcomes for the population. For this reason, more recent monitoring experiments have based the rewards on outcomes rather than on attendance. Basing incentives on actual health outcomes is difficult, however, for reasons discussed in Miller and Babiarz (2014), in particular, the fact that provider behavior is only one of many factors mattering for health outcomes, and the fact that health outcomes can be difficult and expensive to measure. Given this, outcomes over which performance-pay contracts can be written tend to relate to the utilization or coverage of health services, e.g. the share of children who are immunized, the share of pregnant women seeking prenatal care, the share of deliveries that take place at the facility, etc. The potential downside of contracting over such pre-specified indicators is that providers may devote too much effort to those, at the expense of activities related to non-contracted indicators which may be as important for the production of health but simply harder to measure.

At the time of writing, we are aware of two economic field experiments that have tested the impact of performance-based incentives in the health sector, one in Indonesia and one in Rwanda. In both cases, the incentives were at the group level (not individual) and the performance mattered for the total budget available to the providers, rather than for their personal gain. (As far as we know individual performance-based incentives for health workers have not yet been experimented with.)

The Indonesia experiment estimated the effect of incentivized community-based block grants (Olken, Onishi and Wong 2014). The aim was to improve both health and education. The program, known as Generasi, gave villages annual block grants of $8,000 to $14,000 to be used towards either health or education programs. Villages were encouraged to use the funds to make progress on 12 pre-specified maternal and child health indicators, including prenatal visits, delivery by trained midwives, childhood immunizations, and growth monitoring. For the experiment, conducted jointly with the Government of Indonesia, 264 subdistricts were randomized into either a comparison group or one of the two versions of the Generasi program: the "incentivized" version with a pay-for-performance component, or the otherwise identical, "non-incentivized" version without pay-for-performance incentives. In the first year of the program, villages in all groups received program funds based on their size and demography. In the second year, the allocation rule staid the same for the non-incentivized villages, but for the incentivized villages 20 percent of the funds were distributed based on village's performance on the 12 indicators during the last year.

Impacts were measured over two years. The incentives led to an increase in the labor hours of midwives, the major providers of maternal and child health services in the area. Likely as a result, the targeted maternal and child indicators were somewhat higher in incentivized villages than in non-incentivized villages but the effect was modest, with a gap of just 0.04 standard deviations on average. The main impacts were on the number of prenatal visits (+8.2 percent) and regular monthly weight checks for children under five (+4.5 percent). The effect of the incentives varied with the baseline levels of service delivery, however – the effect was stronger in the poorer, off-Java provinces. Interestingly, no detrimental effect of the incentive scheme was found on non-targeted indicators (to the extent they could be measured).

The Rwanda experiment was conducted in partnership with the Rwanda Government as it launched a national pay for performance scheme to supplement primary health centers budgets (Basinga et al., 2011). As a pilot, the program was launched first in 80 randomly assigned facilities, with 86 facilities assigned to a comparison group. Under the program, facilities received payments as a function of their performance on 14 maternal and child health-care output indicators, including most of those in the Indonesia study. Performance was assessed as follows: facilities in the program had to submit monthly activity reports which were then audited against the facility's records. The specific payment amounts differed for each service, between US$0.09 for an initial prenatal visit and US$4.59 for an institutional delivery. Facilities in the control group received funding as a function of their size and the demographic characteristic of their catchment area. The researchers find large impacts on some of the targeted indicators. In particular, the incentives led to a 23% increase in the number of institutional deliveries, a 56% increase in the number of preventive care visits by children aged 0 to 2 years and 132% increase in the number of preventive care visits by older children. They also found a 0.16 standard deviations increase in prenatal quality as measured by compliance with Rwandan prenatal care clinical practice guidelines, but no change in the quantity of prenatal care sought or in rates of full immunization among children.

One of the mechanisms underlying the effects appear to be an increase in provider productivity. The researchers measured productivity as "the gap between provider knowledge and actual practice of appropriate prenatal care clinical procedures" (Gertler and Vermeersch, 2013). This gap appears substantial in the control group: while providers know 63 percent of the appropriate clinical protocols for prenatal care on average, they appear to only deliver about 45 percent of the appropriate protocols. This 18 percentage point gap was reduced by 4 percentage points in the incentivized facilities. The gap is much larger to start with among providers who have skills to start with, and the impact of the incentives is larger for those.

Bottom-Up: Beneficiary control

While monitoring coupled with incentives (whether carrots or sticks, and whether input or output based) can be successful at improving provider performance, these programs can be costly to implement. The monitoring costs can become prohibitive in remote areas, and as discussed earlier they often generate the problem of who monitors the monitor. For this reason, an obvious alternative would be to make the monitor the person who is the direct beneficiary of the gains to be had -- the patient. In other words, citizens, as *clients* of healthcare providers, have a direct interest in seeing the performance improve and this should translate into a willingness to expand some monitoring effort or cost. The difficulty here is that of monitoring ability: how do patients know if their doctor is making the right diagnosis? The right prescription? Health care is one of the domains, along with auto repair, plumbing, etc., where the client can have difficulty evaluating the performance of the informed expert providing the service. For this reason, the impact of increasing beneficiary control by itself may be limited. The existing experimental evidence to date confirms this. It comes from a two experiments conducted in Uganda.

In a first experiment conducted in nine districts of Uganda, Bjorkman and Svensson (2009) partnered with an NGO that focused on increasing local accountability of health providers. The experiment was conducted with 50 communities (with one facility each), 25 treatment and 25 control. In the treatment communities, the NGO first created "report cards" on the quality of services at the health facility, based on information generated through facility audits as well as household interviews conducted by the researchers. A unique report card was established for each facility, and it contained (1) information on key areas subject to improvement, including utilization, access, absenteeism, and patient-clinician interaction; and (2) comparisons vis-à-vis other health facilities and with the national standard for primary health care provision. The report cards were written in the local vernaculars and included graphics to help the non-literate. The NGO then facilitated three sets of meetings: a provider staff meeting, a community meeting and an interface meeting. The staff meeting was fairly short and consisted in sharing and discussing the content of the report cards. The community meeting gathered around 150 community members (the NGO made sure all stakeholders were represented, including the young, elderly, women). Participants were asked to critically review the quality of the health services available to them locally in an open discussion, and through this process the NGO disseminated the information on the report cards. Participants were then encouraged to identify concrete steps the local providers could take to improve quality, as well as concrete actions community members could take to monitor the providers take those steps. The discussion and proposed solutions were summarized at the end of the meeting in an action plan. The content of the action plans differed across communities, but the researchers note that high rates of absenteeism, long waiting time, weak attention of health staff, and differential treatment were common to many of the 25 communities in the treatment group. The interface meeting encouraged community members and health workers to discuss patient rights and provider responsibilities. At the end of the inferface meeting, the community and the facility staff reached an agreement on what the way forward. This shared action plan was called a "community contract" spelled out concrete steps for the provider to take and concrete ways through the community members would monitor them. The NGO came back after six months to conduct two additional meetings, a community and an interface meeting, to discuss the progress to date and fine-tune the action plan.

To estimate the impact, the researchers conducted surveys one year after the first set of three meetings had taken place. They surveyed households (the same households surveyed prior to the intervention and whose information was used for the report cards), as well as health staff. They also collected administrative records from the facilities and performed visual checks on them. They find large impacts on both the quality of care and on health outcomes. A year after the first round of meetings, health facilities in treatment communities had taken significant steps to reduce wait time, in particular through the introduction of numbered waiting cards (20% of the treatment facilities had them compared to only 4% of the comparison facilities) and a 13% reduction in absenteeism, leading to a reduction in wait time of 12 minutes on average. This is despite the fact that utilization of general outpatient services was 20 percent higher in the treatment group, with households shifting away from traditional healers and self-treatment and towards the local public facility. In particular, immunizations increased for all age groups, especially newborns, and prenatal care attendance also increased, contributing to a 0.14 z-score increase in infant weight and a remarkable 33 percent reduction in under-5 mortality. Encouragingly, these large effects persisted over time. Four years after the intervention, researchers went back to collect new data (Bjorkman-Nyqvist, de Walque and Svensson 2014). They found that the treatment communities still had significantly higher rates of utilization by households, better adherence to clinical guidelines by providers, and better health outcomes (reduced child mortality and increased weight-for-age and height-for-age for children).

The results of this first experiment suggest that beneficiary control can work – but it is notable that in this study beneficiaries *were given information* that they could act on, as well as, even if implicitly, information on how to stay informed (the report cards gave them concrete items to look for when monitoring). Does beneficiary control work absent this fairly costly information provision? The second experiment, by the same researchers in the same Ugandan context, suggests that the answer is no (Bjorkman-Nyqvist, de Walque and Svensson 2014). In new communities, the researchers designed an intervention that mimicked everything in the first experiment, except for the report cards. Thus, in the new intervention (called "participation only", in contrast with the "participation and information" treatment of the first experiment) the staff and community meetings started without any quantitative or even qualitative information being provided by the NGO. They found no impact of this "participation only" intervention. The authors conclude that information is key, and theorize that it is because it enables users to better distinguish between health workers' effort and factors that also matter for outcomes but are outside the health workers' control. This makes monitoring possible (since the user knows what to focus its monitoring efforts on), and this beneficiary monitoring is what leads to improved health workers' performance.

4.2. Improving the quality of Informal Providers

Although the experiments described above suggest that increasing the accountability of public health providers can lead to increases in their productivity, the extent to which this will affect citizens' health outcomes depend on how the "market share" of these public providers. Services such as prenatal care are rarely provided outside regulated facilities, but curative primary care is commonly provided by private-sector, informal providers with at best minimal medical training. The large role of poorly trained "quack" doctors has been well documented in India, in particular (see Das and Hammer, 2014, and references therein). In sub-Saharan Africa, it is common for households to procure medication through retail sector drug shops without consulting public providers first (Cohen, Dupas and Schaner 2015). In such contexts, while increasing widespread availability of quality public care may be the goal in the long run, improving the quality of the care and/or medical advice provided by informal providers may be needed in order to improve health outcomes in the short run. In this section and the next, we discuss two recent economic experiments, one in India and one in Uganda, of programs designed to improve the quality of services and products available outside the formal sector.

In West Bengal, India, Banerjee et al. (2015) estimate the impact of offering training to existing informal providers (IPs) on the quality of the care they provide. The training program included 72 sessions of 2 hours, spread over 9 months (4 hours per week), and was taught by certified medical doctors, but no training certificate was issued upon completion of the training. The training covered multiple illnesses. Emphasis was placed on basic medical conditions, triage, and avoidance of harmful practices, accompanied by frequent patient simulations. Informal providers could continue operating their clinics throughout the training since the training demanded only 4 hours of their time each week. Nevertheless, take-up of the training was not universal. Out of 360 providers initially asked whether they were interested in the training program, 304 (84.5%) initially signed up. Half of those 304 were then randomly assigned to start the training immediately (the treatment group). Of those 152 IPs in the treatment group, 20 (13%) quit the program within three sessions, bringing take-up to just above 70%. The attendance rate over the training period was 64% among those IPs that took up the program. To measure impacts, the researchers used unannounced standardized patients – these are individuals recruited from the local community and trained to present consistent cases of illness to providers. This method is considered the "gold standard" in care quality measurement because it does not suffer from observation and recall bias, and because it generates estimates of both the quality of the diagnosis (since illnesses are pre-specified in the study design) and of the treatment prescribed conditional on the diagnosis. In the West Bengal experiment, standardized patients were trained to depict symptoms of either angina, asthma, or dysentery in a child asleep at home. These three conditions require different dimensions of care (angina requires referral, asthma requires identification of a chronic lung disease, and dysentery requires the provision of ORS) which were all supposed to be impacted by the training. The data collection through SPs started three months after the completion of training. SPs were completed for 267 of the 304 providers in the study sample. Additional data collected through clinical observations generated results consistent with those of the SP visits. The researchers found a significant, positive impact of the training on the quality of care. Being assigned to the training group improved case-specific checklist adherence by 4.2 percentage points (from a base of 27.3% in the control group) and the likelihood of correct treatment by 7.8 p.p. (from 52% to 59.8%). Prescription of antibiotics (unnecessary in all cases) remained unchanged at very high levels (close to 50%), though interestingly such unnecessary or even harmful practice is even more common in the public sector. Patients may thus mistakenly expect to be prescribed medicines in all cases, leading trained informal providers to continue prescribing them in order to keep their clients satisfied. Overall though, the results of this West Bengal experiment suggest that training existing IPs can improve the quality of care for rural populations with little access to care from fully qualified providers in either the public or private sector.

When existing providers are absent, or unwilling to go for training, an alternative is to encourage entry of high quality providers. Bjorkman-Nykvist et al. (2014) evaluate the impact of market-based community health care program led by two NGOs, BRAC and Living Goods, in Uganda. In treatment villages (107 villages randomly chosen out of 204), the NGOs recruited and trained women (one per village) to be a "Community Health Promoter" as well as an incentivized sales agent – conducting home visits to not only educate households on essential health behaviors but also sell preventive and curative health products at 20-30% below prevailing retail prices, earning a margin on product sales. Data collected from households three years after the rollout of the intervention in treatment villages suggests that the introduction of these trained informal providers considerably increased care seeking and resulted in a 25% reduction in under-five mortality.

One potential channel through which the community health promoter program improved outcomes could have been by influencing other actors to improve the quality of services and products that they provide/sell. In a companion paper, the researchers document that this may have been the case with respect to drug quality (Bjorkman-Nyqvist, Svensson and Yanagizawa-Drott 2013). The quality of the antimalarials that community health promoters were allowed to sell was strictly monitored by the NGOs, and the authors argue that this reduced the likelihood that antimalarials sold at drug shops in the treatment villages were counterfeits, through a pro-competitive effect. Exploiting the randomized assignment of the program across villages, and using data on drug quality from a subset of villages, they estimate that the introduction of the community health promoter in the village led to an increase in the share of authentic artemisinin-based antimalarials sold by incumbent drug shops of 11-13 percentage points, corresponding to a roughly 50% decrease in the share of fake drugs.

The take-away from this experiment is not clear-cut however. Its results suggest that subsidizing high quality products in the private sector improves health outcomes, but the cost of such subsidization could be prohibitive, especially if one includes the cost that the implementing NGOs had to incur for quality control purposes. What's more, whether this subsidization requires that the promoters are paid through their sales rather than a fixed salary is unclear. The NGOs running the programs consider the implicit piece-rate pay system to be a critical feature of their model which they call "entrepreneurial", but the experiment was not designed to estimate the role of the incentive-pay in observed impacts.

5. Conclusion

We have surveyed the large and growing literature using field experiments to study issues of health access, usage and impacts in low income countries. There has been a veritable explosion of research in this area: 20 years ago, there was nearly no experimental research in development economics, and we are in a position today where a long chapter such as this one is only able to cover a relatively small share of the research literature. As we have argued, research progress has been particularly pronounced in the study of the demand for health products and services, the quality of health care, and in certain aspects of the question of health impacts. However, there remain a number of glaring gaps in the literature, and promising areas for future investigation, and we briefly survey these topics in this concluding section.

One of the most important areas of inquiry is how current adult health status affects contemporaneous labor productivity, however, this is also an areas where there remain relatively few well-designed field experiments (with Thomas et al 2003, 2006 being notable exceptions). One possible explanation is that relatively few field experiments on health in low income settings have taken place with private sector commercial partners, those who would be most likely to have access to a large pool of workers who might serve as participants in such a health intervention study. (As discussed above, most experiments on health in development economics have been conducted with government or NGO partners.) This remains a topic of great intellectual and public policy importance, and one where further research could have high returns.

A related limitation of existing research is the relative lack of work studying the long-run impacts of earlier health investments. While research evidence is beginning to accumulate in this area – including the long-term follow-ups to the well-known INCAP (Hoddinott et al 2008) and Kenya deworming (Baird et al 2015) cases – few studies are successfully able to combine field experimental research designs with long-term longitudinal data collection with high tracking rates. Such studies require long time horizons and prospective planning, not to mention large research budgets, and few have been successfully carried out. Putting in place the data collection plans to follow-up the participants in the large number of recent health experiments in development going forward could have a large research payoff. In the absence of such studies, most of our understanding about the long-run impacts of child health status comes from studies that rely on natural experiments (such as weather variation, as in the Maccini and Yang 2009 study from Indonesia), or variation in exposure to conflict or other large-scale political shocks (as in the work of Alderman et al 2006, Bundervoet et al 2009, Leon 2012, among many other recent papers), but in these cases the interpretation of resulting impacts is arguably less transparent than in a well-designed field experiment.

Another broad area where further research would be useful is in the area of mental health. While a growing number of studies, including many surveyed above, include measures of mental health, depression, anxiety and wellbeing as outcomes of particular health interventions (including in the WISE project discussed in Thomas et al 2003, 2006 study, the KLPS data used in Baird et al 2015, among others), relatively few studies by economists explore the effect of mental health interventions on economic or other life outcomes. This is despite the fact that there is growing evidence that many mental health problems are widespread in both high and low income societies, and that there is a strong correlation in the cross-section between mental health and socio-economic outcomes (see the evidence in Das et al 2008 and the references therein). The growing emphasis on psychological issues in economics overall and in the study of poverty, and in particular how particular psychological or neurophysiological processes affect the decision-making of the poor (for instance, see recent work by Mani et al 2013 and Haushofer and Fehr 2014), suggests that this is an area that is also ripe for further intellectual exploration.

Beyond deepening our understanding of links between health outcomes (along various dimensions) and economic outcomes, it would also be valuable to direct more research energy to understand the large-scale health systems reforms that have occurred in many low income countries during the past decade. For instance, many less developed countries have expanded government supported health insurance programs – including in Ghana, India, Indonesia, Mexico, and Rwanda, among many other countries – but relatively little academic health economics research has examined the performance, structure, and incentives produced by these reforms, or the resulting behavioral responses by households and individuals. This is in stark contrast to the health economics literature focusing on wealthy countries like the U.S., where much research energy has focused precisely on broader institutional and organizational issues in the health sector. While there is already some important work in this area (e.g., Gertler and Vermeersch 2013, Miller and Babiarz 2014), development economists working on health could learn much from the public finance and industrial organization economists working on related health systems issues in the U.S., Europe, and other high income regions.

Over the past two decades, as field experiments have become an increasingly common tool used by applied researchers in development economics and health (and other fields of economics), we have often looked to the methods and approaches used in clinical trials as a model on which to base our own work. That was certainly the case with much of the early experimentation in economics (as discussed in Duflo et al 2007), and has continued with the increasing use of study pre-registration and pre-analysis plans over the past few years (Miguel et al 2014). However, the increasing sophistication of the methods used to study health issues in low income countries, and the innovations in research design and measurements described above, now make it more likely in our view that much of the learning will flow in the opposite direction, from economics and other social science disciplines back into medical research. This appears especially likely given the growing awareness in global health research that the study of health policies, systems and individual behavior cannot be approached in the same way as traditional medical treatment efficacy trials (for a discussion of the rise of “implementation science” in health research, see Madon et al 2007, Mwisongo et al 2011). It is thus our hope that the discussion in this chapter will not only be of interest to development economists already working on health issues in low income settings, but will also prove useful to scholars in other fields or disciplines who are eager to engage with these issues.

References

Alderman, H., J. Hoddinott, and B. Kinsey. (2006). “Long term consequences of early child malnutrition”, *Oxford Economic Papers*, 58(3): 450-474.

Alderman, H., J. Konde-Lule, I. Sebuliba, D. Bundy, A. Hall. (2006). “Increased weight gain in preschool children due to mass albendazole treatment given during ‘Child Health Days’ in Uganda: A cluster randomized controlled trial”, British Medical Journal, 333, 122-6.

Alderman, Harold. (2007). “Improving nutrition through community growth promotion: Longitudinal study of nutrition and early child development program in Uganda”, World Development, 35(8), 1376-1389.

Ali M, Emch M, von Seidlein L, Yunus M, Sack DA, et al. (2005) Herd immunity conferred by killed oral cholera vaccines in Bangladesh: a reanalysis. Lancet 366: 44–49. doi:10.1016/S0140-6736(05)66550-6.

Ali M, Emch M, Yunus M, Sack D, Lopez AL, et al. (2008) Vaccine Protection of Bangladeshi Infants and Young Children Against Cholera: Implications for Vaccine Deployment and Person-to-Person Transmission. Pediatr Infect Dis J 27: 33–37. doi:10.1097/INF.0b013e318149dffd.

Ali M, Sur D, You YA, Kanungo S, Sah B, et al. (2013) Herd protection by a bivalent-killed-whole-cell oral cholera vaccine in the slums of Kolkata, India. Clin Infect Dis: cit009. doi:10.1093/cid/cit009.

Almond, D., and J. Currie, Human Capital Development before Age Five. NBER Working Paper #15827, (2010).

Anderson, Michael L. 2008. “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects.” Journal of the American Statistical Association 103 (484): 1481–95.

Angelucci, Manuela, and Vincenzo Di Maro. (2015). “Program evaluation and spillover effects”, unpublished work paper, University of Michigan.

Angrist, Joshua D, and Jörn-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. doi:10.1257/jep.24.2.3.

Ashraf, Nava, Jim Berry and Jesse Shapiro (2010). “Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia.” *American Economic Review* 100(5)

Ashraf, N., B.K. Jack and E. Kamenica. (2013) “Information and subsidies: Complements or substitutes?” *Journal of Economic Behavior and Organization* 88: 133-139.

Baird, S, J Hamory Hicks, and E Miguel. (2008). “Tracking, Attrition and Data Quality in the Kenyan Life Panel Survey Round 1 (KLPS-1)”, University of California CIDER Working Paper.

Baird S, De Hoop J, Özler B (2013) Income shocks and adolescent mental health. J Hum Resour 48: 370–403.

Baird, Sarah, Aislinn Bohren, Craig McIntosh, and Berk Ozler. (2014). “Designing Experiments to Measure Spillover Effects”, unpublished working paper, George Washington University.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel. (2015). “Worms at Work: Long-run impacts of a child health investment”, unpublished working paper, University of California, Berkeley.

Banerjee, Abhijit, Abhijit Chowdhury, Jishnu Das, Reshmaan Hussam (2015). “The Impact of Training Informal Providers on Clinical Practice in West Bengal, India: A Randomized Controlled Trial”. Working paper.

Banerjee, A, A Deaton, and E Duflo (2004). “Health Care Delivery in Rural Rajasthan,” *Economic and Political Weekly* 39, no. 9: 944-949.

Banerjee, Abhijit V. ., Esther Duflo and Rachel Glennerster (2008). “Putting A Band-Aid On A Corpse: Incentives For Nurses In The Indian Public Health Care System”. J Eur Econ Assoc; 6(2-3): 487–500.

Banerjee, Abhijit , Esther Duflo, Rachel Glennerster and Dhruva Kothari (2010). “Improving Immunization Coverage in Rural India: A Clustered Randomized Controlled Evaluation of Immunization Campaigns with and without Incentives.” *BMJ* 340:c2220.

Barber, Sarah L., and Paul J. Gertler (2010). “Empowering Women: How Mexico's conditional cash transfer programme raised prenatal care quality and birth weight.” *Journal of development effectiveness* 2, no. 1 (2010): 51-73.

Basinga, Paulin, Paul J Gertler, Agnes Binagwaho, Agnes L B Soucat, Jennifer Sturdy, and Christel M J Vermeersch (2011). “Effect on Maternal and Child Health Services in Rwanda of Payment to Primary Health-Care Providers for Performance: An Impact Evaluation.” *Lancet* 377: 1421-28.

Barham, Tania. (2011) “A healthier start: the effect of conditional cash transfers on neonatal and infant mortality in rural Mexico.” *Journal of Development Economics* 94, no. 1: 74-85.

Behrman, Jere R., Maria Cecilia Calderon, Samuel Preston, John Hoddinott, Reynaldo Martorell and Aryeh D. Stein. (2009). “Nutritional Supplementation of Girls Influences the Growth of their Children: Prospective Study in Guatemala”, American Journal of Clinical Nutrition, 90 (November), 1372-1379.

Benjamin-Chung, Jade, Jack Colford, David Berger, Benjamin Arnold, Veronica Jimenez, Diana Tran, Ashley Clark, Eugene Konagaya, Lauren Falcao, Jaynal Abedin, Alan Hubbard, Stephen Luby, Edward Miguel. (2015). “The Identification and Measurement of Health-Related Spillovers in Impact Evaluations: A Systematic Review”, unpublished working paper.

Berry, Jim, Greg Fischer and Raymond Guiteras (2012). “Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana”. Mimeo, London School of Economics.

Bhattacharya, Debopam, Pascaline Dupas, and Shin Kanaya. (2013). “Estimating the Impact of Means-tested Subsidies under Treatment Externalities with Application to Anti-Malarial Bednets”, unpublished working paper, University of Oxford.

Bjorkman-Nykvist, M., A. Guariso, J. Svensson, D. Yanagizawa-Drott. (2014). “Evaluating the impact of the Living Goods entrepreneurial model of community health delivery in Uganda: A cluster-randomized controlled trial.” Working paper.

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

Bjorkman-Nyqvist, Martina, Jakob Svensson and David Yanagizawa-Drott (2013). “The Market for (Fake) Antimalarial Medicine: Evidence from Uganda”. Working paper.

Bjorkman-Nyqvist, Martina, Damien de Walque and Jakob Svensson (2014). “Information is Power: Experimental Evidence of the Long-Run Impact of Community Based Monitoring,” World Bank Policy Research Paper Series No.7015.

Bleakley, Hoyt. Health, Human Capital, and Development. Annual Review of Economics 2, 283-310 (2010).

Bobonis G, Miguel E, Sharma CP. (2006). “Iron deficiency, anemia and school participation,” Journal of Human Resources, 41(4), 692-721,

Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg. 2012. Star Wars: The Empirics Strike Back. SSRN Scholarly Paper ID 2089580. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2089580.

Bundervoet, T. P. Verwimp, and R. Akresh. (2009). “Health and civil war in rural Burundi”, *Journal of Human Resources*, 44(2): 536-563.

Bundy, D.A.P. (1988). “Population Ecology of Intestinal Helminth Infections in Human Communities.” Philosophical Transactions of the Royal Society of London. Series B. 321 (1207), 405-420.

Bundy, D, Guyatt J. (1996) Schools for Health: Focus on Health, Education, and the School-Age Child. Parasitology Today; 12:1-16.

Bundy, D.A.P., Chan, M-S., Medley, G.F., Jamison, D., and Savioli, L. (1998). “Intestinal Nematode Infections.” In Health Priorities and Burden of Disease Analysis: Methods and Applications from Global, National and Sub-national Studies. Harvard University Press for the World Health Organization and the World Bank.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." The Quarterly journal of economics 127(4): 1755-1812.

Chinkhumba, Jobiba, Susan Godlonton and Rebecca Thornton (2014). The Demand for Medical Male Circumcision. *American Economic Journal: Applied Economics* 2014, 6(2): 152–177.

Chong, Alberto, Dean Karlan, Marco Gonzalez-Navarro and Martin Valdivia (2013). “Effectiveness and Spillovers of Online Sex Education: Evidence from a Randomized Evaluation in Colombian Public Schools.” Working paper.

Christensen, Garret, and Edward Miguel. (2015). “Transparency and Reproducibility in Economics Research”, unpublished working paper, U.C. Berkeley.

Cohen, J. and P. Dupas (2010). “Free Distribution or Cost-Sharing? Evidence from a randomized malaria experiment.” *Quarterly Journal of Economics* 125: 1-45.

Cohen, J., P. Dupas and S. Schaner (2015). “Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment.” *American Economic Review* 105(2): 609-645.

Cohen, Jessica, Günther Fink, Kathleen Maloney, Katrina Berg, Matthew Jordan, Theodore Svoronos, Flavia Aber & William Dickens (2015). “Introducing rapid diagnostic tests for malaria to drug shops in Uganda: a cluster-randomized controlled trial”. *Bulletin of the World Health Organization*, Volume 93, Number 3, 133-208.

Cooper E, Fitch L (1983) Pertussis: Herd Immunity and Vaccination Coverage in St. Lucia. The Lancet 322: 1129–1132. doi:10.1016/S0140-6736(83)90637-2.

Cox DR (1958) *Planning of experiments*. Oxford, England: Wiley. 308 p.

Croke, Kevin. (2014). “The long run effects of early childhood deworming on literacy and numeracy: Evidence from Uganda”, unpublished working paper, Harvard University.

Dal-Ré, Rafael, John P. Ioannidis, Michael B. Bracken, Patricia A. Buffler, An-Wen Chan, Eduardo L. Franco, Carlo La Vecchia, and Elisabete Weiderpass. 2014. “Making Prospective Registration of Observational Research a Reality.” Science Translational Medicine 6 (224): 224cm1.

Das, J., J. Hammer and K. Leonard (2008). “The Quality of Medical Advice in Low-Income Countries” *Journal of Economic Perspectives*, 22(2): 93-114. 2008.

Das, J., Q.T. Do, J. Friedman, and D. McKenzie. (2008). “Mental health patterns and consequences: Results from survey data in five developing countries”, World Bank Policy Research Working Paper #4495.

Das, J., and J. Hammer (2014). “Quality of Primary Care in Low-Income Countries: Facts and Economics”. *Annual Review of Economics* Vol. 6: 525–553.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. “Replication in Empirical Economics: The Journal of Money, Credit and Banking Project.” The American Economic Review 76 (4): 587–603.

DeLong, J. Bradford, and Kevin Lang. 1992. “Are All Economic Hypotheses False?” Journal of Political Economy 100 (6): 1257–72.

Dickson R, Awasthi S, Williamson P, Demellweek C, Garner P. Effect of Treatment for Intestinal Helminth Infection on Growth and Cognitive Performance in Children: Systematic Review of Randomized Trials”. British Medical Journal; 320, 1697-1701, 2000.

Dizon-Ross, R., P. Dupas and J. Robinson (2013). “Governance and Effectiveness of Health Subsidies”. Mimeo, JPAL.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, and Trena M. Ezzati. (1991) “The Item Count Technique as a Method of Indirect Questioning: a Review of its Development and a Case Study Application,” in: Beimer, P.B., Groves, R.M., Lyberg L.E., Mathiowetz N.A., Sudman S. (Eds.), Measurement Errors in Surveys. John Wiley & Sons, Inc., Hoboken, New Jersey, pp. 185-211.

Duflo, Esther, Pascaline Dupas and Michael Kremer (December 2014). “Education, HIV and Early Fertility: Experimental Evidence from Kenya”. Forthcoming, *American Economic Review*.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using randomization in development economics research: A toolkit." Handbook of development economics 4 (2007): 3895-3962.

Dupas, P. (2009). “What matters (and what does not) in household’s decision to invest in malaria prevention?”, *American Economic Review* 99(2): 224-230.

Dupas, P. (2011a). “Health Behavior in Developing Countries”, *Annual Review of Economics*, 3: 425-449.

Dupas, P. (2011b). “Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya”, *American Economic Journal: Applied Economics*, 3 (1), 1-36.

Dupas, P. (2014a). “Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment”. *Econometrica* 82(1), pp. 197-28

Dupas, Pascaline (2014b). “Getting Essential Health Products to Their End Users: Subsidize, but How Much?” *Science* Vol. 345, Issue 6202, pp. 1279-1281, 12 September 2014

Dupas, P. V. Hoffmann, M. Kremer and A. Zwane. (2013). “Micro-Ordeals, Targeting and Habit Formation”. Mimeo, JPAL.

Dybvig PH, Spatt CS (1983) Adoption externalities as public goods. J Public Econ 20: 231–247.

Easterbrook, P. J, R Gopalan, J. A Berlin, and D. R Matthews. 1991. “Publication Bias in Clinical Research.” The Lancet, Originally published as Volume 1, Issue 8746, 337 (8746): 867–72.

Eble, Alex, Peter Boone, and Diana Elbourne. (2015). “On minimizing the risk of bias in randomized controlled trials in Economics”, unpublished working paper, Brown University.

Epidemiology. 2010. “The Registration of Observational Studies—When Metaphors Go Bad:,” July, 1. doi:10.1097/EDE.0b013e3181eafbcf.

Fafchamps, Marcel, David McKenzie, Simon Quinn and Chris Woodruff (2014). “The Flypaper effect”.

Fine PE (1993) Herd immunity: history, theory, practice. Epidemiol Rev 15: 265–302.

Finkelstein, Amy, et al. "The Oregon Health Insurance Experiment: Evidence from the First Year." 2012. The Quarterly journal of economics 127(3): 1057-1106.

Fitzsimons E, Malde B, Mesnard A, Vera-Hernández M (2012) Household responses to information on child nutrition: experimental evidence from Malawi. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2034133. Accessed 22 April 2014.

Forleo-Neto E, Oliveira CF de, Maluf EMCP, Bataglin C, Araujo JMR, et al. (1999) Decreased Point Prevalence of Haemophilus influenzae Type b (Hib) Oropharyngeal Colonization by Mass Immunization of Brazilian Children Less Than 5 Years Old with Hib Polyribosylribitol Phosphate Polysaccharide—Tetanus Toxoid Conjugate Vaccine in Combination with Diphtheria-Tetanus Toxoids—Pertussis Vaccine. J Infect Dis 180: 1153–1158. doi:10.1086/315018.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. “Publication Bias in the Social Sciences: Unlocking the File Drawer.” Science 345 (6203): 1502–5.

Gerber, Alan S., Donald P. Green, and David Nickerson. 2001. “Testing for Publication Bias in Political Science.” Political Analysis 9 (4): 385–92.

Gerber, Alan, and Neil Malhotra. "Do statistical reporting standards affect what is published? Publication bias in two leading political science journals." Quarterly Journal of Political Science 3.3 (2008): 313-326.

Gerber, Alan S., and Neil Malhotra. "Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?." Sociological Methods & Research (2008).

Gertler, Paul, and Christel Vermeersch (2013). “Using Performance Incentives to Improve Health Outcomes”. NBER WP 19046.

Glewwe, Paul, and Edward Miguel. (2008). “The Impact of Child Health and Nutrition on Education in Less Developed Countries”, Handbook of Development Economics Volume 4, (eds.) T. Paul Schultz and John Strauss, Elsevier.

Godlonton, Susan, Alister Munthali and Rebecca Thornton (2014) “Responding to Risk: Circumcision, Information, and HIV Prevention.” Forthcoming in the Review of Economics and Statistics.

Godlonton S, Thornton R (2012) Peer effects in learning HIV results. J Dev Econ 97: 118–129.

Gong, Erick. (2015). “HIV Testing and Risky Sexual Behavior”, *Economic Journal*, 125: 32-60.

Grossman, Michael. On the Concept of Health Capital and the Demand for Health. Journal of Political Economy 80, 223-255 (1972).

Hanna, Rema, and Iqbal Dhaliwal (2013). “Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India”. Working paper.

Haushofer, Johannes, and Ernst Fehr. (2014). “On the psychology of poverty”, Science, 344, 862-867.

Hedges, Larry V. 1992. “Modeling Publication Selection Effects in Meta-Analysis.” Statistical Science 7 (2): 246–55.

Hedges, Larry V., and Jack L. Vevea. 1996. “Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model.” Journal of Educational and Behavioral Statistics 21 (4): 299–332.

Hoddinott, John, John A. Maluccio, Jere R Behrman, Rafael Flores and Reynaldo Martorell. (2008). “The Impact of Nutrition During Early Childhood on Income, Hours Worked, and Wages of Guatemalan Adults”, Lancet, 371 (February), 411-416.

Hoffmann, Vivian (2009). “Intrahousehold Allocation of Free and Purchased Mosquito Nets", *American Economic Review* 99(2): 236-41.

Jamison, Julian, Dean Karlan and Pia Raffler (2013). Mixed Method Evaluation of a Passive mHealth Sexual Information Texting Service in Uganda (May 2013) Information Technologies & International Development, 9(3).

Joshi S, Schultz TP (2013) Family Planning and Women’s and Children’s Health: Long-Term Consequences of an Outreach Program in Matlab, Bangladesh. Demography 50: 149–180.

Karlan, Dean, and Jonathan Zinman. (2011). “List randomization for sensitive behavior: An application for measuring use of loan proceeds”, *NBER Working Paper #17475.*

Karlan, Dean, Greg Fischer, Maggie McConnell and Pia Raffler (2014) “To Charge or Not to Charge: Evidence from a Health Products Experiment in Uganda”.

Kremer, M. and E Miguel (2007). “The Illusion of Sustainability.” Quarterly Journal of Economics 112

Kremer, Michael, Edward Miguel, Sendhil Mullainathan, Clair Null and Alix Zwane (2011). Social Engineering: Evidence from a Suite of Take-up Experiments in Kenya. Mimeo, Emory University.

Kvalsig JD, Cooppan RM, Connolly KJ. The effects of parasite infections on cognitive processes in children. Ann Trop Med Parasitol 1991;73:501–6.

LaBrie, Joseph W., and Mitchell Earleywine. 2000. “Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-Count Technique. Journal of Sex Research 37:321–26.

Laine, Christine, Richard Horton, Catherine D. DeAngelis, Jeffrey M. Drazen, Frank A. Frizelle, Fiona Godlee, Charlotte Haug, et al. 2007. “Clinical Trial Registration — Looking Back and Moving Ahead.” New England Journal of Medicine 356 (26): 2734–36.

LaLonde, Robert J. 1986. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” The American Economic Review 76 (4): 604–20.

Leamer, Edward E. 1983. “Let’s Take the Con Out of Econometrics.” *American Economic Review* 73 (1): 31–43.

Leon, Gianmarco. (2012). “Civil conflict and human capital accumulation: The long-term effects of political violence in Peru”, *Journal of Human Resources*, 47(4): 991-1022.

Loder, E., T. Groves, and D. MacAuley. 2010. “Registration of Observational Studies: The next Step towards Research Transparency.” BMJ 340 (feb18 2): c950.

Ma, Xiaochen, Sean Sylvia, Matthew Boswell, and Scott Rozelle (2014). Ordeal Mechanisms and Training in the Provision of Subsidized Products in Developing Countries. Mimeo, Stanford University.

Maccini, Sharon, and Dean Yang. 2009. "Under the Weather: Health, Schooling, and Economic Consequences of Early-Life Rainfall." American Economic Review, 99(3): 1006-26.

Madon, T., K.J. Hofman, L. Kupfer, and R.I. Glass. (2007). “Implementation science”, *Science*, 318 (5857): 1728-1729.

Maluccio, John A., John Hoddinott, Jere R. Behrman, Reynaldo Martorell, Agnes Quisumbing, Aryeh D Stein. (2009). “The Impact of Improving Nutrition During Early Childhood on Education among Guatemalan Adults.” Economic Journal, 199 (537) 734-763.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. (2013). “Poverty Impedes Cognitive Function,” Science, 341, 976-980.

Martorell R, Habicht JP, Rivera JA. History and Design of the INCAP Longitudinal Study (1969-1977) and its Follow-up (1988-89). Journal of Nutrition; 125: 1027S-1041S, 1995.

Mathieu, Sylvain, Isabelle Boutron, David Moher, Douglas G. Altman, and Philippe Ravaud. (2009). "Comparison of registered and published primary outcomes in randomized controlled trials", JAMA, Sept. 2, 2009, 302(9), 977-984.

McEwan, P. J. (2014). “Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments.” Forthcoming, Review of Educational Research.

Meredith et al. (2013). “Keeping the doctor away: Experimental Evidence on Investment in Preventive Health Products.” *Journal of Development Economics* 105: 196-210.

Miguel, Edward, and Michael Kremer (2004). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities”, *Econometrica.*

Miguel, Edward, and Michael Kremer (2014). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, Guide to Replication of Miguel and Kremer (2004)”, *CEGA Working Paper #39.*

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al. 2014. “Promoting Transparency in Social Science Research.” Science 343 (6166): 30–31. doi:10.1126/science.1245317.

Miller, G., K. S. Babiarz, Pay-for-performance incentives in low- and middle-income country health programs. In *Encyclopedia of Health Economics*, Vol. 2, A. J. Culyer, Ed. (Elsevier, San Diego, CA, 2014), pp. 457–483.

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

Mwisongo, A, L Wang, T Madon, S Owusu-Agyei and MÁ González Block. (2011) Current and foreseeable themes in implementation research for disease control. Chapter 9, Implementation research for the control of infectious diseases of poverty. Geneva: World Health Organization.

Nayyar, G., J. G. Breman, P. N. Newton, and J. Herrington, 2012, ”Poor-quality Antimalarial Drugs in Southeast Asia and Sub-Saharan Africa”, The Lancet Infectious Diseases, 12(6):488-496.

Neumark, David. 2001. “The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design The Employment Effects of Minimum Wages.” Industrial Relations: A Journal of Economy and Society 40 (1): 121–44.

Nokes C, Grantham-McGregor S, Sawyer A, Cooper E, Bundy D. Parasitic helminth infection and cognitive function in school children. Proceedings: Biological Sciences, 247(1319):77–81. 1992.

Nokes C, van den Bosch C, Bundy D. The Effects of Iron Deficiency and Anemia on Mental and Motor Performance, Educational Achievement, and Behavior in Children: A Report of the International Nutritional Anemia Consultative Group. USAID: Washington DC. 1998.

Olken, Benjamin, Junko Onishi and Susan Wong (2014). Should Aid Reward Performance? Evidence From a Field Experiment on Health and Education in Indonesia. A*merican Economic Journal: Applied Economics* 6(4): 1-34.

Ozier, Owen. (2014). “Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming”, World Bank Policy Research Working Paper #7052.

Paul JR, Horstmann DM, Riordan JT, Opton EM, Niederman JC, et al. (1962) An oral poliovirus vaccine trial in Costa Rica. Bull World Health Organ 26: 311–329.

Pitt, M.M., M. R. Rosenzweig, N. Hassan, Human Capital Investment and the Gender Division of Labor in a Brawn-Based Economy. American Economic Review 102, 3531–3560 (2012).

Pollitt E, Hathirat P, Kotchabhakadi N, Missel L, Valyasevi A. Iron deficiency and education achievement in Thailand. Am J Clin Nutr 1989(3);50:687–97.

Pollitt E, Gorman K, Engle P, Martorell R, Rivera J. Early supplemental feeding and cognition. Monographs of the Society for Research in Child Development, Serial No. 235, Chicago: University of Chicago Press. 1993.

Pop-Eleches C, Thirumurthy H*, Habyarmina J, Graff Zivin J, Goldstein M, DeWalque D, MacKeen L, Haberer J, Sidle J, Ngare D, Bangsberg D. Mobile phone technologies improve adherence to antiretroviral treatment in resource-limited settings: a randomized controlled trial of text message reminders. AIDS 2011; 25(6):825-834.

Prina, Silvia, and Heather Royer (2014). "The Importance of Parental Knowledge and Social Norms: Evidence from Weight Report Cards in Mexico". Forthcoming, *Journal of Health Economics*.

Robinson, Jon, and Ethan Yeh. (2011). "Transactional Sex as a Response to Risk in Western Kenya," *American Economic Journal: Applied Economics* 3 (1): 35-64.

Robinson, Jon, and Ethan Yeh. (2012). "Risk-Coping through Sexual Networks: Evidence from Client Transfers in Kenya," *Journal of Human Resources* 47 (1): 107-145.

Rosenbaum PR (2007) Interference Between Units in Randomized Experiments. J Am Stat Assoc 102: 191–200. doi:10.1198/016214506000001112.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638–41. doi:10.1037/0033-2909.86.3.638.

Rubin DB (1990) Formal mode of statistical inference for causal effects. J Stat Plan Inference 25: 279–292. doi:10.1016/0378-3758(90)90077-8.

Seshadri S, Gopaldas T. Impact of iron supplementation on cognitive functions in preschool and school-aged children: the Indian experience. Am J Clin Nutr 1989;50(3):675–86.

Schulz, Kenneth F., and David A. Grimes. 2005. "Multiplicity in randomised trials II: subgroup and interim analyses." The Lancet 365.9471: 1657-1661.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve: A Key to the File-Drawer." Journal of Experimental Psychology: General 143 (2): 534–47.

Singal, Amit G, Peter D R Higgins, and Akbar K Waljee. (2014). "A Primer on Effectiveness and Efficacy Trials", *Clinical and Translational Gastroenterology*, 5, e45; doi:10.1038/ctg.2013.13

Soemantri AG, Pollitt E, Kim I. Iron deficiency anemia and education achievement. Am J Clin Nutr 1989; 50(3):698–702.

Soewondo S, Husaini M, Pollitt E. Effects of iron deficiency on attention and learning processes of preschool children: Bandung, Indonesia. Am J Clin Nutr 1989;50(3):667–74.

Tarozzi, Alessandro, Aprajit Mahajan, Brian Blackburn. Dan Kopf, Lakshmi Krishnan, 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 DC, Maayan, N, Soares-Weiser, K. Garner, P. (2012) "Deworming drugs for treating soil-transmitted intestinal worms in children: effects on nutrition and school performance." Cochrane Database of Systematic Reviews.

The Lancet. 2010. "Should Protocols for Observational Research Be Registered?" 375 (9712): 348. doi:10.1016/S0140-6736(10)60148-1.

Thirumurthy H, Goldstein M, Graff Zivin J. "The economic impact of AIDS treatment: labor supply in western Kenya." *Journal of Human Resources,* 2008; 43:511-552.

Thomas, Duncan, et al. (2003). "Iron Deficiency and the Well-Being of Older Adults: Early Results from a Randomized Nutrition Intervention", unpublished manuscript, UCLA.

Thomas, Duncan, et al. (2006). "Causal Effect of Health on Labor Market Outcomes: Experimental Evidence", unpublished manuscript, UCLA.

Thornton, Rebecca (2008). The Demand for, and Impact of, Learning HIV Status. *American Economic Review*, 98(5): 1829–1863, 2008.

Vermeersch C, Kremer M. School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation. Unpublished working paper, World Bank and Harvard University.

Vivalt, Eva. (2014). "How much can we generalize from impact evaluations?", unpublished working paper, NYU.

World Health Organization. (1993). The Control of Schistosomiasis. Second Report of the WHO Expert Committee. (Technical Report Series 830). WHO, Geneva.

Ziegelhöfer Z (2012) Down with diarrhea: Using fuzzy regression discontinuity design to link communal water supply with health. Graduate Institute of International and Development Studies Working Paper. Available: http://www.econstor.eu/handle/10419/77433. Accessed 30 July 2014.

Zivin JG, Thirumurthy H, Goldstein M (2009) AIDS treatment and intrahousehold resource allocation: Children's nutrition and schooling in Kenya. J Public Econ 93: 1008–1015.

Zwane, Alix Peterson, Jonathan Zinman, Eric Van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, Dean Karlan, Richard Hornbeck, Xavier Giné, Esther Duflo, Florencia Devoto, Bruno Crepon, Abhijit Banerjee. (2011). "Being surveyed can change later behavior and related parameter estimates", *Proceedings of the National Academy of Sciences*.