Using RCTs to Estimate Long-Run Impacts in Development Economics

Adrien Bouguen[*], Yue Huang[†], Michael Kremer[‡] and Edward Miguel[§]

December 3, 2018

Abstract

We assess evidence from randomized control trials (RCTs) on long-run economic productivity and living standards in poor countries. We first document that several studies estimate large positive long-run impacts, but that relatively few existing RCTs have been evaluated over the long-run. We next present evidence from a systematic survey of existing RCTs, with a focus on cash transfer and child health programs, and show that a meaningful subset can realistically be evaluated for long-run effects. We discuss ways to bridge the gap between the burgeoning number of development RCTs and the limited number that have been followed up to date, including through new panel (longitudinal) data, improved participant tracking methods, alternative research designs, and access to administrative, remote sensing, and cell phone data. We conclude that the rise of development economics RCTs since roughly 2000 provides a novel opportunity to generate high-quality evidence on the long-run drivers of living standards.

[*]Department of Agricultural and Resource Economics, University of California, Berkeley, Berkeley, CA 94720, USA; email: abouguen@berkeley.edu

[†]Department of Agricultural and Resource Economics, University of California, Berkeley, Berkeley, CA 94720, USA; email: yuehuang@berkeley.edu

[‡]Department of Economics, Harvard University, Cambridge, MA 02138, USA, and NBER; email: mkremer@fas.harvard.edu

[§]Department of Economics, University of California, Berkeley, Berkeley, CA 94720, USA, and NBER; email: emiguel@berkeley.edu

Contents

1 INTRODUCTION

Development economics is an eclectic and methodologically rich field, featuring important contributions that utilize cross-country data, historical information, administrative records, and, increasingly, original survey data to understand the determinants of long-run living standards in poor countries. The roles played by human capital investments, well-functioning credit markets, and cash transfers have figured prominently in debates within the field (Gennaioli et al., 2012; Banerjee et al., 2015c; Haushofer & Shapiro, 2016), as have specific programs and policies that combine elements of these and other approaches (Banerjee et al., 2015b). Notably, development economists have also pioneered novel field experimental methods over the past twenty years, carrying out thousands of randomized controlled trials (RCTs), often in close collaboration with low-income country governments and non-governmental organizations.

This review article surveys what we have learned about the determinants of long-run living standards from this growing body of RCTs in development economics, and argues that these studies provide an exceptional opportunity to generate high-quality evidence on the impacts of a range of international development interventions on economic productivity and living standards.

For quantitative evidence on the rapid increase in the use of RCTs in development economics over the past 20 years, Figure 1 presents the cumulative number of registered RCTs in development economics conducted from 1995 to 2015 from the American Economic Association RCT Registry[1]. These data are likely to be lower bounds on the total number of relevant studies, since not all RCTs are registered, but the patterns in the data remain clear. Following influential early RCTs, such as the Mexico PROGRESA study (Skoufias & McClafferty, 2001), the Kenya deworming study (Miguel & Kremer, 2004) and the early education studies described in Kremer (2003), there was a surge in the use of RCTs in development economics in the decade of the 2000's. The data indicate that cash transfer programs and health interventions constitute a large share of these studies[2]. Our impression as researchers active in the field is that the pace of new RCTs in development economics has only accelerated since 2015.

The timing of this surge in development economics RCTs opens up an intriguing possibility: it has been roughly twenty years since these early interventions from the late 1990's and early 2000's were conducted, allowing researchers to begin to assess truly long-run impacts. For child health and education programs, beneficiaries in the early RCTs are now adults, allowing an assessment of long-run impacts on labor productivity, consumption, and living standards. Given the large numbers of RCTs launched in the 2000's, every year that goes by means that more and more RCT studies are "aging into" a phase where the assessment of long-run impacts becomes possible.

Beyond the opportunity presented by early RCTs in development economics, there are also many RCTs conducted by researchers in international public health that are promising. Figure 2 presents the cumulative number of RCT studies in the cash transfer and child health areas that

[1]Data were extracted from the AEA Registry (https://www.socialscienceregistry.org) on August 1, 2018. We extracted studies conducted in low- and middle-income countries, according to World Bank definitions. We focused on completed studies, and do not include RCTs that are ongoing, in the design phase, or withdrawn. This database is not comprehensive since study registration only became the norm around a decade after RCTs became common in development economics, and registration is voluntary; that said, many research institutions and journals are actively promoting registration of both ongoing and completed studies. We omit counts after 2015 since most such studies are ongoing.

[2]We considered a trial to be in the cash transfer or health category if its abstract contains the keyword "cash transfer" or "health", respectively.

have been published in both public health and economics journals, from the AidGrade database[3]. It is apparent that experimental research methods were widely adopted at least a decade earlier in public health (and related fields) than in economics, and that public health features an even larger body of evidence in terms of the raw count of studies.

Several of these studies generated exogenous variation in child nutrition and health that have already laid the groundwork for long-run evaluations. For instance, the famous INCAP child nutrition experiment in Guatemala was initiated in 1969 by public health researchers (Martorell et al., 1995), and 35 years later, economists followed up on the original sample and estimate significant gains in male wages, improved cognitive skills, and even some positive inter-generational effects (Hoddinott et al., 2008; Behrman et al., 2009; Maluccio et al., 2009), as we detail below.

Some early studies have the advantage of observing truly long-term changes, and have generated invaluable insights into the economic mechanisms underlying program effects, as well as documenting broader technological and institutional changes. Back in 1974–75, Christopher Bliss and Nicholas Stern led an extraordinary data collection effort in an Indian village, Palanpur. Along with other researchers, they surveyed this village intensively across several decades, and documented how lives and livelihoods changed from the 1970s to today, including local experiences with the Green Revolution and intensification of agriculture, structural transformation and increasing market integration (Lanjouw et al., 2018). A limitation of these early studies is their relatively small sample size of households and focus on a specific geographic region.

Another route to assess long-term program or policy impacts that has been more common is to exploit natural experiments. This strategy is common in the economic history literature. For instance, Bleakley (2007) exploits the introduction of a hookworm-eradication campaign in the U.S. South, combined with the cross-area differences in pretreatment infection rates, to form an identification strategy, and shows that the eradication campaign had long-lasting impacts on income and return to schooling. Similarly, Acemoglu & Johnson (2007) exploit the international epidemiological transition, which led to potentially exogenous differential changes in mortality from tuberculosis, pneumonia, malaria, and various other diseases; countries with larger baseline disease burden thus experience a larger reduction in mortality. Using predicted mortality as an instrument for life expectancy, the authors estimated the effects of life expectancy on population and GDP. Almond (2006) alternatively exploits the timing of the 1918 influenza pandemic, which arrived unexpectedly in the fall of 1918 and had largely subsided by January 1919, and shows large negative educational and labor market effects on cohorts in utero during the pandemic. Evidence from natural experiments are particularly compelling when the policy or program variation studied is truly random. Bleakley & Ferrie (2013), for example, take advantage of the 1832 Cherokee Land Lottery in the state of Georgia, to assess the long-term impact of large shocks to wealth. These studies have provided valuable insights into the long-term impacts of various cash transfer or health interventions, but natural experiments such as these are hard to come by. Rich historical census or other records are also necessary for researchers to link participants' treatment status to later outcomes, and those records tend to be less available in low-income contexts.

Changes from natural policy variations have also made it possible for researchers to study long-run impacts of cash transfer and child health interventions in wealthy countries. This literature

[3] We extracted data from the AidGrade project (http://www.aidgrade.org) in August 2018. AidGrade is a meta-analysis database focusing on 10 types of development aid programs. Notably, they use a somewhat different definition of sector than ours: they restrict attention to health interventions in deworming, HIV/AIDS education, micronutrients, school meals, bed nets, safe water storage, and water treatment, and not all studies focus specifically on children.

in general shows large, positive and persistent effects of such programs; while we briefly discuss it here, the extensive literature in high-income countries is not the focus of this article, and we refer interested readers to the recent surveys in Almond et al. (2017) and Hoynes & Schanzenbach (2018). In the United States, the Earned Income Tax Credit (EITC) program has been shown to improve beneficiaries' academic achievement, education attainment, employment and earnings in the long run (Chetty et al., 2011b; Bastian & Michelmore, 2018). Bastian & Michelmore (2018) estimate that an additional $1,000 in EITC exposure when a child is 13–18 years old increases adult earnings by 2.2%, with the primary channel being induced increases in pretax family earnings. Similarly, the Food Stamp Program improved child health in the medium run (East, 2018), reduced metabolic syndrome conditions such as obesity, heart disease, and diabetes in adulthood, and increased long-term education and earnings for women (Hoynes et al., 2016). The U.S. Mothers' Pension program implemented during 1911–1935 benefited the male children of the recipients up to 70 years later: the program increased longevity by one year, reduced the probability of being underweight by half, increased educational attainment by 0.34 years, and increased income in early adulthood by 14% on average, all substantial gains (Aizer et al., 2016). A notable exception is the Seattle-Denver Income Maintenance Experiment, which did not appear to generate long-run benefits (Price & Song, 2016). Public health interventions in the U.S., especially in early childhood, also generate long-run gains. For instance, the successful hookworm-eradication campaign in the American South increased school enrollment, attendance, literacy and income roughly 30 years later (Bleakley, 2007)[4]. Many of these studies exploit large-scale policy changes and leverage rich administrative data and census records, which are often not readily available in poor countries. As a result, much of the evidence from development economics that we survey in this article relies on original data collection, often including household surveys.

The remainder of this article proceeds as follows.

Section 2 summarizes and evaluates the growing body of evidence from RCTs on the long-term impacts of international development interventions, and finds that most (though not all) provide evidence for positive and meaningful effects on individual economic productivity and living standards. Most of these studies examine existing cash transfer, child health, or education interventions, and shed light on important theoretical questions such as the existence of poverty traps (Bandiera et al., 2018) and returns to human capital investments in the long term. One notable pattern in the existing body of evidence is the finding that impacts often differ substantially by respondent gender, arguably as a result of the different educational and labor market opportunities facing females and males in most low-income countries. Another is that several existing human capital investment programs, in both health and education, appear to have high rates of return, making them potentially attractive for public policy. We observe some heterogeneity in rates of returns for different age cohorts, echoing the literature on the attractiveness of early childhood interventions (Heckman, 2006), but we note that interventions targeting kids already in school often present high returns. We caution that the studies we summarized may not be representative of all the relevant interventions, because projects that attract enough interest and resources for long-term evaluations can be selected on certain traits by both researchers and donors. Many of these characteristics are unobservable and not well understood, thereby potentially generating publication biases.

[4]Early childhood interventions such as the Perry Preschool project lead to increases in high school graduation and college attendance rates, and some positive impacts on economic outcomes, criminal behavior, drug use, and marriage for women (Anderson, 2008). The Head Start program significantly reduced child mortality rates (Ludwig & Miller, 2007), and improved long-term education and health, closing one-third of the gap between children with median and bottom quartile family income (Deming, 2009).

Section 3 implements a systematic survey exercise and evaluates which existing randomized controlled trials are likely to be amenable to long-term follow-up research, with a particular focus on cash transfer and child health programs, which as we have shown are particularly abundant in the literature. We first consulted existing meta-analysis and survey articles and extracted hundreds of existing experimental studies in these two areas. We then implemented a rigorous screening procedure to identify the studies that could feasibly — and productively — be followed up in future research, after accounting for research design and data challenges, such as a lack of statistical power, high attrition rates or differential attrition across treatment arms, and phase-in designs that dampen cross-arm differences in program exposure. Fortunately, even after screening, we identify dozens of existing RCTs in the cash transfer and child health areas that appear to be attractive candidates for long-term follow-up studies today, where we typically use a follow-up period of roughly a decade to mean the long-run. We view this identification of studies that appear promising for long-run evaluation as a public good for the development economics research community.

Section 4 presents a methodological discussion on promising approaches to estimating long-term impacts, both among existing RCTs as well as approaches that can be taken prospectively to make long-run follow-up surveys more successful. We first discuss the assumptions under which it is possible to identify long-run treatment effects using a phase-in research design. We provide lessons from our experience in conducting long-term tracking studies, as well as innovative data approaches. An important methodological question is whether it is worthwhile to conduct follow-up research for RCTs that demonstrated limited short-run impacts, or whether it is safe to assume that any effects fade out in such cases. We discuss evidence from several existing studies that long-run impacts may exist even in the absence of clear-cut short-run effects. There are plausible conceptual reasons for such a pattern: if education and experience are complements in the labor market, the magnitude of program impacts can grow over time (Brunello & Comi, 2004). Policies aimed at improving education can even have negative impacts on beneficiaries' labor market outcomes in the short term, as they may remain in school or in training, or they may experience a longer job-searching period as they search for certain jobs (such as jobs in the public sector or formal sector). In such cases, the absence of long-run evidence could lead to the erroneous conclusion that the benefits of a human capital investment are small. The need for truly long-run labor market data may be particularly important for females, who often have lower labor market attachment during peak child-bearing years, before fully re-entering the labor force in mid-life.

Section 5 discusses the implications of this evidence for development economics. We conclude that the rise of development RCTs over the past two decades provides an exciting opportunity for scientific progress, by generating credible evidence on the determinants of living standards over the long-run. We predict and hope that the trickle of early studies that exploit RCTs to generate long-run evidence will become a flood in the coming years.

2 WHAT HAVE WE LEARNED? A REVIEW OF THE EXPERIMENTAL EVIDENCE

Relatively little is currently known about the long-run impacts of many common interventions in international development. A systematic review by the World Bank (Tanner et al., 2015) focusing on early childhood interventions was able to identify only a single study that reported later employment and labor market outcomes (Gertler et al., 2014). Contemporaneous work by Molina Millán et al. (2018a) focusing on conditional cash transfers also concludes that very few studies are able to confidently assess the later employment and labor market impacts of the transfers, as many beneficiaries are still in school and not yet in the labor force. As full-time students usually have lower earnings, estimates obtained when only a portion of the participants have entered the labor market could understate the true long-run benefits of an intervention, or even get the sign wrong, especially if the intervention increases schooling and delays labor market entry. Moreover, as noted in the introduction, if labor market experience and education are complements, even early estimates obtained when all participants are in the labor force could understate the true long-run benefits of the intervention, if individual labor productivity grows more rapidly over time for the more educated, for instance. This raises the possibility that very long-run evaluations may be necessary to confidently assess true programs impacts and cost-effectiveness.

In this section, we assess the evidence from the emerging body of literature that exploits RCTs to estimate long-run impacts of development interventions. One pattern that emerges from the handful of existing studies is that human capital interventions appear to be particularly effective at boosting long-run economic outcomes. For instance, direct investments in child health, such as deworming (Baird et al., 2016a), nutritional supplementation (Hoddinott et al., 2008), and perinatal interventions (Charpak et al., 2016) have all been found to generate meaningful impacts on adult labor productivity. Certain investments in education, including cognitive stimulation in early childhood (Gertler et al., 2014; Kagitcibasi et al., 2009) and scholarship programs (Bettinger et al., 2018) also yield positive returns. Interventions that aim to improve child education, nutrition and health by leveraging a conditional cash transfer similarly appear to have persistent effects on earnings in some cases (Barham et al., 2017), although not in others: Molina Millán et al. (2018b) find no meaningful impacts, possibly because their sample population is still relatively young.

The other set of RCTs that estimate long-run impacts examine unconditional cash transfers and various entrepreneurial grant assistance programs. These programs typically have quite large short-term effects on labor and firm productivity — see, for example, Blattman et al. (2013). Yet, most gains appear to fade out after several years (Blattman et al., 2018b,a; Araujo et al., 2017). Similar patterns are sometimes observed in medium-run follow-up studies (Baird et al., 2016b). One exception, which we discuss further below, is provided by multifaceted programs that provide assets to poor households as well as training and other forms of support (Banerjee et al., 2016; Bandiera et al., 2017, 2018), which appear to have more persistent effects.

Table 1 summarizes all the studies (to the best of our knowledge) that satisfy our screening criteria. For inclusion a study had to

1. be in a relevant category (cash transfer or child health interventions),
2. have randomized treatment,
3. report outcomes at least (roughly) ten years after the intervention started,
4. report labor market or living standards outcomes.

While we mainly focus on long-run impacts of cash transfers and child health interventions, for completeness we also briefly discuss other relevant studies in the main text (though some are omitted from Table 1 due to our inclusion criteria). In this review, we interpret "long-run impacts" to be persistent effects on the labor market or living standards outcomes of the program beneficiaries, over a period of roughly 10 years. We focus on labor market outcomes because they directly reflect individual productivity, and largely determine future household living standards in most cases. Many studies document short to medium-run schooling gains, but these may or may not translate into higher earnings due to institutional or other constraints, hence the importance of directly assessing labor market outcomes. It is notable that many of the studies discussed in Table 1 are new unpublished working papers (at the time of writing this article).

2.1 Long-run Impacts of Cash Transfers

Cash transfer programs can achieve large persistent impacts if (1) the poor have high returns to physical capital, and business grants relax constraints in their ability to borrow, save and mitigate risk, thereby improving living standards; or (2) the poor have high returns to human capital, and cash transfers or direct education and health interventions promote investments in education and health, thereby improving living standards. Even if cash transfers do not lead to persistent impacts on consumption, the rate of return and welfare impacts of the programs could potentially be large. Suppose, for example, that people are credit constrained, and have a high rate of return to a good purchased with the transfer that is not permanent but persists for several years, e.g., a motorcycle that they use as a taxi, or a metal roof that allows them to avoid purchasing grass for thatching, but that this good completely depreciates before the final endline measurement. Suppose also that they allocate all of the income generated by the good before it depreciated into immediate consumption, so there were no persistent welfare gains at the time of endline measurement. If the net present value of the temporary consumption gains due to the transfer were sufficiently large relative to the size of the transfer, the program may have nonetheless been highly beneficial. The total welfare impacts can be recovered if measurements are collected sufficiently frequently, although that is typically not the case. In this sub-section, we assess the accumulating evidence on the magnitude and persistence of the effects of cash transfer programs, beginning with unconditional cash transfer programs.

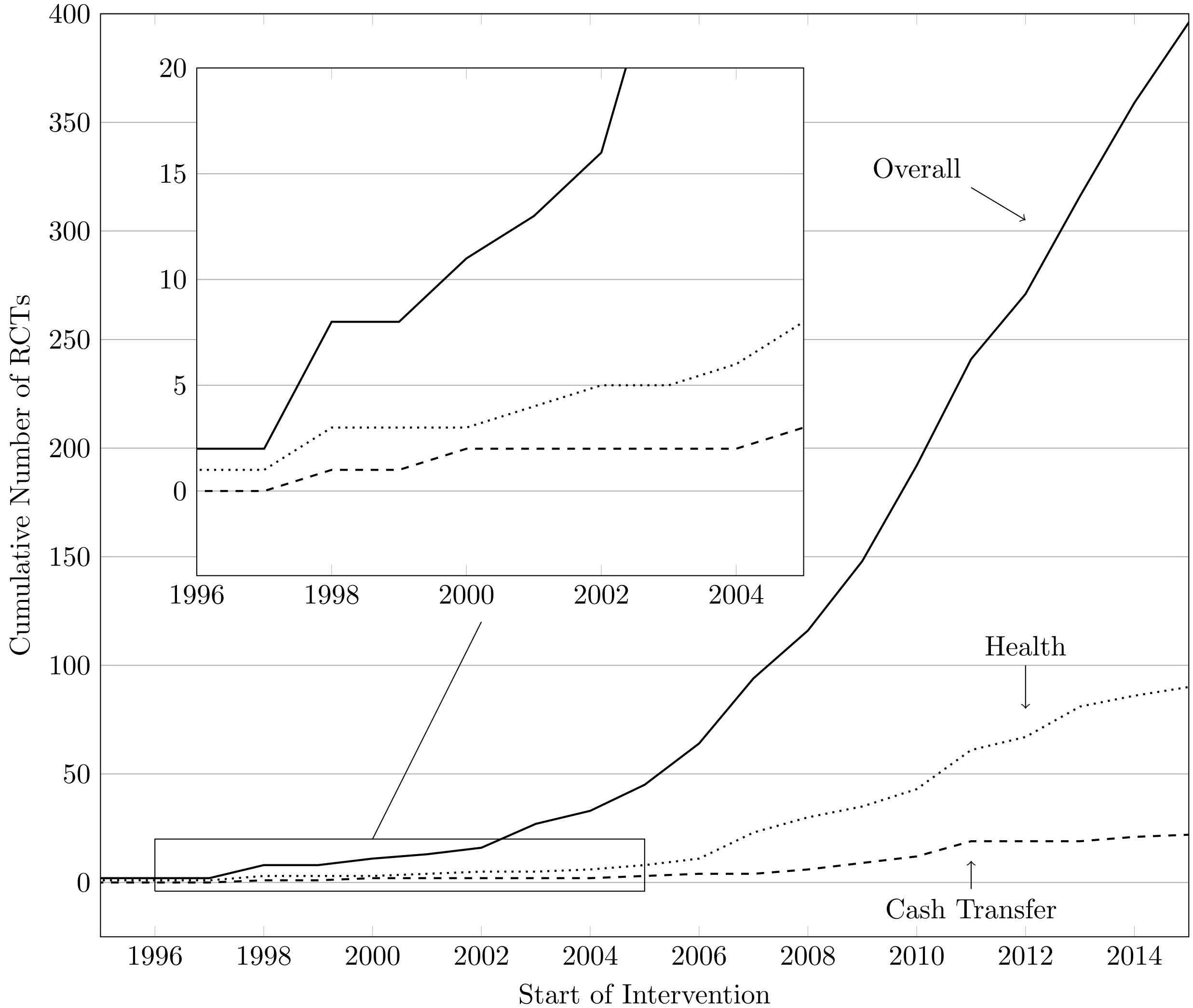

Figure 1: Cumulative Number of Completed RCTs in Low- and Middle-Income Countries from 1995 to 2015 in the AEA RCT Registry (https://www.socialscienceregistry.org)

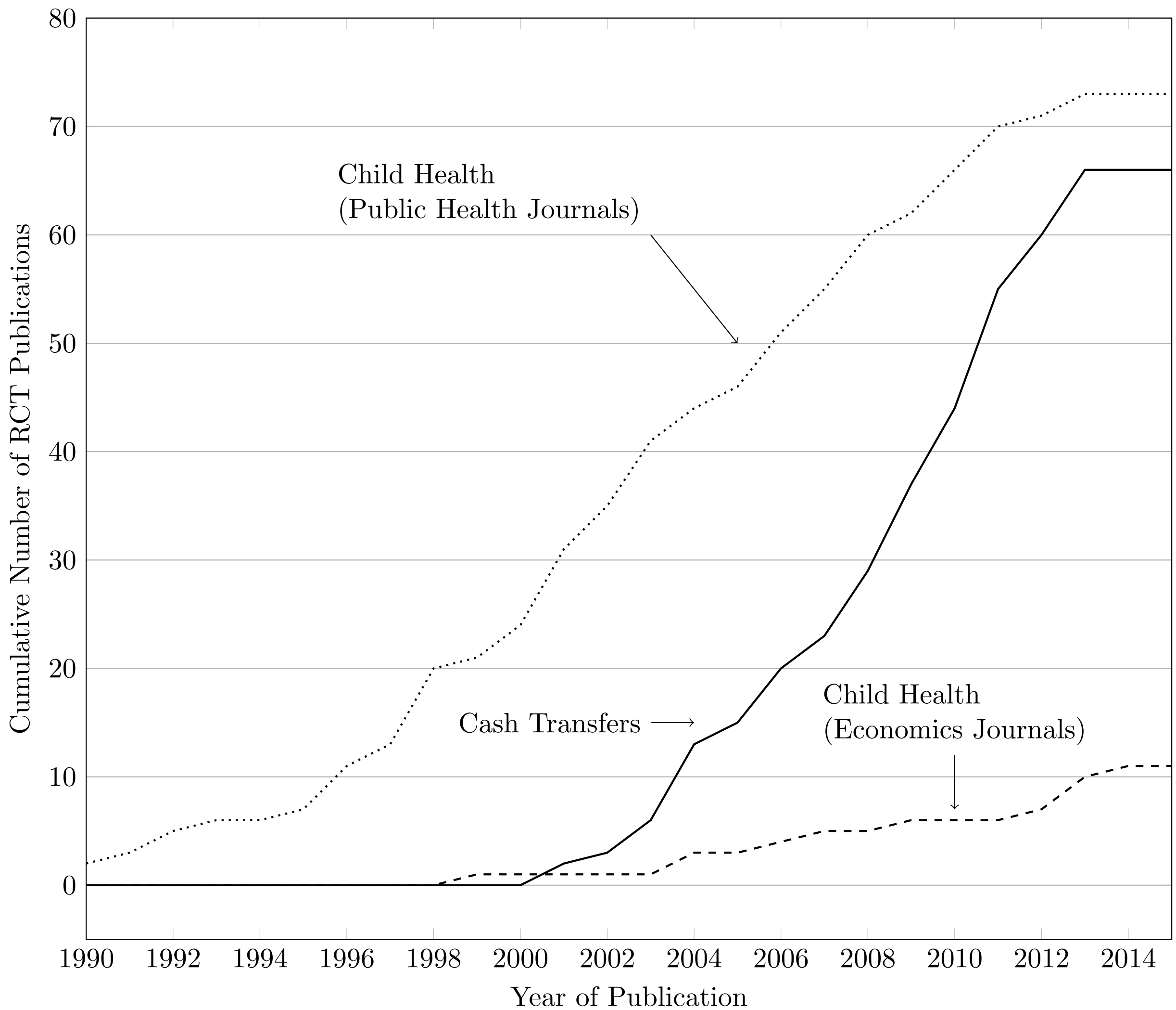

Figure 2: Cumulative Number of RCT Publications in Cash Transfer and Child Health in the AidGrade Database (http://www.aidgrade.org)

Table 1: Existing Evidence on Long-Run Impacts of Development RCTs

	Study	Country	Inter-vention	Start of Inter-vention	Years to Follow-up	Data	Attrition Rate	Long-Run Impacts
1	INCAP (Hoddinott et al., 2008)	Guatemala	Nutrition	1969	35	Survey	40%	Men experienced a 46% increase in average wages.
2	TEEP (Kagitcibasi et al., 2009)	Turkey	Stimulation	1983	22	Survey	49%	More education and more pretigious jobs. Higher consumption but not statistically significantly.
3	Jamaica (Gertler et al., 2014)	Jamaica	Stimulation	1986	22	Survey	19%	Increased earnings by 25% for beneficiaries who were growth-stunted as toddlers.
4	KMC (Charpak et al., 2016)	Colombia	Perinatal	1993	21	Survey	38%	Increased school attendance, labor force participation and wages.
5	PACES (Bettinger et al., 2018)	Colombia	Scholarship	1994	20	Admin	3%[a]	An 8% increase in earnings. Effects are driven by those who applied to vocational schools.
6	KLPS (Baird et al., 2016a, 2018)	Kenya	Deworming	1998	15	Survey	16%	More education, esp. for females. More nonagricultural self-employment. Increased total earnings by 13% and consumption by 23%, with effects concentrated among men.
7	RPS (Barham et al., 2017) (Barham et al., 2018)	Nicaragua	CCT	2000	10	Survey	10%	Positive effects on temporary migration and earnings, through increasing education and learning for men and improving nutrition and reproductive health for women.
8	PRAF II (Molina Millán et al., 2018b)	Honduras	CCT	2000	13	Survey[b]	N/A[c]	Improved education and increased the probability of migration. Effects on wages or earnings are unclear. Some negative effects on hours worked for women.

Table 1: Existing Evidence on Long-Run Impacts of Development RCTs (Continued)

	Study	Country	Inter-vention	Start of Inter-vention	Years to Follow-up	Data	Attrition Rate	Long-Run Impacts
9	BDH (Araujo et al., 2017)	Ecuador	UCT	2004	10	Survey	14%	No impacts on learning, small impacts on education and no impacts on labor market outcomes.
10	YOP (Blattman et al., 2018b)	Uganda	Grant	2007	9	Survey	13%[d]	No impacts on employment, earnings, and consumption.
11	THP (Banerjee et al., 2016)	India	Grant	2007	7	Survey	$< 1\%$	Positive effects on consumption, assets, income, food security, financial stability, time spent working, and physical and mental health.
12	TUP (Bandiera et al., 2017)	Bangladesh	Grant	2007	7	Survey	15%	Positive effects on consumption, productive assets and savings.
13	Ghana Scholarship (Duflo et al., 2018)	Ghana	Scholarship	2008	8	Survey	5%	More tertiary education; higher probability of obtaining more desirable jobs (e.g. jobs in the public sector, jobs with more benefits); reduced fertility for women.
14	Bangladesh Child Marriage (Buchmann et al., 2018)	Bangladesh	CCT	2008	9	Survey	13%	The conditional incentives delayed marriage and childbearing, increased education, and had insignificant effects on income-generating activities.

INCAP: the Institute of Nutrition of Central America and Panama; TEEP: the Turkish Early Enrichment Project; KMC: Kangaroo Mother Care; PACES: the Programa de Ampliacion de Cobertura de la Educacion Secundaria; KLPS: Kenya Life Panel Survey; RPS: Red de Proteccion Social; PRAF II: Programa de Asignacion Familiar Phase II; BDH: Bono de Desarrollo Humano; YOP: Youth Opportunities Program; THP: Targeting the Hard Core Poor Program; TUP: Targeting the Ultra-Poor Program.
UCT: Unconditional Cash Transfer; CCT: Conditional Cash Transfer.

[a] The authors obtain identification numbers that are valid for 97% of the sample. These IDs are then matched to each administrative datasets, with different match rates.

[b] For labor market outcomes, the authors use national household survey data, assigning treatment status to individuals based on their municipality of birth.

[c] Attrition rates are challenging to calculate because new samples were drawn and treatment status was assigned to individuals based on their municipality of birth. These samples do not necessarily correspond to the originally sampled households at the time of the intervention.

[d] Differential attrition observed across treatment and control groups.

2.1.1 Unconditional Cash Transfers

Evidence on the long-term impacts of unconditional cash transfers remains scarce, as they were fairly uncommon in the early wave of development RCTs in the 1990's and early 2000's. One exception is Araujo et al. (2017), who study the long-term effects of the Ecuador Bono de Desarrollo Humano (BDH, translation: Human Development Voucher) program. As in many development RCTs, the control group began receiving treatment three years after the treatment group, most likely dampening estimated program impacts and leading estimates to be lower bounds on true effects, relative to a trial design with a never-treated control group. The authors find that 10 years after the program, children in the early treatment group did not improve learning outcomes in late childhood. Using a regression discontinuity design to exploit the poverty index cutoff, they also show that cash transfers received in late childhood modestly increased the proportion of young women who completed secondary school, but did not affect their education and work choices after graduation. Taken together, there are limited detectable long-run impacts of a fairly generous unconditional cash transfer program.

The fade out of effects from unconditional cash transfers have also been observed in the medium run by Baird et al. (2016b)[5]. In a cash transfer program with both unconditional and conditional transfer arms in Malawi, they find that female unconditional cash transfer recipients show a modest delay in the timing of marriage, fertility and HIV infection, but that these effects fade out after roughly two years. There are some differences across the unconditional and conditional cash transfer arms; for instance, for girls who had already dropped out of school when the study began, two years of conditional cash transfers do produce meaningful increases in educational attainment and lead them to marry significantly more educated husbands, which may lead to long-run benefits.

2.1.2 Entrepreneurial Grants

The long-run effects of entrepreneurial grant programs are similarly mixed, possibly due to the very heterogeneous nature of the interventions and populations studied. Blattman et al. (2018b) study the Youth Opportunities Program (YOP), an entrepreneurial grant program launched in Uganda in 2008. The program granted hundreds of small groups $400 per person to "kick-start" microenterprises. The program increased average earnings by 38% and consumption by 10% after 4 years (Blattman et al., 2013), but after 9 years, the control group had completely caught up with the treatment group in terms of employment, earnings, and consumption. There are lasting effects on assets and occupational choice, suggesting some persistent economic gains. YOP beneficiaries and their children also show little to no health or education gains, except for modest improvements in physical functioning among children of female recipients. These results are consistent with the findings of Blattman et al. (2018a)[6], which finds that a program in Ethiopia that provided grants of $300 plus basic business consulting raised incomes by one third in the first year, but that employment and earnings largely converged across the treatment and control groups within 5 years.

While both unconditional cash transfer programs and entrepreneurial grant programs appear to initially help the poor accumulate assets, evidence from the limited number of studies at hand is broadly consistent and indicate that these assets are generally gradually run down over time,

[5]We omit this study from Table 1 because it only reports medium-run outcomes at a time horizon of less than a decade.

[6]We omit this study from Table 1 because it only reports medium-run outcomes at a time horizon of less than a decade.

generating little permanent impacts on poverty. One possible explanation is that neither type of program directly ties the transfers to human capital investments. A notable exception are programs that tie asset transfers to more intensive training and support, namely, the multifaceted assistance projects targeted at the extremely poor studied by Bandiera et al. (2017) and Banerjee et al. (2016). Banerjee et al. (2016) find that an asset transfer combined with support for 18 months in India generated impacts that persisted and even grew over seven years. Positive effects are found across all categories of outcomes, including consumption, assets, income, food security, financial stability, time spent working, and physical and mental health. Bandiera et al. (2017) find similar evidence for persistent effects of a program transferring livestock and training, valued at $1120 in 2007 PPP, to the ultra poor in Bangladesh. Like Blattman et al. (2018b), they find persistent effects on assets and occupational choice (in particular livestock rearing) seven years after the program, but in contrast to other work, they also find persistent effects on consumption. Some of the estimates for impacts after seven years are likely to be lower bounds since the control group was subsequently phased into treatment.

Using the same exogenous variation induced by the BRAC TUP program (Bandiera et al., 2017), Bandiera et al. (2018) present evidence for the existence of poverty traps in this setting, providing a potential explanation for the persistent gains they find. Conceptually, poverty traps can occur when there are increasing returns to scale to factors that can be accumulated, and when credit markets are imperfect. They find that in their sample, individuals' assets exhibit a bi-modal distribution: after the intervention, those above a certain threshold accumulate assets at a decreasing rate, while individuals below that threshold lose assets at an increasing rate. This is consistent with theoretical predictions from a poverty trap model with multiple equilibria. While the authors provide evidence for a poverty trap in their setting, few other cash transfer programs seem to generate persistent effects. It is, of course, possible that the assistance provided in other programs was simply too small to move recipients over the threshold needed to escape the poverty trap. Yet both the NGO Give Directly (Haushofer & Shapiro, 2018) and the 19th century Georgia land lottery (Bleakley & Ferrie, 2013) provided very large-scale transfers and the evidence from neither suggests poverty traps. Moreover, empirical wealth distributions are typically unimodal, rather than following the bimodal distribution observed in Bandiera et al. (2018)'s setting, so the existence of poverty traps may be specific to certain contexts.

What explains the differences in impacts between these "ultra-poor" programs and other enterprise grant and assistance projects? While there is no definitive answer, there are several plausible interpretations, beyond the possibility that this particular setting and population had conditions that led to a poverty trap. Differences in targeting (i.e., the poor versus the extremely poor) could play a role. Finally, the multifaceted and intensive training in the ultra-poor programs may have induced greater human capital accumulation or addressed behavioral barriers to saving. Further research is needed to provide more definitive answers.

2.1.3 Conditional Cash Transfers

Despite the proliferation of evaluations of conditional cash transfer programs, especially in Latin America, high-quality experimental evidence on their long-term impacts remains limited. The estimation of long-run impacts is also complicated by the fact that many RCTs employ phase-in designs, in which the control group later receives treatment, sometimes after only a year or so. As we discuss below (Section 4), phase-in designs are likely to yield lower bounds on the treatment effects that would be obtained with a pure control group (that never received treatment), somewhat changing the interpretation of effect estimates and also making null results harder to interpret.

Barham et al. (2017) and Barham et al. (2018) evaluate a three-year conditional cash transfer program in Nicaragua, which was later phased in. They exploit the fact that although both the "early treatment" and "late treatment" groups received three years' worth of cash transfers, in the latter group the boys largely missed the transfers that were most likely to prevent them from dropping out of school, and the girls missed the transfers during their potentially critical early teenage years, around the onset of puberty. After 10 years, among children aged 9–12 years at the start of the program, young men in the early treatment group show increased schooling and learning, which translated into more engagement in wage work, higher rates of temporary migration for better paying jobs, and higher earnings; young women in the early treatment group reached sexual maturity later, had lower BMI scores, and started sexual activities later, resulting in lower overall fertility in young adulthood. Despite modest effects on education and learning outcomes for young women, they experienced similar earnings and labor market participation gains as men.

Buchmann et al. (2018) evaluate a program aimed at reducing child marriage and teenage childbearing and increasing girls' education in Bangladesh. The authors cross-randomized a six-month empowerment program, and a financial incentive to delay marriage. Nine years after the program started (and four and a half years after the end of the program), girls randomized to receive conditional cash incentives get married later and are less likely to bear a child as teenagers. Unlike cash transfer programs that are conditional on school enrolment or attendance, this program also benefits vulnerable girls that are already out of school at baseline. The authors find that empowerment programs alone have no discernible effect on marriage, but do improve education. The empowerment program also increases income-generating activities, in particular labor market participation, among older girls.

Evidence from other related interventions is less conclusive. PROGRESA/Oportunidades is the pioneering Mexican conditional cash transfer program that has served as a model for many other programs in Latin America and beyond, and was evaluated with an RCT in its pilot phase. The program started in 1997 and offered monetary transfers to households conditional on investing in education, health, and nutrition of the children (e.g., attendance at school, regular clinic visits). Many studies have exploited the experimental design to assess medium- and long-run impacts (see Parker & Todd (2017) for a survey) but the long-run studies are limited by the original evaluation sample's research design: policymakers decided that the original control group villages would be phased into treatment a mere 18 months after the treatment communities, creating a relatively short gap between the early and late treatment groups. Moreover, participant attrition in follow-up survey rounds was relatively high and differential across treatment arms. Exploiting this data, Behrman et al. (2011)[7] show that in the medium-run by 2003 (six years after the start of the program), the greater exposure (of 18 months) to the program in early treatment PROGRESA/Oportunidades communities significantly increases schooling for both genders and decreases participation in the labor force for boys, but not for girls.

Given the limitation of the research design and follow-up survey data, long-run impacts of PROGRESA/Oportunidades on labor force participation, wages and earnings are mostly estimated non-experimentally,[8] and are mostly large and positive. Adhvaryu et al. (2018) focus on a small cohort who were 18 at the time of the 2003 survey, and show that PROGRESA has significant

[7]We omit PROGRESA/Oportunidades from Table 1 because the follow-up studies that are based on the original randomization only report medium-run outcomes.

[8]We omit these studies from Table 1 because they do not rely on RCT based estimators.

impacts on the probability of stable employment immediately following high school completion among disadvantaged children (proxied by rainfall shocks in early childhood), but no impact on children with greater endowments. Parker & Vogl (2018) use a difference-in-differences strategy and leverage both the spatio-temporal variation in program roll-out at the municipal level and cohort variation in the age at which children were treated. They find that childhood exposure to PROGRESA improves educational attainment, geographic mobility, labor market outcomes, and household economic outcomes in early adulthood: the program increased mean labor force participation by 30–40% and labor income by 50% for women. Kugler & Rojas (2018) exploit similar sources of variation, combined with propensity score weighting, and estimate significant positive impacts of the program on both the likelihood and quality of employment.

In contrast, Molina Millán et al. (2018b)'s evaluation of the Honduras PRAF II conditional cash transfer program estimates more ambiguous effects on beneficiaries' labor market outcomes. They use national census microdata and assign individuals their treatment status based on their municipality of birth, the unit of randomization in the RCT. Transfer recipients have significantly higher schooling attainment 13 years after the start of the program, with a notable increase in the likelihood of attaining university studies. Program receipt also more than doubles the probability of international migration among young men. However, impacts on labor market outcomes are less clear-cut: they find no significant treatment effects on wages or earnings, except for some negative effects on women's hours worked. The labor market results are difficult to interpret as some young adults are still transitioning into the labor market (for instance, the students enrolled in university), and thus further follow-up surveys could be useful to more reliably assess impacts on lifetime earnings.

There is some evidence that the mode of delivery for cash transfers may be important in determining schooling and other outcomes. For instance, Barrera-Osorio et al. (2017)[9] leverage administrative data to analyze the Colombian Conditional Subsidies for School Attendance (Subsidios Condicionados a la Asistencia Escolar) program in 2005. The experiment has three treatment arms: the "basic" bimonthly transfers; the "savings" treatment where families were forced to save a portion of the transfers until they make school enrollment decisions; and a "tertiary" transfer that is conditional on tertiary school enrollment. While the various cash transfer arms are all effective in boosting short-run secondary school enrollment, only the "savings" treatment improves longer-term educational outcomes, particularly tertiary enrollment.[10]

2.1.4 Scholarship Programs

Scholarship or school voucher programs are closely related to some conditional cash transfer interventions, in that the award is conditional on school attendance (although not identical, since scholarship funding can only be spent directly on education). There are now several long-run RCT evaluations of scholarship programs. As we survey here, these tend to show both meaningful gains in educational attainment and subsequent benefits in the labor market.

Bettinger et al. (2018) evaluate the PACES voucher program in Colombia. The program used a lottery to assign vouchers for private secondary schools among applicants from public elementary schools in the poorest two socioeconomic strata in Colombia. The authors sampled the lottery win-

[9]We omit this study from Table 1 because it does not report labor market outcomes.

[10]We refer interested readers to the contemporaneous work by Molina Millán et al. (2018a), which also summarizes and evaluates research on conditional cash transfers, while also bringing in more evidence from non-experimental studies and projects focusing mainly on schooling outcomes.

ners and losers in 1994 and matched their IDs to five different administrative datasets, including a rich set of educational, financial and labor market outcomes. Up to 20 years later, when applicants' average age was roughly 33, the voucher winners had completed significantly more tertiary education, had experienced lower teen fertility, had annual formal earnings that were 8% higher than lottery losers, and had greater access to formal consumer credit and better credit scores. Notably, these impacts on formal sector earnings are entirely driven by applicants to vocational (as opposed to academic) schools. Effects on formal earnings and payroll taxes are also concentrated among the top 40% of the sample distribution: much of the voucher effect appears to work through increasing the odds that winners make it into the "middle class". A fiscal calculation based on impacts on formal sector earnings and payroll taxes shows that the program is likely to generate large and positive public finance benefits.

Duflo et al. (2018) evaluate a 2008 secondary school scholarship program in Ghana, and also find positive effects on some labor market outcomes. The program randomized full scholarships for public high schools among rural youth who had gained admission but did not immediately enroll. Five years after receipt of the scholarship, winners show increased educational attainment and improved cognitive skills, and also engage in more preventative health behaviors; sample females are less likely to have become pregnant. Nine years after the program, treated individuals were significantly more likely to have public sector jobs, jobs which tend to be characterized by a high wage premium, more benefits, and greater job security. They are also more likely to have jobs with benefits, or jobs they characterize as permanent. Yet there were no significant differences between scholarship winners and losers in total earnings, log earnings conditional on positive earnings, or hours worked for those observed to be working. These null effects on earnings should be interpreted with caution, however, as the confidence intervals are fairly wide and cannot exclude either zero effects or very large private returns. Moreover, a non-trivial portion of participants are still receiving tertiary education, and that portion is significantly higher in the treatment group, raising the possibility that treatment effects may grow over time. This pattern is particularly important if we believe that, as Bettinger et al. (2018) showed, gains from these programs tend to be concentrated at the top of the distribution. The authors caution that while these results indicate positive private returns to education to the extent that education simply helps people get access to jobs with rents, it may not generate similar social returns. It is also worth noting that the impact of the scholarships on obtaining public sector jobs only became apparent over time, likely because many of the positions require tertiary education.

Taken together, the findings from the existing long-run randomized evaluations of both conditional cash transfer programs and scholarship programs indicate that very long follow-up periods — often of greater than a decade — may be necessary to confidently estimate program impacts on lifetime earnings. This is due to the fact that many beneficiaries are still in school in their 20's, and that certain positions (such as public sector or formal private sector jobs) have rising wage profiles that only become apparent over time. These issues may be particularly important for females in low income countries, many of whom also have lower labor market participation in early adulthood than they will exhibit later on in life.

A lack of statistical power is also often a challenge in long-term impact evaluation. Income is typically measured with considerable noise, especially in low- and middle- income settings. Measurement concerns are exacerbated by the fact that actual income is highly skewed, that most people obtain a large proportion of their income from self-employment or informal activities in low income countries, often with strong seasonal variability, and that these income sources may be subject to important reporting biases. For these reasons, in some settings, other socio-economic indicators, such as jobs with benefits, "permanent" jobs and public sector jobs as in Duflo et al. (2018), may sometimes be more informative about long-run living standards than snapshot income measures. Increasing the frequency of measurements may also be helpful in averaging out measurement errors when dealing with such noisy outcomes (McKenzie, 2012).

Finally, we highlight concerns regarding the "file drawer problem" and publication bias in assessing the studies surveyed here. It is possible that studies that delivered null results or less interesting findings are less likely to be published, or even to be written up in the first place. Even within published studies, there may be concerns that outcomes with statistically significant results are emphasized over potentially more meaningful outcomes where impacts are less pronounced. This is a particularly important issue given the latitude that researchers often have in selecting results to report across a range of outcomes or sample sub-groups. In subsequent sections, we discuss the importance of collecting comprehensive follow-up measurement across a wide range of experiments, ideally with prespecified outcomes and statistical tests, which could be expected to deliver a more complete picture about the overall impacts of a particular intervention and of the body of evidence as a whole.

2.2 Long-run Impacts of Child Health Interventions

The literature on the short-term impacts of child health interventions is vast, spanning public health, economics, education, psychology and nutrition. The RCT evidence on long-run economic impacts of health is far more limited. The limited existing evidence finds generally positive impacts of child health interventions on adult productivity.

2.2.1 Deworming

As discussed in Section 1, the more traditional approach to studying long-term impacts is leveraging historical natural experiments for (hopefully) quasi-random variation in treatment. In the deworming case, Bleakley (2007) studied the successful eradication of hookworm disease from the American South and found large positive long-run educational and socio-economic impacts. These results are echoed by experimental evidence in developing countries. The Kenya deworming study (Miguel & Kremer, 2004) evaluates an experiment starting in 1998 that randomized 75 schools into an intervention group of free deworming drug treatment and worm prevention health education, and control groups. The control group schools were phased into deworming treatment 2 to 3 years after the early treatment groups, a larger gap between early treatment and late treatment groups than was observed in the experimental PROGRESA/Oportunidades evaluation, for instance. In the short-run, Miguel & Kremer (2004) estimate increased school participation rates, and reductions in worm infections among those who directly received drugs as well as evidence for treatment externalities, but no significant improvements in students academic or cognitive test scores.

There have since been multiple follow-up survey rounds of the a representative subsample of the deworming sample, in what is called the Kenya Life Panel Survey (KLPS), starting in 2003. These panel (longitudinal) surveys have been characterized by relatively high respondent effective tracking rates[11], of approximately 83.9 percent (among those still alive), with tracking rates balanced across the treatment and control groups. In the second follow-up round (KLPS-2) collected during

[11]The effective tracking rate (ETR) is a function of the regular phase tracking rate (RTR) and intensive phase tracking rate (ITR) as follows: $ETR = RTR + (1 - RTR) \times ITR$.

2007–2009 roughly 10 years after the start of the deworming project, Baird et al. (2016a) find that deworming program beneficiaries showed increased educational attainment, especially among women (women were 25% more likely to have attended secondary school) while labor supply increased among men (men worked 17% more hours each week), with accompanying shifts in labor market specialization. Since the deworming treatment is inexpensive (at less than US$1 per person per year), the authors estimate a large annualized financial internal rate of return of 51.0% when accounting for health spillovers.

There is new evidence of similarly large impacts on economic productivity and living standards in the third KLPS follow-up survey round (KLPS-3), which was collected during 2011–2013, approximately 15 years after the start of the Primary School Deworming Project. Baird et al. (2018) show that respondent tracking rates were similarly high, at 84 percent and once again balanced across treatment and control groups. Treatment group respondents still have higher total earnings, with an average gain of 13%, which once again implies an extremely high rate of return to school-based deworming program spending. This KLPS round also features a detailed consumption expenditure module, which allows for more reliable assessment of household living standards. The data indicate that consumption is also significantly higher in the treatment group, with an average effect of 23%. The gains in both total earnings and consumption are considerably larger among males, echoing results from KLPS-2. Beyond economic productivity and living standards, treatment group beneficiaries are significantly more likely to live in a city than the control group, have improvements in certain home characteristics (including improving flooring and greater likelihood of being connected to the electricity grid), and also show gains in subjective wellbeing, specifically a question that asks about happiness. Taken together, the Kenya deworming project provides evidence of meaningful long-run gains in economic productivity and living standards along multiple dimensions at both 10 and 15 years following the start of the intervention.

While the evidence on the benefits of deworming in the labor market comes primarily from the Miguel & Kremer (2004) sample, there is evidence on deworming's educational and cognitive impacts in a related sample. Ozier (2018)[12] estimates large cognitive gains 10 years after the start of treatment among children who were 0 to 2 years old when the Kenya deworming program was launched and who lived in the catchment area of a treatment school. These children were not directly treated themselves but were in position to benefit from positive within-community externalities generated by mass school-based deworming. Ozier (2018) estimates average test score gains of 0.3 standard deviation units, which is equivalent to roughly half a year of schooling. It is worth noting that the Baird et al. (2016a) sample (who were already enrolled in primary school) do not experience improvements in test scores, which is consistent with the hypothesis that nutritional interventions are particularly effective in improving child cognition in critical early periods. These patterns among two distinct samples across multiple time points taken together indicate that the treatment effects found in Ozier (2018), Miguel & Kremer (2004), Baird et al. (2016a) and Baird et al. (2018) are unlikely to be driven by chance.

As we discuss below, many other studies show that early childhood interventions in utero or before age three can have large positive impacts (Gertler et al., 2014; Hoddinott et al., 2008). Evidence from the Kenya deworming project suggest that health interventions among somewhat older school-aged children can also have sizable long-run impacts on labor market outcomes through a combination of impacts on education, nutrition and health status.

[12] We omit this study from Table 1 because it does not report labor market outcomes.

2.2.2 Nutritional Supplementation

The earliest experimental evidence on long-run returns to child health interventions comes from the well-known INCAP experiment in rural Guatemala. Between 1969 and 1977, two nutritional supplements — a high-protein energy drink versus a low-energy drink devoid of protein — were made available twice daily in the village and randomly assigned to pre-school children in four villages (Hoddinott et al., 2008; Maluccio et al., 2009; Behrman et al., 2009). Researchers find evidence of a 46% gain in adult wages for males who were exposed to the nutritional supplement before 3 years of age (Hoddinott et al., 2008). They also find improved cognitive skills among both men and women (Maluccio et al., 2009), and even some positive inter-generational effects on the nutrition of the female beneficiaries' children up to 35 years later . This is a highly unusual and exceptional data collection effort, and it provides evidence that childhood health and nutrition gains can have large returns in terms of adult labor productivity.

Through the lens of more recent studies, the pioneering INCAP study also has some limitations. First, it has a small effective sample size of just four villages (since the intervention did not vary within villages), and it is unclear if all the existing studies fully account for the intra-cluster correlation of respondent outcomes in their analyses, thus perhaps leading them to overstate the statistical significance of estimated effects. Second, within each village, receipt of the nutritious drink was voluntary, so those who were treated were not a random sample of the population within each village. In this case, the most convincing estimation strategy may be an intention to treat analysis, yet some studies report the direct effects of receiving nutritious drinks on outcomes, potentially introducing selection bias. Finally, sample attrition is a 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% by the time of the 35 year follow-up survey.

A public health study, Prado et al. (2017)[13] follows up on the sample from a more recent experiment, the Supplementation with Multiple Micronutrients Intervention Trial (SUMMIT), which provides maternal supplementation with multiple micronutrients (MMN) or iron and folic acid (IFA) in Indonesia. The MMN intervention provided the same nutrients as IFA, plus various vitamins, zinc, copper, selenium and iodine, which are thought to have benefits for development in utero. The project has a massive sample size of 31,290 women enrolled in the trial during 2001–2004. The authors find that the children (who were 9–12 years old at the time of the follow-up survey) had better cognition and academic achievement if their mothers had been assigned to MMN instead of IFA. This opens up the possibility of longer-term labor market gains, although these are yet to be established in this sample. Unfortunately, as with the INCAP study, sample attrition in the SUMMIT sample is substantial: only 62% of participants were re-enrolled in the follow-up, among which a representative subset of children were selected for cognitive testing.

2.2.3 Cognitive Stimulation

The well-known Jamaica experiment (Gertler et al., 2014) carried out during 1986–1987 provides some of the earliest and most compelling evidence on the long-run benefits of early childhood psychosocial stimulation in a low-income country. The intervention targeted growth-stunted toddlers and consisted of weekly visits over a 2-year period by community health workers who taught parenting skills and ways to interact with children to develop cognitive and socio-emotional skills. The

[13] We omit this study from Table 1 because it does not report labor market outcomes.

authors found that 20 years later, the intervention increased participants' full-time job earnings by a massive 25%. For non-temporary jobs, the gains are even higher, at 48%.

These labor market gains could result from increased parental investments in children, increased schooling, and from migration. At the end of the 2-year intervention, the researchers find that the treatment increased the quality of parental interaction and investment in children, as measured by the HOME inventory (Caldwell et al., 1984). These effects faded out in mid-to-late childhood (at age 7 and 11) but then did ultimately translate to more years of schooling attainment, again illustrating that the absence of effects at one time point does not preclude finding effects later. The authors also find suggestive evidence that the treated group tends to migrate more, and that migrants earned substantially more than those who stayed in Jamaica.

The Jamaica study achieved a fairly low attrition rate of 18.6%, which is much lower than several other early experiments described in this section, including the 40% attrition in the INCAP experiment and 49% in the Turkish Early Enrichment Project (TEEP) discussed below. One important limitation, however, is its modest sample size of 129. Another caveat is that the authors were only able to track 14 out of 23 migrants in the sample, and treatment group individuals were over-represented among the 14 migrants tracked. This differential attrition of migrants across treatment arms could potentially bias treatment effect estimates upward.

Despite the large positive gains to small-scale psychosocial stimulation programs, some efforts to scale up these interventions have been less successful. Andrew et al. (2018) studied a scalable psychosocial stimulation intervention, implemented at larger scale and using the institutional infrastructure of existing government services. Two years after the program ended, they found no effects on child test scores, cognition, behavior, stimulation in the home environment or maternal depressive symptoms. The authors note that it is possible that intervention effects may appear later on, and long-term effects are unknown.

Another early RCT in the psychology literature, the Turkish Early Enrichment Project (Kagitcibasi et al., 2009) provides further evidence on an early childhood stimulation intervention carried out during 1983–1985 among children aged 4–6 from deprived backgrounds. The intervention randomized children into one of three alternative care environments: an educational day care center, a custodial day care center, or the home. Half of the mothers in each care environment were randomly assigned to receive parenting training related to cognitive stimulation. The 22-year follow-up analysis grouped all treatment arms together into "any stimulation" and found that treated participants had more favorable outcomes in terms of educational attainment, occupational status, and integration into modern urban life, such as owning a computer. The effects of the enrichment treatment on consumption were positive but not statistically significant. Two limitations are the high sample attrition rate of 49% mentioned above, as well as the fact that assignment to the different preschool environments was not entirely random, but determined in part by availability at the workplace, possibly leading to some selection bias.

2.2.4 Perinatal Interventions

There are many RCTs involving perinatal interventions in public health, but they have received relatively little attention from economics researchers to date. While it is unusual for public health studies to collect long-run employment and labor market outcomes, Charpak et al. (2016) do so. They study the 20-year impacts of a kangaroo mother care (KMC) intervention[14] in Colombia and

[14] Kangaroo mother care is an intervention designed for preterm and low birth weight infants, consisting of (1) continuous skin-to-skin contact between mother and infant; (2) exclusive breastfeeding when possible; and (3) timely

find that the intervention increased beneficiaries' school attendance, and later wages and labor force participation. However, sample attrition was again substantial, unfortunately: the authors were only able to survey 441 participants (62% of all the original participants), including 264 participants weighing less than 1800g at birth, who were thought to be most likely to gain from the intervention. Another potential methodological concern is the fact that statistical significance levels were not adjusted for multiple hypothesis testing.

The Promotion of Breastfeeding Intervention Trial (PROBIT) in Belarus randomized 31 maternity hospitals and affiliated polyclinics to either the control arm or the intervention, which aimed at increasing breastfeeding duration and exclusivity, during 1996–1997. In a follow-up survey carried out 16 years later, Martin et al. (2017)[15] successfully followed up 79.5% of the 17,046 breastfeeding mother-infant pairs who participated in the original trial. They do not find any effects of the intervention on the obesity or blood pressure levels of the infant beneficiaries (who were young adults at the follow-up survey). However, it remains an open question whether this intervention impacts other health outcomes, or any cognitive and economic outcomes in the long run.

Bhalotra et al. (2017) evaluate an intervention that provided psychotherapy to perinatally depressed mothers in rural Pakistan. The intervention successfully reduced depression at the time. Seven years later, it also increased women's financial empowerment, control over household spending, as well as time- and monetary-intensive parental investments, especially on girls. These investments have the potential to translate into later gains in cognition, education and labor market outcomes, although longer-term effects are unknown.

2.3 Differential Impacts by Gender

Substantial heterogeneity in treatment effects along gender lines is common across several of the interventions that we survey in this article. However, the literature does not seem to converge on whether it is men or women who consistently gain more from the interventions, or on the mechanisms driving the differences. Here we highlight the findings from the long-run studies that we review in Table 1, and call for further research to help reconcile these findings with each other, as well as with predictions from economic theory.

Baird et al. (2016a) and Baird et al. (2018) observe that school deworming treatment effects in both total earnings and consumption are larger in magnitude among males 10 to 15 years after the intervention, although differences are not always statistically significant. In contrast, women who were eligible for deworming as girls are 25% more likely to have attended secondary school, halving the gender gap, and they reallocate time away from traditional agriculture and 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 (Baird et al., 2016a). The authors argue that these results are broadly consistent with the theory of human capital presented in Pitt et al. (2012), in which time allocation depends on how the labor market values both improved human capital and improved raw labor capacity, and this may vary by gender in low-income "brawn-based" economies. In particular, Pitt et al. (2012) present evidence consistent with a model in which exogenous health gains 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.

(early) discharge with close follow-up (Charpak et al., 2016).

[15] We omit this study from Table 1 because it does not report long-run labor market outcomes.

Unlike for primary school deworming, Barham et al. (2017) and Barham et al. (2018) find that a conditional cash transfer program in Nicaragua generated similar effects on earnings and labor market participation for both men and women, and they uncover quite different underlying causal mechanisms that in many ways are the reverse of those identified in the KLPS. Unlike with deworming, both education and learning gains here are concentrated among males, in a context where boys typically drop out of school at younger ages than girls. Women experienced at most modest effects on education and learning, but improved nutrition and reproductive health during teenage years, which the authors argue could translate into labor market gains.

Duflo et al. (2018) study the impacts of secondary school scholarships in rural Ghana and find larger effects on learning and progress to tertiary education among females. In particular they note that the "marginal" males (who were only sent to secondary school because of the scholarship) were much less likely to go on to tertiary education than inframarginal males, while marginal females were just as likely to go on to tertiary education as inframarginal females. They argue that families may typically already send academically promising boys to senior secondary schools even in the absence of scholarships, but that there may be heterogeneity among households in their treatment of girls, with some but not all households sending promising girls to school in the absence of scholarships. The "marginal" girls could therefore have higher underlying ability than similarly marginal boys.

Bettinger et al. (2018) examine the effects of private secondary school scholarships in Colombia, and observe large positive effects on the probability of ever enrolling in tertiary education (including vocational schools and universities), formal credit access, and formal sector earnings, with the strongest scholarship impacts among vocational school applicants, as noted above. Within the vocational sub-population, there are larger effects among males; within the academic sub-population, females seem to benefit more, although differences across gender tend not to be statistically significant.

Several other studies find more positive long-run effects for males. Hoddinott et al. (2008) find that in the INCAP experiment (the nutritional intervention in Guatemala), all effects are concentrated among men and effects for women are typically smaller and not statistically significant. Molina Millán et al. (2018b) find that the conditional cash transfer in Honduras leads to increased international migration for young men, by 3 to 7 percentage points, with the effects being smaller for women. There is also evidence that, among those who received the conditional cash transfers, women, but not men, reduced their labor supply. The authors caution that this does not necessarily imply negative labor market impacts for women, as the beneficiaries are still transitioning from school into the labor market. Finally, Blattman et al. (2018b) find that the treatment effects of an entrepreneurial grant in Uganda have largely faded out after 9 years. However, among the few impacts that persisted, effects on durable asset ownership are higher among men, whereas effects on occupational choice (such as engagement in a skilled trade) are higher among women.

Further theoretical and conceptual work will likely be needed to make sense of these findings by gender, and additional empirical research will be important to understand which patterns are robust across settings. It will be particularly useful to follow effects over a longer time period, and to relate any differences to patterns of marriage, fertility, and female labor force participation across study environments, as well as to patterns of occupational segregation and gender wage gaps.

3 WHAT CAN WE LEARN? OPPORTUNITIES AND LIMITATIONS

The large number of experimental cash transfer and child health studies conducted during the late 1990's and the 2000's provide an opportunity to conduct long-term follow-up studies, as described in Section 1. But how feasible is this opportunity in practice? In this section, we systematically survey and evaluate the opportunities and limitations of the existing pool of cash transfer and child health RCT studies.[16]

3.1 Cash Transfers

We focus here on unconditional or conditional cash transfer experimental studies that examine impacts on either the living standards or economic productivity of individuals and households.

3.1.1 Study Screening Criteria

Appendix A provides a detailed description of the screening procedure and justifications for our selection criteria. Study selection was based on six main criteria, namely, for inclusion a study had to:

1. have randomized treatment,
2. have been implemented before 2010 (to allow for long-run follow-up),
3. have sufficient statistical power (and relatedly, a sufficiently large sample size),
4. be properly implemented (in ways we make precise in Appendix A),
5. have sufficient differential exposure to the intervention across treatment arms,
6. and have the potential for a reasonably high respondent tracking rate.

Among the 170 publications extracted from the seven meta-analysis studies identified during our review, 18 cash transfer studies appear eligible for long-term follow-up research (see Table 2). If we additionally exclude the six studies that have already benefited from a long-term follow-up of labor market outcomes, 12 studies appear to be particularly promising for new long-term studies. We think of these 12 studies as "low-hanging fruit" for the research community. Yet, the fact that the majority of existing cash transfer RCTs end up being excluded due to important design or data limitations also indicates that many past experiments have, unfortunately, not been set up to allow for longer-term evaluation. In Section 4 below, we discuss several approaches that could improve this yield rate for future experiments.

[16]The overall screening strategy was carried out as part of the Long-term Impact Discovery (LID) project financed by GiveWell and co-chaired by Prashant Bharawadj (UCSD) and Craig McIntosh (UCSD). We thank both of them for their leadership in the project and their crucial intellectual contribution to this section of the paper. The LID project does not focus on education interventions, but in our view there are also likely to be abundant opportunities for conducting long-term impact evaluations in education given the large number of education RCTs. Assessing the existing pool of education RCTs is beyond the scope of this article.

3.1.2 Eligible Studies

Table 2 describes the 18 RCTs that meet all the selection criteria and are considered attractive for conducting a long-term follow-up study. Among these experiments, four (denoted by the acronyms AAC, NCTPP, SCAE and ZOMBA, see Table 2 for references and details) present particularly favorable features: all had interventions that were well implemented; none featured a phase-in design; and no long-term follow-up survey has yet been conducted. Two of these RCTs feature both an unconditional cash transfer study arm and a conditional cash transfer arm (namely, NCTPP, and ZOMBA), presenting a particularly fruitful setting for comparing the long-run impacts of CCTs and UCTs.

The table also provides two important pieces of information about the selected studies that may guide future decisions regarding whether or not to conduct a long-term follow-up. First, in the column "Phase-in Design" we document whether the original control group subsequently received treatment. Although phase-in studies with sufficient time lag between early treatment and late treatment groups should not be excluded *a priori*, following up on phase-in studies with a relatively short lag presents some challenges for both estimation and interpretation. We discuss this issue in Section 4.1.

Second, we also report on the short-term impacts of each intervention on the living standards, education, health and labor market outcomes of household adults (see Table 2, column "Short-Term Impacts"). Conducting follow-up surveys just for studies with large and positive short-term impacts may be tempting, and may even be justified at times, yet focusing solely on these studies can have several undesirable consequences. First, cherry-picking only the most "favorable" studies for follow-up surveys will generate a set of estimated long-term impacts that may be representative of studies that yielded short-run impacts, but would be unrepresentative of the set of studies as a whole. For scientific progress, it would be more useful to conduct follow-up studies for multiple RCTs in this table, perhaps in a coordinated fashion (with common survey instruments, etc.) in order to create a more complete picture of long-run impacts.

Table 2: Selected Cash Transfer Studies for Potential Long-term Follow-up

	Study Acronym	Country	Type	Start of Inter-vention	Phase-in Design[a]	Already Followed-Up (> 5 years)	Short-Term Impacts			
							Economic	Education	Health	Adult Labor Market
1	PROGRESA (Behrman et al., 2005)	Mexico	CCT	1998	yes	yes	+	+	+	0
2	PRAF II (Galiani & McEwan, 2013)	Honduras	UCT	2000	no	no		+		0
3	RPS (Maluccio & Flores, 2005)	Nicaragua	CCT	2000	yes	yes	+	+	+	−
4	BDH (Paxson & Schady, 2010)	Ecuador	UCT	2003	yes	yes		+	+	
5	PAL (Cunha, 2014)	Mexico	UCT	2003	no	no	+		+	
6	SCAE (Barrera-Osorio et al., 2011)	Colombia	CCT	2005	no	yes	+	+		
7	AAC (Macours et al., 2012)	Nicaragua	CCT	2005	no	no	+	+	+	
8	YOP (Blattman et al., 2013)	Uganda	UCT	2006	no	yes				
9	MDICP (Kohler & Thornton, 2011)	Malawi	CCT	2006	no	no			0	
10	BRAC TUP (Bandiera et al., 2017)	Bangladesh	UCT	2007	yes	yes	+			+
11	NCTPP (Akresh et al., 2016)	Burkina Faso	Both	2008	no	no	+	+	+	
12	TASSYR (Benhassine et al., 2015)	Morocco	Both	2008	yes	no	+	+		
13	ZOMBA (Baird et al., 2011)	Malawi	Both	2008	no	no	+	+		
14	Women Plus (Green et al., 2015)	Uganda	UCT	2009	yes	no	+			+
15	Respect (Damien de Walque, 2014)	Tanzania	CCT	2009	no	no			+	
16	CGP Zambia (Natali et al., 2016)	Zambia	UCT	2010	yes	no	+			+
17	TASAF (Evans et al., 2014)	Tanzania	CCT	2010	yes	no	+	+	+	
18	BONO (Benedetti et al., 2016)	Honduras	CCT	2010	yes	no	+	+	+	+

UCT: Unconditional Cash Transfer; CCT: Conditional Cash Transfer.

+ indicates significant and positive effects, − indicates significant and negative effects, 0 indicates non-significant effects.

[a] For cases where it is unclear whether there is a phase-in design, we write "no" here, but more precisely, this means not to our knowledge.

If one were confident that studies which yielded no short-run impact also had no long-run impact or that, for example, effects fade out monotonically over time, one might be able to recover estimates or bounds on long-run impacts more broadly. However, as noted above, there is evidence that the effects of certain development interventions can "re-surface" in the long-run even after an apparent fade-out of short-run impacts.[17] The mechanisms underlying this phenomenon are still not well understood. One possibility is that the short- and medium-run surveys fail to adequately capture competencies, such as individual socio-emotional skills or job referral networks, that may eventually generate positive impacts. Consequently, failing to follow up samples in which short- to medium-run impacts are modest (or non-existent) may lead us to erroneously conclude interventions were unsuccessful when in fact they do improve long-run living standards. A bottom line lesson is that a wide and representative range of studies should be evaluated for long-run impacts, and studies should not be ruled out for long-term follow-up because they do not find economically meaningful or statistically significant short-term impacts.

3.2 Child Health Interventions

The child health literature is even more expansive than the body of cash transfer studies, and its boundaries less clearly defined. We consider studies that aim to improve the overall health of a child from in utero through adolescence.[18] Our criteria include physical health interventions, as well as psychological stimulation and preschool age child development interventions. The selection criteria does not include education studies beyond preschool unless the intervention specifically included a health component.[19]

3.2.1 Study Screening Criteria

We implement a strategy similar to that employed in the cash transfer literature to identify existing child health RCTs that could potentially benefit from a long-term follow-up study. We identified a total of 378 publications and, based on the same criteria used for cash transfers, restrict the selection to 77 eligible studies; details are provided in Appendix B. As indicated in Appendix B, these studies are grouped into five main categories (namely, nutrition, perinatal, sanitation, specific diseases, and stimulation). Studies in the nutrition literature, listed in Table 3, represent the largest group (32 studies), while the other categories, as shown in Table B.1, include 45 studies. In the rest of this section, we focus, for reasons of space, on studies in the nutrition literature, and leave a detailed discussion of other categories for Appendix B.

[17] For instance, Gertler et al. (2014) report a lack of medium-run impacts, and Banerjee et al. (2016) report effects that grow over time. Deming (2009) and Chetty et al. (2011a) show that the Head Start program and the Tennessee STAR experiment in the U.S. improved participant outcomes in adulthood, despite initial "fade-out" of test score gains.

[18] We focus on interventions that address a public health issue and affect a meaningful proportion of children. For instance, stunting is estimated to impact 24.3% of the children under 5 for less developed regions (Unicef et al., 2018) and the prevalence of malaria is estimated at 9.13% for low Socio-Demographic Index regions in 2017 (Institute for Health Metrics and Evaluation, 2018). Interventions that aim to address specific syndromes or diseases (such as genetic defects which affect very small proportions of newborns) were thus excluded from our review. More details on the inclusion criteria are provided in Appendix B.

[19] As noted above, there is a large pool of education RCTs in development economics but assessing their suitability for long-run follow-up impact evaluation is beyond the scope of this article.

3.2.2 Eligible Nutrition Studies

Studies of two interventions clearly stand out in Table 3: vitamin A and mixed supplementation studies. Mixed supplementation includes both Multiple Micronutrient (MMN) supplementation and Lipid based Nutrient Supplements (LNS), as we discuss in more detail below.

Since the mid-1980's, vitamin A interventions have attracted considerable attention among nutritionists. A seminal study by Sommer et al. (1986a) among 480 villages in Indonesia suggested that vitamin A supplementation could be a highly effective strategy for reducing mortality (-34%). Since the mid 1980's, multiple RCTs (Sommer, 2008) confirmed the positive effects of vitamin A, with an effect size varying between -50% to -34% (though 2 out of 16 studies found no significant impact on health, see Table 3). While a more recent and large-scale study has led to some questions regarding these magnitudes (Awasthi et al., 2013a), there remains a broad consensus that vitamin A delivered to vitamin A deficient children or pregnant woman is likely to be an effective strategy for reducing mortality.

Yet, how these early health benefits translate into subsequent motor, cognitive ability, or long-run economic productivity impacts in long-run remains almost entirely unknown, as no such long-run studies based on experimental data exist (to our knowledge). This appears to be a promising area for future research. Table 3 provides some additional information on the vitamin A studies that could feasibly be followed up today. The data presented in the table appears to confirm that vitamin A's short-term impact on health outcomes, and particularly child mortality, is positive overall.

It is possible that an intervention that affects mortality could pose methodological problems for researchers examining long-run outcomes, due to the possibility of selection (or "survivorship") bias. Yet we do not believe this would be a major concern in practice. In the Indonesia data in Sommer et al. (1986a), for instance, mortality amounts to only 1% of the total attrition and 0.2% of the differential attrition in a 13-month follow-up. Thus we do not believe that concerns about differential mortality across treatment arms should deter researchers from following up on populations that took part in vitamin A RCTs.

Another potential methodological challenge posed by the nutritional supplementation RCTs is imperfect compliance in the control group: due to ethical concerns, in certain trials project health staff examined control group participants and opted to provide treatment to control group children with severe nutritional problems. This practice makes estimated treatment effects challenging to interpret, and seems likely to dampen estimated effects. We flag studies that follow this approach in Table 3 and Appendix Table B.1 (see notes).

Mixed supplementation interventions (namely, MMN and LNS) constitute the second largest group of nutrition studies that we identified, as listed in Table 3. Widespread research interest in MMN appears to be more recent than for vitamin A, with most studies dating back only to the mid-1990's. Many of these studies found short-run evidence that MMN supplementation, distributed early on, positively impacts child motor and cognitive development (Eilander et al., 2009). Prado et al. (2017) even report positive medium-term impacts on cognition at age 9–12, but to our knowledge, the impact of MNM on long-run living standards and labor market outcomes has never been estimated with experimental data, creating another promising opportunity for research. The Lipid-based Nutrient literature is even more recent (starting in the early 2000's). Most studies estimate large positive short-run impacts of such interventions, with gains even larger than those found for MMN interventions (Matias et al., 2017).

Table 3: Selected Child Nutrition Studies for Potential Long-term Follow-up

	Study	Country	Description	Start of Inter-vention	Clustered RCT	Sample Size[a]	Age of Children	Short-Term Impacts: Health	Short-Term Impacts: Cognition
1	Sommer et al. (1986b)	Indonesia	VA	1983	yes	450	12–71 mo	+	
2	Rahmathullah et al. (1990)[b]	India	VA	1985	yes	206	6–60 mo	+	
3	Vijayaraghavan et al. (1990)	India	VA	1987	yes	84	1–5 y	0	
4	Herrera et al. (1992)[b]	Sudan	VA	1988	no	28,753	9–72 mo	+	
5	Stansfield et al. (1993)	Haiti	VA	1988	no	11124	6–83 mo	−	
6	Dibley et al. (1996)[b]	Indonesia	VA	1989	no	1,407	6–47 mo	+/−	
7	Ross et al. (1993) (VAST)[b]	Ghana	VA	1989	yes	185	6–90 mo	+	
8	Barreto et al. (1994)	Brazil	VA	1990	no	1240	6–48 mo	+	
9	West Jr et al. (1991)[b]	Nepal	VA	1991	yes	261	6–72 mo	+	
10	Shankar et al. (1999)	Papua NG	VA	1995	no	480	6–60 mo	+	
11	Jinabhai et al. (2001)	South Africa	VA & Deworming	1995	no	579	8–10 y	+	0
12	Sempértegui et al. (1999)[b]	Ecuador	VA	1996	no	400	6–36 mo	+	
13	Lind et al. (2004)	Indonesia	Iron & Zinc	1997	no	680	6–12 mo	+	+
14	Rahman et al. (2001)[b]	Bangladesh	VA & Zinc	1997	no	800	12–35 mo	+/−	
15	Solon et al. (2003)	Philipines	MMN & Deworming	1998	no	831	6–14 y	+	+
16	Sivakumar et al. (2006)	India	MMN	1999	yes	20	6–16 y	+	+
17	Awasthi et al. (2013b) (DEVTA)	India	VA & Deworming	1999	yes	72	6–72 m	+	
18	Group (2008) (SUMMIT)	Indonesia	MMN	2001	yes	262	in utero	+	
19	Manger et al. (2008)	Thailand	MMN	2002	no	569	5–13 y	+	+
20	Faber et al. (2005)	South Africa	MMN	2002	no	361	6–12 mo		+
21	Sazawal et al. (2010)	India	MMN	2002	no	1,257	1–4 y	+	
22	NEMO Study Group (2007) (NEMO)	Indonesia	MMN & Fatty Acid	2003	no	384	6–10 y		+
23	Long et al. (2006)	Mexico	VA & Zinc	< 2005[c]	no	736	6–15 mo	+/−	
24	Aboud et al. (2009)	Bangladesh	Responsive Feeding	2007	yes	37	8–20 mo	0	+
25	Suchdev et al. (2012)[b]	Kenya	MMN	2007	yes	60	6–35 mo	+	
26	Aboud & Akhter (2011)	Bangladesh	MMN & Responsive Feeding	2008	yes	45	8–20 mo	+	+

Table 3: Selection of Child Nutrition Studies for Long-term Follow-up (Continued)

	Study	Country	Description	Start of Inter-vention	Clustered RCT	Sample Size[a]	Age of Children	Short-Term Impacts Health	Cognition
27	Veenemans et al. (2011)	Tanzania	Zinc & MMN	2008	no	612	6–60 mo	0	
28	Maleta et al. (2015) (iLiNS-DOSE)[b]	Malawi	LNS & Milk	2009	no	1,932	5–7 mo	0	
29	Aboud et al. (2009) (iLiNS-DYAD)	Ghana	MMN & LNS	2009	no	1,320	in utero	+	
30	Attanasio et al. (2014)	Colombia	Stimulation & MMN	2010	yes	96	12–24 mo	0	+
31	Hess et al. (2017) (iLiNS-BF)	Burkina Faso	MMN	2010	yes	36	6–27 mo	+	
32	Mazumder et al. (2015) (Neovita)	India	VA	2010	no	44984	newborn	0	

VA: Vitamin A; MMN: Multiple Micro-Nutrient; LNS: Lipid-based Nutrient Supplement. mo: month; y: year.
+ indicates significant and positive effects, − indicates significant and negative effects, 0 indicates non-significant effects. +/− indicates coexistence of significant positive and negative effects (including side effects).

[a] We report the number of clusters for clustered RCTs, and the number of households or individuals for non-clustered RCTs.

[b] In these RCTs, participants in treatment and control arms are regularly examined during the trial, and those with severe conditions (e.g., severe Vitamin A deficiency) are then treated; this practice may change the interpretation of estimated treatment effects.

[c] The authors did not mention when the intervention was conducted, but we infer that it was before 2005 when the paper was submitted.

Although some appear promising, the bulk of MMN and LNS RCTs are still too recent for a long-term follow-up on economic outcomes and are thus excluded from Table 3. However, many will become viable candidates in the coming years.

4 HOW CAN WE DO BETTER? RESEARCH DESIGN AND DATA

This section contains a discussion of the following two questions: (1) how can researchers most effectively assess the long-run impacts of an intervention that has already been conducted, and (2) how can researchers design experiments and data collection to improve the feasibility of studying long-run impacts? We discuss these two intertwined issues in the context of both research design, as well as data collection and usage.

4.1 Research Design

The most important building block of a randomized controlled trial is the experimental design. One type of design that is common in field experiments in economics, especially among those that we review in this article, is the phase-in design.[20] A phase-in design is where treatment groups first receive the interventions, and then "control" groups receive the same interventions later.

A phase-in design ensures greater similarity across the treatment and comparison group in the (eventual) distribution of assistance, arguably relaxing some ethical concerns, and may also increase the local political acceptability of a project. It is also a natural design choice when real-world programs are being piloted or gradually rolled out: randomizing the order of program expansion generates treatment and control groups. These experiments include many of the earliest and most influential studies in development economics, including some of those that have already carried out long-run follow-ups. Examples of phase-in designs include the prominent PROGRESA/Oportunidades experiment (Parker & Todd, 2017), the deworming program in Kenya (Miguel & Kremer, 2004) and the "graduation" programs (Bandiera et al., 2017).

One might be tempted to exclude phase-in experiments when trying to learn about long-term impacts, due to concerns that long-term effects are not identifiable when there are no pure control groups left. However, we demonstrate that under certain assumptions detailed below, it is possible to identify long-term treatment effects in the presence of a phase-in design, as long as measurements are taken sufficiently frequently. We also show that the variance of treatment effect estimates will grow linearly over time, at a rate that varies inversely to the difference in the duration of treatment between the treatment and control groups (as denoted by T below).

Consider a setup similar to that of Borusyak & Jaravel (2016), in which a panel of units (individuals or clusters) $i = 1, \cdots, 2N$ are randomized into two (equally sized) groups $j = 0, 1$, which are the control group (or "late treatment" group) and the treatment group (or "early treatment" group), respectively. Suppose that the treatment group receives the treatment at period $T_0 = 0$ and the intervention is phased into the control group after $T_1 = T$ periods. First consider a situation in which the outcomes Y_{ijt} are observed $(K + 1)$ times, at $t = 0, T, \cdots, KT$ ("calendar time"). Following the event-study notation, denote "relative time" to treatment $K_{jt} = t - T_j$[21]. This denotes

[20] Phase-in RCT designs — also called stepped-wedge designs — appear to be less common in the health research literature.

[21] The setup here ensures that for both the treatment and the control group, $K_{jt} \in \{0, T, \cdots, KT\}$ and take the

the amount of time group j has already been exposed to the intervention at time t. We specify the data-generating process to be

$$Y_{ijt} = \alpha_t + \sum_{k=0}^{K} \tau_k \mathbf{1}\{K_{jt} = kT\} + \epsilon_{ijt} \tag{1}$$

and make the following assumptions:
Assumption 1, Stable Dynamic Effects The pattern of dynamic treatment effects (the τ_k terms) is the same in the treatment group and the control group. This holds if the dynamic treatment effects do not interact with (calendar) time, for example.
Assumption 2, Validity of Randomization Absent the intervention, the outcomes of the units in the treatment and control groups follow the same trends.
Assumption 3, Stable Unit Treatment Value Assumption The intervention on the treatment units does not have effects on the outcomes of the control units.

The first assumption is standard for event-study designs (Borusyak & Jaravel, 2016), but it is somewhat restrictive, as discussed below. The last two are standard assumptions for most RCTs.

Note that this flexible setup imposes minimal assumptions on the dynamics of the treatment effects. The treatment effects can be increasing or decreasing over time, and can even reverse signs after a certain period.

Under the three assumptions described above, and most importantly, the stable dynamic effects assumption, the difference between treatment and control groups before the program rolls-out to the control group identifies the effects of the program in the first T periods. These estimates can then be used to compute the counterfactual of the control group (if they had been left untreated) to back out long-term impacts after full program roll-out. The long-term effects at $t = 2T$, for example, would be the sum of the difference between treatment and control groups at $t = T$ and at $t = 2T$; in other words, the counterfactual outcome for the control group at $t = 2T$ is simply its actual value minus the estimated effect T periods after the treatment, which is imply the difference between the treatment and control groups at time $t = T$. With the same logic, the long-term effects at $t = 3T$ would be the sum of the difference between treatment and control groups at $t = T$, at $t = 2T$, and at $t = 3T$. One can extend this to $t = KT$, the completed period for which we have measurements of the outcomes, although intuitively, summing up these treatment effect estimates will lead to larger standard errors as t grows.

An important result is that sufficiently frequent measurement is essential. Identification is possible only if the measurements are carried out at least every T periods, otherwise one simply cannot identify the effects in the initial few periods, and cannot compute longer term effects using the approach described above. However, in the case where the initial measurement is done after phase-in of the control group, if we are willing to make the assumption that the effects of additional exposure is non-negative, the difference between treatment and control groups provides a lower bound of the true treatment effect.

When we run the regression of the form

$$Y_{ijt} = \alpha_t + \sum_{k=0}^{K} \hat{\tau}_k \mathbf{1}\{K_{jt} = kT\} + \epsilon_{ijt} \tag{2}$$

same set of values. When this is not the case, interpolation is necessary, as is the case for Bandiera et al. (2017). They took measurements in year 2, 4 and 7, and their control group was treated in year 4. They interpolate between 2- and 4-year estimates of effects for the treatment group to derive a counterfactual 3-year effect for the control group, in order to estimate treatment effects after 7 years.

we recover the $\hat{\tau}_k$'s. The variance of these estimators is (under standard assumptions, in particular, homoskedasticity)

$$\mathrm{Var}(\hat{\tau}_k) = \frac{2}{N}(k+1)\sigma^2 \tag{3}$$

where σ is the residual standard error of the regression[22]. It is clear that the variance of treatment effect estimates grows linearly over time (namely, as observations are farther from the time of control group phase-in), as opposed to staying relatively stable over time, as would be the case for a non-phase-in RCT design.

Despite reduced precision for (absolute) long-term estimates in a phase-in design, this approach actually yields more precise estimates for the differential effects. These estimates may be of particular interest if one is interested in testing certain hypotheses, such as whether effects grow or fade out over time. This is because these estimates are taken directly from comparing the treatment and control groups at a point in time, and are not computed by summing up or differencing multiple estimates. For example, suppose we want to know whether the treatment effect after T periods is the same as the effect after $2T$. In a standard non-phase-in RCT design, one would have to test the equality between the treatment effect estimates in $t = 2T$ versus $t = T$. In a phase-in design, however, one can take the treatment effect estimate at $t = 2T$ directly, yielding more precise estimates than the former method.

Bandiera et al. (2017) employ a related approach. They evaluate an intervention that was phased in to control groups after four years, and compute a range of estimates for the treatment effects after seven years. Rather than calculating standard errors using an analogue of the procedure above, they check for robustness by using the 25th, 50th and 75th percentile Quantile Treatment Effect estimates on the 3-year effects to create counterfactuals for the phased-in controls 7 years after the program had started. As they measured 2-year and 4-year effects in practice, they need to impose some additional assumptions for interpolation to get the 3-year effects. Adjusting the standard errors with our calculation above leads to somewhat wider confidence intervals than with their approach; one can reject the hypothesis of no long-term effects, however, so the results remain robust under the approach outlined in this article. Note that the phase-in design allows them to demonstrate that effects are in fact increasing over time, even though standard errors on the 7 year effect are fairly large.

While the economics literature generally assumes that the path of dynamic effects does not vary with time (Borusyak & Jaravel, 2016), in many contexts, the dynamic path of treatment effects would vary with either the age of participants, or other factors that are time-varying, such as the prevalence and intensity of a disease. Identification of long-run effects will still be possible if there is a sufficient sample size and sufficient variation in child age (or prevalence and intensity of a disease) among the treatment and control samples, to separately identify the dynamic path of effects for children of different ages (or in contexts with different prevalence and intensity of the disease). However, this would be impossible if time and age (or prevalence and intensity) are perfectly correlated. For example, if all the treatment and control group individuals are 4 years old at $t = 0$ when the treatment group receives a health intervention, and if the control group receives the intervention at $t = 3$ (three years later), to estimate long-term effects, we would have to make the perhaps implausible assumption that effects on 4-year-olds are the same as effects on 7-year-olds.

[22] The derivation is shown in Appendix C.

While we show that long-run effects may be econometrically identified even with phase-in designs, they will at best be estimated with more noise, and so our view is that experimental research designs with pure control groups are generally preferable to phase-in designs, when it is ethically and politically feasible to use them.

The other basic building block of a randomized controlled trial is an adequate sample size. However, many trials are underpowered to detect modest yet economically meaningful treatment effects, partly because researchers often face a trade-off between the number of treatment arms and statistical power. Croke et al. (2016), for example, showed that out of the 22 studies that estimate the impacts of mass deworming, the median sample size for non-clustered RCTs is only 198 individuals, and the median sample size for clustered RCTs is 80 clusters. For assessing long-term impacts, concerns about power are particularly relevant, because sample attrition may further erode statistical power. One may combine data from individual papers and conduct meta-analyses in order to gain more statistical power and make progress in this area. Study sample size plays a role in our selection criteria, as described in Section 4, and Appendix A and B.

4.2 Data

4.2.1 Follow-up Surveys

Table 1 illustrates that follow-up surveys are the most common source of data used to conduct long-run evaluations of RCTs in international development. The choice to use survey data appears to often be made out of necessity: in most low-income countries, relevant administrative data at the individual level is either non-existent or difficult to obtain. Even when they exist and are accessible, administrative records may only capture a small share of the outcomes of interest to development economists. For instance, few low-income countries rigorously measure informal economic activity, self-employment earnings, or subsistence agricultural production, and even when they do, data may only exist for a small subset of the population (which may not overlap with the population studied in an RCT). It is not a coincidence that the rise in field experiments and original survey data collection in development economics have gone hand in hand over the past twenty years.

Individual or household level surveys have many strengths, but also key limitations. The most important upside of original survey data is the researcher's ability to design her own questions to effectively answer the question at hand. Many recent household surveys in development economics collect highly detailed measures of demographic, educational, health, psychological, and labor market and enterprise outcomes. The richness of original survey data, and the fact that questions can be tailored to particular study goals, allows researchers to probe the mechanisms underlying any intervention impacts, and explore heterogeneity in treatment effects across subgroups. It has become a rite of passage for young development economists to spend extended periods of time in the field designing and piloting survey questions, improving the implementation of data collection processes, and sitting in on countless surveys with trained enumerators. In our view, a positive byproduct of these real-world experiences is often a better understanding of the study setting.

Two frequent downsides of original survey data collection are cost and attrition. Relative to the cost of simply downloading existing administrative records, original survey data collection of thousands of respondents is extremely expensive, with typical project budgets running into the hundreds of thousands of dollars. (Of course, downloading relevant administrative data is usually simply not an option in development economics.) Second, follow-up surveys often suffer from considerable sample attrition. As illustrated in Table 1, several prominent existing long-run follow-ups feature high attrition rates, including 40% in the INCAP nutritional supplement study, 49% in the TEEP cognitive stimulation study and nearly 40% in Progresa (Behrman et al., 2011). Sample attrition appears to be particularly severe in settings where migration — both domestic and international — is common, and among adolescent and young adult populations that are particularly mobile geographically as they seek out educational, labor market and family opportunities.

Fortunately, several more recent long-term tracking efforts, such as the Indonesia Family Life Survey (IFLS) (Strauss et al., 2016; Thomas et al., 2001, 2012), the Kenya Life Panel Survey (KLPS) and the Ghana study mentioned above (Duflo et al., 2018) report much lower sample attrition rates. These surveys all devote considerable resources to tracking and re-contacting original participants, which is critical for reducing non-random attrition and improving data quality. To illustrate, the IFLS5 round tracked 92% of the original households after 21 years. This is despite the high geographic mobility of the baseline respondents: in the fourth wave in 2007, over one-third had moved from the community in which they were interviewed at baseline. For KLPS3, the effective tracking rate is 84% after 15 years and is not significantly different between the deworming treatment and control groups (Baird et al., 2018). Encouragingly, Table 1 indicates that several recent studies have even higher survey respondent tracking rates over periods of roughly a decade.

How have these projects improved long-term tracking and achieved such low attrition rates? In the next few paragraphs, we document several key lessons from the pioneering IFLS project (Thomas et al., 2001, 2012). Several of the authors of the current article also have first-hand experience in respondent tracking from KLPS and the Ghana study, and it is also worth stating several lessons that we have learned along the way (Baird et al., 2008).

A first key lesson is that the detailed contact information of the respondent, as well as of their close relatives and neighbors, should be collected as early as possible in the data collection effort. Starting from the first wave, IFLS began collecting the current residential locations of all households, a sketch map with landmarks and a description of how to find the location, landline and mobile phone numbers, email addresses, people who would likely know their whereabouts in the future and their contact information, whether respondents are planning to move and the likely destinations, and so on (Thomas et al., 2012). When tracking respondents, a field team needs as many "leads" as possible. By the time several years have passed since an intervention started, it may simply be too late to gather this type of data on respondents who are already on the move. Similarly, it is important to renew contact with respondents relatively frequently — in our experience, at least every few years — to prevent residential location information from becoming stale.

A second observation is that respondent tracking has become considerably easier over the past decade or so in many low-income countries as mobile phone penetration has expanded, becoming nearly universal in many societies. At the start of early KLPS follow-up rounds (approximately 15 years ago), launching a tracking round meant revisiting the original villages and schools of the school deworming project; today, a follow-up round is launched with a barrage of cell phone calls and texts to respondents and their relatives, to figure out if they have moved and to set up in-person interviews. In the Ghana study, the research team even provided cell phones to respondents at baseline to facilitate later follow-up contacts (although this step may become unnecessary over time as larger shares of individuals own mobile phones). The cost savings and logistical gains for researchers generated by new communication technologies have been immense.

Third, we have observed that respondent tracking in KLPS actually appears to become somewhat easier as respondents age out of their 20's and into their 30's, as many individuals appear to settle into more stable family, work, and residential arrangements. If a panel survey data collection effort can "get through" the more difficult adolescent and young adult period unscathed, there is hope for more consistently high tracking rates in midlife and beyond.

Fourth, in many low-income countries, including Kenya, there is substantial mobility across national borders. The KLPS project has always had a policy of tracking respondents who move internationally, via phone or Skype surveys, if necessary, in order to limit attrition. While the costs of international tracking can be substantial, it is critical for successful long-run follow-up surveys in many settings. We note that the KLPS survey was launched in a Kenyan region that features a fairly open border with Uganda (and strong family, ethnic and historic ties across the border), which greatly facilitates both international mobility and international tracking; the situation along other borders may be more challenging, for instance, currently when it comes to the case of Mexican and Central American migrants who have moved to the U.S.

Finally, the IFLS research team documents many differences between "movers" and "stayers", including in exhibiting significantly different observed returns to education in IFLS4 (in 2007) (Thomas et al., 2012). This indicates that treatment effect estimates generated in samples that exclude "movers" could be biased. Similarly, in the KLPS-3 analysis described above, deworming treatment has substantial positive long-run impacts on the likelihood of urban migration, which suggests that excluding the subsample of movers from the analysis could again lead to bias. Taken together, investing in tracking study respondents across space will likely be valuable for most long-run research projects.

4.2.2 Administrative Data

Administrative data can be a highly cost-effective alternative to follow-up surveys, in cases where relevant administrative data are available and the baseline surveys contain information that allows them to match to offical records (for instance, a government ID number). Bettinger et al. (2018), for example, achieved very high tracking rates among PACES school voucher lottery participants in Colombia, with 97% of participant identification numbers being valid. These individuals can then be matched to five distinct government administrative datasets with minimal attrition, and no need for costly follow-up surveys since the government is already collecting this data. For labor market outcomes, the authors are able to match roughly participants to formal sector earnings and tax payment records in the 2008–2014 SISPRO dataset (from Colombia's Social Protection Ministry) as well as with Familias en Accion conditional cash transfer and other social protection program eligibility information in the SISBEN survey. For those living in low-income neighborhoods they also obtain self-reported earnings. The administrative data approach in Bettinger et al. (2018) is extremely cost-effective and yields a rich set of outcome data.

It seems clear to us that administrative data should be used when high-quality information on relevant measures is available and can be matched to study participants; the key constraint is that this has rarely been the case in practice, and is especially rare in the poorest developing countries (note that Colombia is a middle income country). In assessing the feasibility of additional long-run follow-up projects, researchers could consider the presence of good administrative data, such as in Colombia, as an important criterion.

When there are no unique identifiers in place to help researchers match records in different datasets, using "probabilistic matching" techniques — matching on individual characteristics such as names, neighborhoods, addresses, birth places, birth dates, etc. can be an attractive alternative.

In Venezuela, Hsieh et al. (2011) matched the list of petition signers who opposed the Hugo Chávez regime to household survey respondents in order to estimate the economic effects of being identified as a Chávez political opponent. Even without an official ID number, the authors successfully matched most records based on locality, exact birth date, and gender.

Yet administrative data also have some drawbacks. As mentioned above, administrative records will typically not contain all of the outcomes or measures that researchers are interested in. Where subsistence agriculture and informal sector economic activities are widespread, as in many low-income countries, administrative data will likely miss important components of total household earnings. To some extent this concern can be ameliorated if there exist proxy means-tested programs (for poverty alleviation), with accompanying administrative records, but the surveys that go into determining eligibility may only be collected infrequently or cover limited geographic areas.

Similar strategies have been pursued in another Latin American environment by Molina Millán et al. (2018b). They evaluate PRAF-II, a conditional cash transfer program in Honduras, using microdata from the national population census and repeated cross-sectional surveys collected more than a decade after the program's start. They assign individual program treatment status based on their municipality of birth, in what is essentially an intention to treat design, given that municipalities were the unit of randomization. However, administrative data here again has some limits: the use of aggregated municipal level data can lead to risk of bias if there is extensive migration and asymmetric mobility across treatment and control areas, for instance.

4.2.3 New Data Sources

An emerging body of studies has leveraged new data and methods from economics, computer science (specifically machine learning), and earth sciences to measure poverty, and these have some promise. In principle, these methods could offer cost-effective and scalable ways to evaluate international development interventions in a timely manner, especially in cases where original data collection is challenging, such as societies experiencing armed conflict. The key caveat to most of these methods is that they are limited in terms of the outcomes that researchers can examine, falling far short of the richness of found in most original development economics household surveys in their measurement of living standards, consumption, and income, and they typically have nothing to say about economically important attitudes, beliefs and expectations, let alone direct health or nutritional measures.

An early application of new data to estimate RCT impacts is the Alix-Garcia et al. (2013) study, which uses Landsat satellite data to study the ecological consequences of the Mexican PROGRESA/Oportunidades program. Remote sensing data appears particularly well-suited to study impacts of cluster-randomized interventions (like this Mexico RCT), where treatment and control geographic areas can be easily identified. Researchers may not always have adequate resolution, or relevant geolocation data, to identify treatment and control households when randomization is done within a village. However, there are exceptions: Burke & Lobell (2017) combined high-resolution satellite imagery (1m Terra Bella imagery) and intensive field sampling on thousands of smallholder maize fields over two years, and they detected positive crop yield responses to fertilizer and hybrid seed inputs; see also Jean et al. (2016). Satellite data has also been used to generate nightlight intensity measures, which have recently become very widely used to proxy for overall local economic activity (Henderson et al., 2012), and once again these could be useful for the evaluation of RCTs where the unit of randomization is fairly large.

More recently, researchers are leveraging cell phone records to assess poverty. The seminal paper by Blumenstock et al. (2015) shows that machine learning methods can be used to predict household wealth and living standards measures from detailed mobile phone meta data in Rwanda. Blumenstock et al. (2018) apply this method to impact evaluation: they recruited mobile phone subscribers in Afghanistan to participate in a 7-month high-frequency phone-based survey, and matched their responses to historical call detail records. They were able to infer the onset and magnitude of positive and negative economic shocks, including the (randomized) receipt of cash transfers. In cases where cell phone meta-data is available to researchers, baseline survey data collection could usefully collect participants' mobile phone numbers, which could later be matched to call detail records, and together with the appropriate prediction methods, these can generate estimates of living standards.

Yet Blumenstock (2018) also warns that these new data sources may suffer from a lack of validation and biased algorithms. For instance, there is some evidence that existing predictive models may work in one institutional context but not be nearly as successful in others. The number of international phone calls made, for example, is a better predictor of wealth in Rwanda than it is in Afghanistan. Predictive model performance also appears to deteriorate rather quickly over time, raising questions about how often the models need to be re-validated, and at what cost in terms of fresh "training data". In addition, the behavioral patterns currently used for prediction may change when individuals become aware that their personal data is being observed and used to generate statistics that affect eligibility for particular government programs, for instance. Moreover, when these predicative models are trained on biased or patchy data, those who are poorly represented (e.g., household too poor to own a smart phone) may be further marginalized, and predictions for important sub-populations largely uninformative.

The bottom line on new data sources is similar to administrative data: they are cheaper to collect than traditional household surveys and should be used when available, but may lack the specific measures needed to test many important economic research hypotheses. As a result, we do not see original household data collection disappearing from the development economics toolkit anytime soon, including in the context of long-run studies.

5 CONCLUSION

In this article, we argue that the coming years provide an exceptional opportunity for development economists to make intellectual progress in understanding the underlying determinants of long-run living standards, by exploiting the large number of development RCTs that have been conducted since the late 1990's. Despite the methodological and data limitations of many early RCTs in development economics and public health, we identify dozens of studies that currently appear amenable to follow-up evaluations, with scores if not hundreds more "aging into" the possibility of long-run evaluation in the coming decade. If the development economics research community is able to seize this opportunity, it has the potential generate considerable scientific progress in our field.

Given the policy relevance and intellectual importance of long-term impact evaluations, we argue that this research agenda should be a top priority for donors and policymakers. Conducting long-term follow-up studies on past RCTs will demand a large amount of funding and coordinated researcher effort to set up successful survey data collection, often across geographical areas and sometimes across academic disciplines. Yet establishing parallel data collection and tracking protocols across multiple interventions could help generate comparable estimates on the long-run impacts of related interventions, leading to greater external validity. There are already models of successful efforts along these lines. Banerjee et al. (2015b), for example, evaluated a multifaceted program targeted at the very poor in six different countries, and a similar effort is underway in the political economy of development through the EGAP Metaketa initiative (EGAP, 2018). Comparable long-term evaluations of multiple international development interventions will advance intellectual understanding of the drivers of long-run living standards, and could generate valuable insights into comparative cost-effectiveness for policymakers.

We also describe patterns in the relatively small but growing body of literature that already takes advantage of experimental variation to study long-run living standards impacts. One emerging pattern is that several human capital interventions — in both health and education — appear to have successfully led to persistent economic productivity gains, often with impressive rates of return (Baird et al., 2016a). In contrast, most interventions aimed at relaxing liquidity constraints and stimulating firm growth appear to be characterized by positive short-term effects that fade out over time (with the exception of "graduation" programs that are characterized by large asset transfers and intensive training and support). This pattern echoes the lack of persistent or meaningful impacts documented in the microcredit literature (see, for example, Banerjee et al. (2015a)). Yet we caution that this pattern is driven by a relatively small number of RCT studies, and must be viewed as suggestive at this time. With the appropriate resources and coordination, the body of evidence on long-run impacts of these and other development interventions is poised to become much more definitive in the coming years.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are particularly grateful to Craig McIntosh and Prashant Bharadwaj for their work on the systematic review of cash transfer and child health projects discussed in this article, under the Long-term Impact Discovery (LID) initiative. GiveWell provided generous financial support for the LID initiative, and Josh Rosenberg of GiveWell gave us many helpful suggestions. We also benefited from suggestions and comments provided by the LID faculty advisory committee at CEGA, including Lia Fernald, Paul Gertler, Marco Gonzalez-Navarro, and Manisha Shah, as well as from Oriana Bandiera, Robin Burgess, Xavier Jaravel, and Rachael Meager.

References

Aboud FE, Akhter S. 2011. A cluster-randomized evaluation of a responsive stimulation and feeding intervention in bangladesh. *Pediatrics* :peds–2010

Aboud FE, Shafique S, Akhter S. 2009. A responsive feeding intervention increases children's self-feeding and maternal responsiveness but not weight gain. *The Journal of nutrition* 139:1738–1743

Acemoglu D, Johnson S. 2007. Disease and development: the effect of life expectancy on economic growth. *Journal of political Economy* 115:925–985

Adhvaryu A, Nyshadham A, Molina T, Tamayo J. 2018. Helping children catch up: Early life shocks and the progresa experiment. Tech. rep., National Bureau of Economic Research

Aizer A, Eli S, Ferrie J, Lleras-Muney A. 2016. The long-run impact of cash transfers to poor families. *American Economic Review* 106:935–71

Akresh R, de Walque D, Kazianga H. 2016. Evidence from a randomized evaluation of the household welfare impacts of conditional and unconditional cash transfers given to mothers or fathers. The World Bank

Alix-Garcia J, McIntosh C, Sims KR, Welch JR. 2013. The ecological footprint of poverty alleviation: evidence from mexico's oportunidades program. *Review of Economics and Statistics* 95:417–435

Almond D. 2006. Is the 1918 influenza pandemic over? long-term effects of in utero influenza exposure in the post-1940 us population. *Journal of political Economy* 114:672–712

Almond D, Currie J, Duque V. 2017. Childhood circumstances and adult outcomes: Act ii. Tech. rep., National Bureau of Economic Research

Anderson ML. 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* 103:1481–1495

Andrew A, Attanasio O, Fitzsimons E, Grantham-McGregor S, Meghir C, Rubio-Codina M. 2018. Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in colombia. *PLoS medicine* 15:e1002556

Araujo MC, Bosch M, Schady N. 2017. Can cash transfers help households escape an intergenerational poverty trap? In *The Economics of Poverty Traps*. University of Chicago Press

Attanasio OP, Fernández C, Fitzsimons EO, Grantham-McGregor SM, Meghir C, Rubio-Codina M. 2014. Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in colombia: cluster randomized controlled trial. *Bmj* 349:g5785

Awasthi S, Peto R, Read S, Clark S, Pande V, et al. 2013a. Vitamin a supplementation every 6 months with retinol in 1 million pre-school children in north india: Devta, a cluster-randomised trial. *The Lancet* 381:1469–1477

Awasthi S, Peto R, Read S, Clark S, Pande V, et al. 2013b. Vitamin a supplementation every 6 months with retinol in 1 million pre-school children in north india: Devta, a cluster-randomised trial. *The Lancet* 381:1469–1477

Baird S, Hamory J, Miguel E. 2008. Tracking, attrition and data quality in the kenyan life panel survey round 1 (klps-1). Tech. rep., Center for International and Development Economics Research Working Paper

Baird S, Hicks JH, Kremer M, Miguel E. 2016a. Worms at work: Long-run impacts of a child health investment. *The Quarterly Journal of Economics* 131:1637–1680

Baird S, Hicks JH, Kremer M, Miguel E. 2018. Worms and wellbeing: 15 year economic impacts from kenya

Baird S, McIntosh C, Özler B. 2011. Cash or condition? evidence from a cash transfer experiment. *The Quarterly journal of economics* 126:1709–1753

Baird S, McIntosh C, Ozler B. 2016b. When the money runs out: Evaluating the longer-term impacts of a two year cash transfer program. Tech. rep., Working Paper

Bandiera O, Balboni C, Burgess R, Ghatak M, Heil A. 2018. Why do people stay poor? Tech. rep.

Bandiera O, Burgess R, Das N, Gulesci S, Rasul I, Sulaiman M. 2017. Labor markets and poverty in village economies. *The Quarterly Journal of Economics* 132:811–870

Banerjee A, Duflo E, Chattopadhyay R, Shapiro J. 2016. The long term impacts of a graduation program: Evidence from west bengal

Banerjee A, Duflo E, Glennerster R, Kinnan C. 2015a. The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics* 7:22–53

Banerjee A, Duflo E, Goldberg N, Karlan D, Osei R, et al. 2015b. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science* 348:1260799

Banerjee A, Karlan D, Zinman J. 2015c. Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics* 7:1–21

Barham T, Macours K, Maluccio JA. 2017. Are conditional cash transfers fulfilling their promise? schooling, learning, and earnings after 10 years

Barham T, Macours K, Maluccio JA. 2018. Experimental evidence of exposure to a conditional cash transfer during early teenage years: Young womens fertility and labor market outcomes. Tech. rep., CEPR discussion paper

Barrera-Osorio F, Bertrand M, Linden LL, Perez-Calle F. 2011. Improving the design of conditional transfer programs: Evidence from a randomized education experiment in colombia. *American Economic Journal: Applied Economics* 3:167–95

Barrera-Osorio F, Linden LL, Saavedra J. 2017. Medium-and long-term educational consequences of alternative conditional cash transfer designs: Experimental evidence from colombia. Tech. rep., National Bureau of Economic Research

Barreto ML, Farenzena G, Fiaccone R, Santos L, Assis AMdO, et al. 1994. Effect of vitamin a supplementation on diarrhoea and acute lower-respiratory-tract infections in young children in brazil

Bastian J, Michelmore K. 2018. The long-term impact of the earned income tax credit on childrens education and employment outcomes. *Journal of Labor Economics* 36:1127–1163

Behrman JR, Calderon MC, Preston SH, Hoddinott J, Martorell R, Stein AD. 2009. Nutritional supplementation in girls influences the growth of their children: prospective study in guatemala. *The American Journal of Clinical Nutrition* 90:1372–1379

Behrman JR, Parker SW, Todd PE. 2011. Do conditional cash transfers for schooling generate lasting benefits? a five-year followup of progresa/oportunidades. *Journal of Human Resources* 46:93–122

Behrman JR, Sengupta P, Todd P. 2005. Progressing through progresa: An impact assessment of a school subsidy experiment in rural mexico. *Economic development and cultural change* 54:237–275

Benedetti F, Ibarrarán P, McEwan PJ. 2016. Do education and health conditions matter in a large cash transfer? evidence from a honduran experiment. *Economic Development and Cultural Change* 64:759–793

Benhassine N, Devoto F, Duflo E, Dupas P, Pouliquen V. 2015. Turning a shove into a nudge? a" labeled cash transfer" for education. *American Economic Journal: Economic Policy* 7:86–125

Bettinger E, Kremer M, Kugler M, Medina C, Posso C, Saavedra J. 2018. School vouchers, labor markets and vocational education

Bhalotra S, Biroli P, Maselko J, et al. 2017. Maternal depression, women's empowerment, and parental investment: Evidence from a large randomized control trial

Blattman C, Dercon S, S. F. 2018a. The long run effects of industrial and entrepreneurial jobs: 5-year evidence from ethiopia

Blattman C, Fiala N, Martinez S. 2013. Generating skilled self-employment in developing countries: Experimental evidence from uganda. *The Quarterly Journal of Economics* 129:697–752

Blattman C, Fiala N, Martinez S. 2018b. The long term impacts of grants on poverty: 9-year evidence from uganda's youth opportunities program. *Working Paper*

Bleakley H. 2007. Disease and development: evidence from hookworm eradication in the american south. *The Quarterly Journal of Economics* 122:73–117

Bleakley H, Ferrie JP. 2013. Up from poverty? the 1832 cherokee land lottery and the long-run distribution of wealth. Tech. rep., National Bureau of Economic Research

Blumenstock J. 2018. Dont forget people in the use of big data for development

Blumenstock J, Cadamuro G, On R. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350:1073–1076

Blumenstock J, Callen M, Ghani T, Keleher N, Shapiro J. 2018. Measuring poverty and vulnerability in real-time

Borusyak K, Jaravel X. 2016. Revisiting event study designs

Brunello G, Comi S. 2004. Education and earnings growth: evidence from 11 european countries. *Economics of Education Review* 23:75–83

Buchmann N, Field E, Glennerster R, Nazneen S, Pimkina Svetlana SI. 2018. Power vs money: Alternative approaches to reducing child marriage in bangladesh, a randomized control trial. Tech. rep.

Burke M, Lobell DB. 2017. Satellite-based assessment of yield variation and its determinants in smallholder african systems. *Proceedings of the National Academy of Sciences* 114:2189–2194

Caldwell BM, Bradley RH, et al. 1984. Home observation for measurement of the environment. University of Arkansas at Little Rock Little Rock

Charpak N, Tessier R, Ruiz JG, Hernandez JT, Uriza F, et al. 2016. Twenty-year follow-up of kangaroo mother care versus traditional care. *Pediatrics* :e20162063

Chetty R, Friedman JN, Hilger N, Saez E, Schanzenbach DW, Yagan D. 2011a. How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics* 126:1593–1660

Chetty R, Friedman JN, Rockoff JE. 2011b. New evidence on the long-term impacts of tax credits. *IRS Statistics of Income White Paper*

Croke K, Hicks JH, Hsu E, Kremer M, Miguel E. 2016. Does mass deworming affect child nutrition? meta-analysis, cost-effectiveness, and statistical power. Working Paper 22382, National Bureau of Economic Research

Cunha JM. 2014. Testing paternalism: Cash versus in-kind transfers. *American Economic Journal: Applied Economics* 6:195–230

Damien de Walque William H. Dow RN. 2014. Rewarding safer sex conditional cash transfers for hiv/sti prevention. *Working Paper*

Deming D. 2009. Early childhood intervention and life-cycle skill development: Evidence from head start. *American Economic Journal: Applied Economics* 1:111–34

Dibley MJ, Sadjimin T, Kjolhede CL, Moulton LH. 1996. Vitamin a supplementation fails to reduce incidence of acute respiratory illness and diarrhea in preschool-age indonesian children. *The Journal of nutrition* 126:434–442

Duflo E, Dupas P, Kremer M. 2018. The impact of free secondary education: Experimental evidence from ghana

East CN. 2018. The effect of food stamps on childrens health: Evidence from immigrants changing eligibility. *Journal of Human Resources* :0916–8197R2

EGAP. 2018. Metaketa initiative

Eilander A, Gera T, Sachdev HS, Transler C, van der Knaap HC, et al. 2009. Multiple micronutrient supplementation for improving cognitive performance in children: systematic review of randomized controlled trials–. *The American journal of clinical nutrition* 91:115–130

Evans D, Hausladen S, Kosec K, Reese N. 2014. Community-based conditional cash transfers in tanzania: results from a randomized trial. The World Bank

Faber M, Kvalsvig JD, Lombard CJ, Benadé AS. 2005. Effect of a fortified maize-meal porridge on anemia, micronutrient status, and motor development of infants–. *The American journal of clinical nutrition* 82:1032–1039

Galiani S, McEwan PJ. 2013. The heterogeneous impact of conditional cash transfers. *Journal of Public Economics* 103:85–96

Gennaioli N, La Porta R, Lopez-de Silanes F, Shleifer A. 2012. Human capital and regional development. *The Quarterly journal of economics* 128:105–164

Gertler P, Heckman J, Pinto R, Zanolini A, Vermeersch C, et al. 2014. Labor market returns to an early childhood stimulation intervention in jamaica. *Science* 344:998–1001

Green EP, Blattman C, Jamison J, Annan J. 2015. Women's entrepreneurship and intimate partner violence: A cluster randomized trial of microenterprise assistance and partner participation in post-conflict uganda (ssm-d-14-01580r1). *Social science & medicine* 133:177–188

Group SS. 2008. Effect of maternal multiple micronutrient supplementation on fetal loss and infant death in indonesia: a double-blind cluster-randomised trial. *The Lancet* 371:215–227

Haushofer J, Shapiro J. 2016. The short-term impact of unconditional cash transfers to the poor: experimental evidence from kenya. *The Quarterly Journal of Economics* 131:1973–2042

Haushofer J, Shapiro J. 2018. The long-term impact of unconditional cash transfers: Experimental evidence from kenya. *Busara Center for Behavioral Economics, Nairobi, Kenya*

Heckman JJ. 2006. Skill formation and the economics of investing in disadvantaged children. *Science* 312:1900–1902

Henderson JV, Storeygard A, Weil DN. 2012. Measuring economic growth from outer space. *American economic review* 102:994–1028

Herrera MG, Nestel P, Weld L, El Amin A, Mohamed KA, Fawzi W. 1992. Vitamin a supplementation and child survival. *The Lancet* 340:267–271

Hess SY, Peerson JM, Becquey E, Abbeddou S, Ouédraogo CT, et al. 2017. Differing growth responses to nutritional supplements in neighboring health districts of burkina faso are likely due to benefits of small-quantity lipid-based nutrient supplements (lns). *PloS one* 12:e0181770

Hoddinott J, Maluccio JA, Behrman JR, Flores R, Martorell R. 2008. Effect of a nutrition intervention during early childhood on economic productivity in guatemalan adults. *The lancet* 371:411–416

Hoynes H, Schanzenbach DW, Almond D. 2016. Long-run impacts of childhood access to the safety net. *American Economic Review* 106:903–34

Hoynes HW, Schanzenbach DW. 2018. Safety net investments in children. Tech. rep., National Bureau of Economic Research

Hsieh CT, Miguel E, Ortega D, Rodriguez F. 2011. The price of political opposition: Evidence from venezuela's maisanta. *American Economic Journal: Applied Economics* 3:196–214

Institute for Health Metrics and Evaluation. 2018. Global health data exchange

Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353:790–794

Jinabhai CC, Taylor M, Coutsoudis A, Coovadia HM, Tomkins AM, Sullivan KR. 2001. A randomized controlled trial of the effect of antihelminthic treatment and micronutrient fortification on health status and school performance of rural primary school children. *Annals of tropical paediatrics* 21:319–333

Kagitcibasi C, Sunar D, Bekman S, Baydar N, Cemalcilar Z. 2009. Continuing effects of early enrichment in adult life: The turkish early enrichment project 22 years later. *Journal of Applied Developmental Psychology* 30:764–779

Kohler HP, Thornton RL. 2011. Conditional cash transfers and hiv/aids prevention: unconditionally promising? *The World Bank Economic Review* 26:165–190

Kremer M. 2003. Randomized evaluations of educational programs in developing countries: Some lessons. *American Economic Review* 93:102–106

Kugler AD, Rojas I. 2018. Do ccts improve employment and earnings in the very long-term? evidence from mexico. Tech. rep., National Bureau of Economic Research

Lanjouw P, Stern N, et al. 2018. How lives change: Palanpur, india, and development economics. Oxford University Press

Lind T, Lönnerdal B, Stenlund H, Gamayanti IL, Ismail D, et al. 2004. A community-based randomized controlled trial of iron and zinc supplementation in indonesian infants: effects on growth and development. *The American journal of clinical nutrition* 80:729–736

Long KZ, Montoya Y, Hertzmark E, Santos JI, Rosado JL. 2006. A double-blind, randomized, clinical trial of the effect of vitamin a and zinc supplementation on diarrheal disease and respiratory tract infections in children in mexico city, mexico–. *The American journal of clinical nutrition* 83:693–700

Ludwig J, Miller DL. 2007. Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly journal of economics* 122:159–208

Macours K, Schady N, Vakis R. 2012. Cash transfers, behavioral changes, and cognitive development in early childhood: Evidence from a randomized experiment. *American Economic Journal: Applied Economics* 4:247–73

Maleta KM, Phuka J, Alho L, Cheung YB, Dewey KG, et al. 2015. Provision of 10–40 g/d lipid-based nutrient supplements from 6 to 18 months of age does not prevent linear growth faltering in malawi–3. *The Journal of nutrition* 145:1909–1915

Maluccio J, Flores R. 2005. Impact evaluation of a conditional cash transfer program: The nicaraguan red de protección social. Intl Food Policy Res Inst

Maluccio JA, Hoddinott J, Behrman JR, Martorell R, Quisumbing AR, Stein AD. 2009. The impact of improving nutrition during early childhood on education among guatemalan adults. *The Economic Journal* 119:734–763

Manger MS, McKenzie JE, Winichagoon P, Gray A, Chavasit V, et al. 2008. A micronutrient-fortified seasoning powder reduces morbidity and improves short-term cognitive function, but has no effect on anthropometric measures in primary school children in northeast thailand: a randomized controlled trial–. *The American journal of clinical nutrition* 87:1715–1722

Martin RM, Kramer MS, Patel R, Rifas-Shiman SL, Thompson J, et al. 2017. Effects of promoting long-term, exclusive breastfeeding on adolescent adiposity, blood pressure, and growth trajectories: a secondary analysis of a randomized clinical trial. *JAMA pediatrics* 171:e170698–e170698

Martorell R, Habicht JP, Rivera JA. 1995. History and design of the incap longitudinal study (1969–77) and its follow-up (1988–89). *The Journal of nutrition* 125:1027S–1041S

Matias SL, Vargas-Vásquez A, Pérez RB, Valdivia LA, Vivanco OA, et al. 2017. Effects of lipid-based nutrient supplements v. micronutrient powders on nutritional and developmental outcomes among peruvian infants. *Public health nutrition* 20:2998–3007

Mazumder S, Taneja S, Bhatia K, Yoshida S, Kaur J, et al. 2015. Efficacy of early neonatal supplementation with vitamin a to reduce mortality in infancy in haryana, india (neovita): a randomised, double-blind, placebo-controlled trial. *The Lancet* 385:1333–1342

McKenzie D. 2012. Beyond baseline and follow-up: The case for more t in experiments. *Journal of development Economics* 99:210–221

Miguel E, Kremer M. 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72:159–217

Molina Millán T, Barham TCJ, Macours K, Maluccio JA, Stampini M. 2018a. Long-term impacts of conditional cash transfers in latin america: Review of the evidence. Tech. rep.

Molina Millán T, Macours K, Maluccio J, Tejerina L. 2018b. Experimental long-term impacts of early childhood and school age exposure to a conditional cash transfer

Natali L, Handa S, Peterman A, Seidenfeld D, Tembo G, et al. 2016. Making money work: unconditional cash transfers allow women to save and re-invest in rural zambia. Tech. rep.

NEMO Study Group. 2007. Effect of a 12-mo micronutrient intervention on learning and memory in well-nourished and marginally nourished school-aged children: 2 parallel, randomized, placebo-controlled studies in australia and indonesia–. *The American journal of clinical nutrition* 86:1082–1093

Ozier O. 2018. Exploiting externalities to estimate the long-term effects of early childhood deworming. *American Economic Journal: Applied Economics* 10:235–62

Parker SW, Todd PE. 2017. Conditional cash transfers: The case of progresa/oportunidades. *Journal of Economic Literature* 55:866–915

Parker SW, Vogl T. 2018. Do conditional cash transfers improve economic outcomes in the next generation? evidence from mexico. Tech. rep., National Bureau of Economic Research

Paxson C, Schady N. 2010. Does Money Matter? The Effects of Cash Transfers on Child Development in Rural Ecuador. *Economic Development and Cultural Change* 59:187–229

Pitt MM, Rosenzweig MR, Hassan MN. 2012. Human capital investment and the gender division of labor in a brawn-based economy. *American Economic Review* 102:3531–60

Prado EL, Sebayang SK, Apriatni M, Adawiyah SR, Hidayati N, et al. 2017. Maternal multiple micronutrient supplementation and other biomedical and socioenvironmental influences on children's cognition at age 9–12 years in indonesia: follow-up of the summit randomised trial. *The Lancet Global Health* 5:e217–e228

Price DJ, Song J. 2016. The long-term effects of cash assistance. Tech. rep., Working Paper, Stanford University

Rahman MM, Vermund SH, Wahed MA, Fuchs GJ, Baqui AH, Alvarez JO. 2001. Simultaneous zinc and vitamin a supplementation in bangladeshi children: randomised double blind controlled trial. *Bmj* 323:314–318

Rahmathullah L, Underwood BA, Thulasiraj RD, Milton RC, Ramaswamy K, et al. 1990. Reduced mortality among children in southern india receiving a small weekly dose of vitamin a. *New England journal of medicine* 323:929–935

Ross D, Dollimore N, Smith P, Kirkwood B, Arthur P, et al. 1993. Vitamin a supplementation in northern ghana: effects on clinic attendances, hospital admissions, and child mortality. *The Lancet* 342:7 – 12. Originally published as Volume 2, Issue 8862

Sazawal S, Dhingra U, Dhingra P, Hiremath G, Sarkar A, et al. 2010. Micronutrient fortified milk improves iron status, anemia and growth among children 1–4 years: a double masked, randomized, controlled trial. *PloS one* 5:e12167

Sempértegui F, Estrella B, Camaniero V, Betancourt V, Izurieta R, et al. 1999. The beneficial effects of weekly low-dose vitamin a supplementation on acute lower respiratory infections and diarrhea in ecuadorian children. *Pediatrics* 104:e1–e1

Shankar AH, Genton B, Semba RD, Baisor M, Paino J, et al. 1999. Effect of vitamin a supplementation on morbidity due to plasmodium falciparum in young children in papua new guinea: a randomised trial. *The Lancet* 354:203–209

Sivakumar B, Vijayaraghavan K, Vazir S, Balakrishna N, Shatrugna V, et al. 2006. Effect of micronutrient supplement on health and nutritional status of schoolchildren: study design. *Nutrition* 22:S1–S7

Skoufias E, McClafferty B. 2001. Is progresa working: Summary of the results of an evaluation by ifpri

Solon FS, Sarol Jr JN, Bernardo AB, Solon JAA, Mehansho H, et al. 2003. Effect of a multiple-micronutrient-fortified fruit powder beverage on the nutrition status, physical fitness, and cognitive performance of schoolchildren in the philippines. *Food and Nutrition Bulletin* 24:S129–S140

Sommer A. 2008. Vitamin a deficiency and clinical disease: an historical overview. *The Journal of nutrition* 138:1835–1839

Sommer A, Djunaedi E, Loeden A, Tarwotjo I, West JR K, et al. 1986a. Impact of vitamin a supplementation on childhood mortality: a randomised controlled community trial. *The Lancet* 327:1169–1173

Sommer A, Djunaedi E, Loeden A, Tarwotjo I, West JR K, et al. 1986b. Impact of vitamin a supplementation on childhood mortality: a randomised controlled community trial. *The Lancet* 327:1169–1173

Stansfield SK, Pierre-Louis M, Augustin A, Lerebours G. 1993. Vitamin a supplementation and increased prevalence of childhood diarrhoea and acute respiratory infections. *The Lancet* 342:578–582

Strauss J, Witoelar F, Sikoki B. 2016. The fifth wave of the indonesia family life survey (ifls5): Overview and field report. Tech. rep.

Suchdev PS, Ruth LJ, Woodruff BA, Mbakaya C, Mandava U, et al. 2012. Selling sprinkles micronutrient powder reduces anemia, iron deficiency, and vitamin a deficiency in young children in western kenya: a cluster-randomized controlled trial–. *The American journal of clinical nutrition* 95:1223–1230

Tanner JC, Candland T, Odden WS. 2015. Later impacts of early childhood interventions: a systematic review. *Washington: Independent Evaluation Group, World Bank Group*

Thomas D, Frankenberg E, Smith JP. 2001. Lost but not forgotten: Attrition and follow-up in the indonesia family life survey. *Journal of Human resources* :556–592

Thomas D, Witoelar F, Frankenberg E, Sikoki B, Strauss J, et al. 2012. Cutting the costs of attrition: Results from the indonesia family life survey. *Journal of Development Economics* 98:108–123

Unicef, et al. 2018. Levels and trends in child malnutrition. Tech. rep., eSocialSciences

Veenemans J, Milligan P, Prentice AM, Schouten LR, Inja N, et al. 2011. Effect of supplementation with zinc and other micronutrients on malaria in tanzanian children: a randomised trial. *PLoS medicine* 8:e1001125

Vijayaraghavan K, Radhaiah G, Prakasam BS, Sarma KR, Reddy V. 1990. Effect of massive dose vitamin a on morbidity and mortality in indian children. *The Lancet* 336:1342–1345

West Jr KP, Katz J, LeClerq SC, Pradhan E, Tielsch JM, et al. 1991. Efficacy of vitamin a in reducing preschool child mortality in nepal. *The Lancet* 338:67–71

APPENDICES

A Study Screening Procedure for Cash Transfer Studies

Figure A.1 documents the selection process for cash transfer studies. We first identified seven recent meta-analysis studies conducted between 2009 and 2017 (de Walque et al., 2017; Bastagli et al., 2016; Baird et al., 2014; Saavedra et al., 2012; Arnold et al., 2011; Owusu-Addo et al., 2018; Fiszbein & Schady, 2009), from which we extracted 170 publications. To further restrict our attention to the most promising studies, we exclude a study if the study exhibits any of the following features.

1. The study is not randomized experiments.
   We focus on RCTs in order to minimize the risks of including studies that could suffer from internal validity issues.[23]

2. The intervention started after 2010.
   The cutoff is somewhat arbitrary, but corresponds roughly to the lower bound of the time necessary for most recipients to enter the labor force.

3. The study is not sufficiently powered.
   Long-term evaluations are likely to suffer from a reduction in statistical precision, notably due to attrition in long-term tracking (see Section 4). A small sample size can lead to lack of statistical power in the long-term evaluation and yield ambiguous and uninformative results. Predicting long-term effect sizes is difficult, given the heterogeneity of interventions and institutional and economic contexts. Long-term impacts can be larger than short-term impacts if the intervention sets the beneficiaries onto a different life trajectory, as in the Jamaica study (Gertler et al., 2014). However, short-term health or cognitive impacts may also dissipate over time, or translate only partly into labor market outcomes. Fading out of effects in certain interventions is also common, such as in the entrepreneurial grant program in Uganda (Blattman et al., 2018b). This possibility requires researchers to follow up with experiments that are at least sufficiently powered at baseline. As a rule of thumb, we choose the famous Jamaica experiment (Grantham-McGregor et al., 1991) with a sample size of 129 as our benchmark study[24]. Assuming that the Type I error $\alpha = 0.05$, the Type II error $\kappa = 0.80$, treatment and control groups are of the same size $P = 0.50$, and full compliance, we estimate the ex-ante Minimum Detectable Effects (MDE) to be about 0.5 σ[25]. We therefore use 0.5 standard deviation (SD) as our power benchmark, which translates into roughly a sample size of 65 in individual per branches randomization designs or 9 clusters per branch for clustered RCT's, assuming an intra-cluster correlation (ICC) of 10%, and a group size of 100. Note that this benchmark is considered as the bare minimum of the sample size: depending on the context, a larger sample size may be necessary.

[23] Including high quality non-randomized studies as in Molina Millán et al. (2018a) would have required us to take a stance on whether the assumptions underlying the non-experimental estimators are valid. Such assessments are often subjective. We restrict the scope of this review to randomized experiments.

[24] We regroup the treatment arms into cognitive stimulation versus control, ignoring the nutritional supplement treatment, as in Gertler et al. (2014).

[25]

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{\sigma^2}{P(1-P)N}} = (0.84 + 1.96)\sqrt{\frac{\sigma^2}{0.5 \times 0.5 \times 129}} \approx 0.5\sigma$$

Using these criteria, we identify 26 RCT's that may be eligible for a follow-up study. We screen these trials further qualitatively, based on the three criteria described below.

1. Initial implementation quality.
   We analyze the quality of the original research design and its implementation. We examine whether there may be major spillover effects, whether the control groups have been accidentally treated or contaminated, whether there are any baseline imbalances, or compromises of the randomization design due to ethical or logistical concerns, and whether there are any other difficulties in implementing the research design.

2. Sufficient difference in treatment and control exposure.
   As discussed in Section 4, many studies employ a phase-in design to ensure fairness in aid distribution, creating difficulty in long-term evaluations. We do not always exclude studies with phase-in designs. Instead, we make a qualitative assessment as to whether the intensity of the intervention in the treatment group is sufficiently larger than that in the control group to allow for long-term impacts' estimation.

3. Plausibility of tracking households and individuals in the long run.
   In our view, the quality of the short-run tracking informs us about whether long-run tracking can be successful. That said, we recognize that attrition rates are not always monotonically increasing, and respondents who were previously attrited could be re-surveyed with more efforts by the research team.

We grade each study "pass", "fail" or "unclear" for each criteria. We then exclude any study with one or more "fail"s. By these standards, we exclude eight studies[26]. CCT China (Mo et al., 2013) and CT-OVC (Asfaw et al., 2014) were excluded because of severe baseline imbalances. HSNP (Merttens et al., 2013) and MPP (Robertson et al., 2013) were excluded because of severe attrition. NMMT (Aker, 2017) was excluded because it did not have a pure control group. PKH (Alatas et al., 2011) was excluded because some of the control localities received PKH funds through an early and unanticipated expansion of the program, while implementation in some of the treatment localities was delayed. Two additional studies were excluded because we consider them to be outside of the scope of cash transfers. GLSS5+ (Karlan et al., 2014) studies rainfall insurances, and only uses cash transfers as a kind of "control" group. ESHE (Duflo et al., 2015) was an in-kind transfer study. The 18 studies described in Table 2 correspond to the final selected studies.

B Study Screening Procedure for Child Health Studies

In this review, we classify child health interventions into five categories: (1) nutritional supplementation; (2) perinatal interventions (e.g., general training and/or support groups for women, breastfeeding, responsive feeding, kangaroo mother care, cord cleansing)[27]; (3) interventions targeted at specific disease (e.g., deworming, HIV prevention or testing, malaria prevention, diarrhoea); (4) sanitation interventions; and (5) psychosocial stimulation.

[26]The spreadsheet documenting the selection process is available upon request. Study acronyms are not necessarily official.

[27]All the nutritional interventions targeting pregnant women are classified as nutritional supplementation.

Figure B.1 demonstrates a similar selection process as that in Appendix A. We first identified 20 meta-studies[28] across all domains of child health (Conde-Agudelo & Díaz-Rossello, 2016; Gertler et al., 2015; De-Regil et al., 2013; Elder et al., 2014; Farnsworth et al., 2014; Hill et al., 2004; iLiNS, 2017; Imdad et al., 2017, 2013; Sinha et al., 2015; Larsen et al., 2017; Levin & Brouwer, 2014; Singh et al., 2013; Speich et al., 2016; Strunz et al., 2014; Yousafzai & Aboud, 2014; Zwane & Kremer, 2007; Eilander et al., 2009; Meager, 2017; Polec et al., 2015). Then, based on a similar set of criteria, we identified 77 potential studies, as shown in Table 3, that includes 32 nutrition studies and Appendix Table B.1 that includes 45 non-nutrition studies.

We note an important difference of our strategy here as compared to that in Appendix A. In the public health literature, we are more able to leverage the meta analyses that sometimes have already implemented similar selection criteria and constructed databases of relevant studies. We rely on the pre-screening process in these systematic reviews to extract a few promising trials from the vast universe of relevant studies. A rather drastic example of this is Imdad et al. (2013), which reviewed 2303 article records on umbilical cord cleansing in developing countries, and eventually identified only three trials that satisfy their screening criteria.

Given how broad the definition for "child health intervention" is, and how expansive the literature is, we cannot be sure that the list of studies presented in Table B.1 encompasses all the existing RCT's that have potentials for future follow-ups — some studies may not have appeared in the meta analyses that we identified. For this reason, we welcome and appreciate additions from fellow researchers.

[28]The total number of meta-analysis identified so far is actually 36. Because of time constraints, we decided to focus on the 20 most recent meta-analysis. As a result, our estimation of the total number of potential health publications is probably underestimated. Updates on the selection process are available upon request.

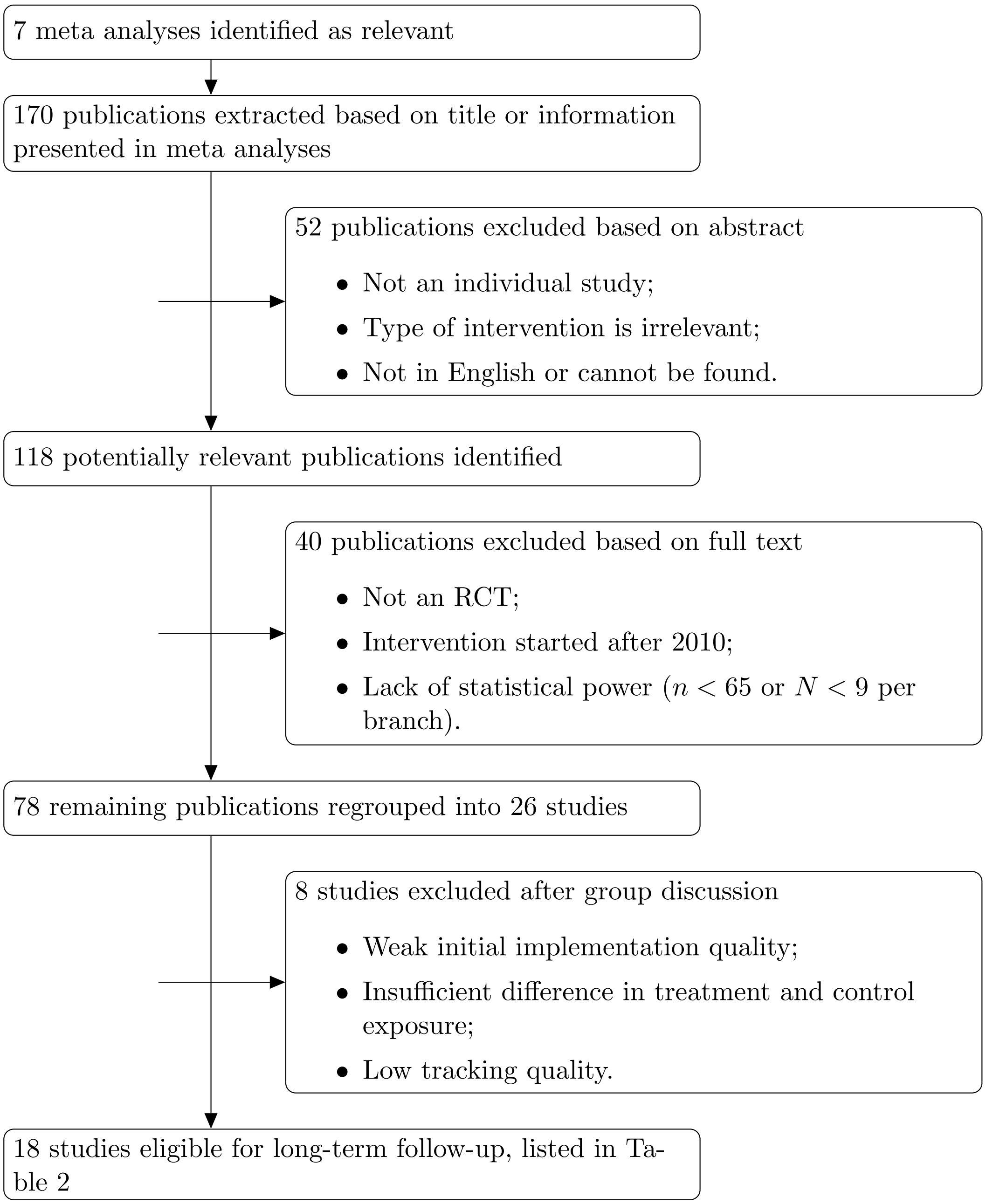

Figure A.1: Flowchart for Study Screening in the Cash Transfer Literature

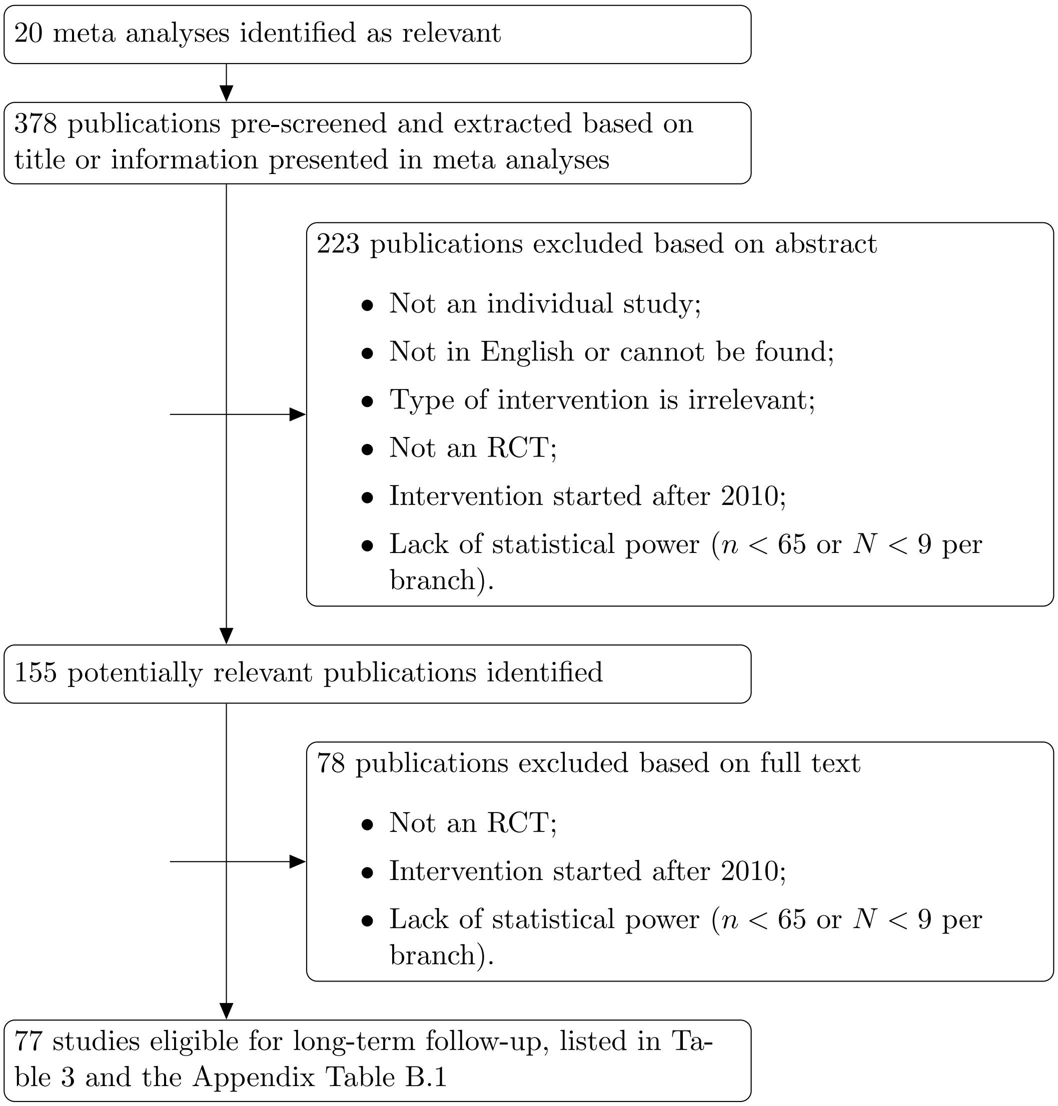

Figure B.1: Flowchart for Study Screening in the Child Health Literature

Table B.1: Selection of Child Health Studies for Long-term Follow-up

	Study	Country	Type	Description	Start of Inter-vention	Clustered RCT	Sample Size[a]	Age of Children	Short-Term Impacts	
									Health	Cognition
1	Charpak et al. (2001)	Colombia	Perinatal	KMC	1993	no	746	newborn	+	
2	Kramer et al. (2001) (PROBIT)	Belarus	Perinatal	Breastfeeding	1996	yes	31	newborn	+	
3	Darmstadt et al. (2008)	Bangladesh	Perinatal	Emollient Therapy	1998	no	497	newborn	+	
4	Mullany et al. (2006)[b]	Nepal	Perinatal	Cord Cleansing	2002	yes	413	newborn	+	
5	Kumar et al. (2008)	India	Perinatal	Newborn Care	2004	yes	39	newborn	+	
6	Pandey et al. (2007)	India	Perinatal	Information on Health Services	2004	yes	105	newborn		
7	Lewycka et al. (2013) (MaiMwana)	Malawi	Perinatal	Women's Group & Peer Counselling	2005	yes	48	newborn	+	
8	Tripathy et al. (2010)	India	Perinatal	Women's Group	2005	yes	36	newborn	+	
9	Roy et al. (2013)	India	Perinatal	Women's Group	2005	yes	36	newborn	+	
10	Chola et al. (2015) (PROMISE-EBF)	Uganda	Perinatal	Breastfeeding	2006	yes	24	0–6 mo	0	
11	El Arifeen et al. (2012)	Bangladesh	Perinatal	Cord Cleansing	2007	yes	133	newborn	+	
12	Persson et al. (2013)	Vietnam	Perinatal	Women's Group	2008	yes	90	newborn	+	
13	Soofi et al. (2012)	Pakistan	Perinatal	Cord Cleansing	2008	yes	187	newborn	+	
14	le Roux et al. (2013)	South Africa	Perinatal	Home Visit	2010	yes	24	1–18 mo	+	
15	Tahir & Al-Sadat (2013)	Malaysia	Perinatal	Telephone Lactation Counselling	2010	no	357	newborn		
16	Reller et al. (2003)	Guatemala	Sanitation	Water Treatment	2001	no	492	< 11 mo	+	
17	Luby et al. (2005)	Pakistan	Sanitation	Handwashing	2002	yes	36	< 15 y	+	
18	Bowen et al. (2007)	China	Sanitation	Handwashing	2003	yes	87	7–8 y	+	
19	Crump et al. (2005)	Kenya	Sanitation	Water Treatment	2003	yes	605	< 2 y	+	
20	Hammer & Spears (2016)	India	Sanitation	Sanitation Campaign	2004	yes	180	< 5 y	+	
21	Kremer et al. (2011)	Kenya	Sanitation	Protected Springs	2005	yes	184	N/A	+	

Table B.1: Selection of Child Health Studies for Long-term Follow-up (Continued)

	Study	Country	Type	Description	Start of Inter-vention	Clustered RCT	Sample Size[a]	Age of Children	Short-Term Impacts: Health	Short-Term Impacts: Cognition
22	Freeman et al. (2012)	Kenya	Sanitation	Handwashing & Water Treatment	2007	yes	135	13–14 y	+	
23	Cameron et al. (2013) (TSSM)	Indonesia	Sanitation	Sanitation Campaign	2008	yes	160	< 5 y	+	
24	Briceño et al. (2015)	Tanzania	Sanitation	Handwashing	2009	yes	181	< 5 y	0	
25	Patil et al. (2014)	India	Sanitation	Sanitation Campaign	2009	yes	80	< 2 y	0	
26	Marsh et al. (1996)	Kenya	Specific Diseases	Bednet	1993	yes	56	< 5 y	+	
27	Miguel & Kremer (2004)	Kenya	Specific Diseases	Deworming	1998	yes	75	6–18 y	+	0
28	Hall et al. (2006)	Vietnam	Specific Diseases	Deworming	1999	yes	80	8–9 y	0	0
29	Alderman et al. (2006)	Uganda	Specific Diseases	Deworming	2000	yes	48	1–7 y	+	
30	Arifeen et al. (2009)	Bangladesh	Specific Diseases	Integrated Management of Childhood Illness	2002	yes	20	< 5 y	+	
31	De La Cruz et al. (2006)	Ghana	Specific Diseases	Bednet Education & Credit	2004	no	1051	< 6 y		
32	Müller et al. (2008)	Bukina Faso	Specific Diseases	Bednet	2006	yes	24	N/A		
33	Chinbuah et al. (2012)	Ghana	Specific Diseases	Antimalaria & Antibiotics	2006	yes	114	2–59 mo	+	
34	Cohen & Dupas (2010)	Kenya	Specific Diseases	Bednet	2007	yes	20	in utero		
35	Tarozzi et al. (2014)	India	Specific Diseases	Bednet	2007	yes	141	N/A	+	
36	Dupas (2009)	Kenya	Specific Diseases	Bednet	2007	no	1289	N/A		
37	Krezanoski et al. (2010)	Madagascar	Specific Diseases	Bednet	2007	yes	21	N/A		
38	Deribew et al. (2010)	Ethiopia	Specific Diseases	Bednet	2009	yes	22	< 5 y	+	
39	Fink & Masiye (2012)	Zambia	Specific Diseases	Bednet	2009	yes	49	N/A	+	
40	Walker et al. (1991)	Jamaica	Stimulation	Stimulation	1986	no	129	9–24 mo	+	
41	Gardner et al. (2003)	Jamaica	Stimulation	Stimulation	1999	no	140	0–2 mo		+
42	Ertem et al. (2006) (Care for Development)	Turkey	Stimulation	Stimulation during Sick Child Visits	2004	no	233	0–24 mo		

Table B.1: Selection of Child Health Studies for Long-term Follow-up (Continued)

	Study	Country	Type	Description	Start of Inter-vention	Clustered RCT	Sample Size[a]	Age of Children	Short-Term Impacts: Health	Short-Term Impacts: Cognition
43	Martinez et al. (2017)	Mozambique	Stimulation	Preschool Construction	2008	yes	76	3–5 y		+
44	Cooper et al. (2009)	South Africa	Stimulation	Home Visit	< 2009	no	449	< 6 mo		+
45	Vazir et al. (2013)	India	Stimulation	Responsive Feeding & Stimulation	< 2012	yes	60	3 mo	+	+

KMC: Kangaroo Mother Care. mo: month; y: year.

+ indicates significant and positive effects, − indicates significant and negative effects, 0 indicates non-significant effects. +/− indicates coexistence of significant positive and negative effects (side effects).

[a] We report the number of clusters for clustered RCT's, and the number of households or individuals for non-clustered RCT's.

[b] In these RCT's, participants in treatment and control arms are regularly examined during the trial, and those with severe conditions (e.g., severe Vitamin A deficiency) are then treated; this practice may change the interpretation of estimated treatment effects.

C Mathematical Appendix

Here we present the derivation of Equation 3 in the text. Consider a setup similar to that of Borusyak & Jaravel (2016), in which a panel of units (individuals or clusters) $i = 1, \cdots, 2N$ are randomized into two (equally sized) groups $j = 0, 1$ which are the control group (or "late treatment" group) and the treatment group (or "early treatment" group), respectively. Suppose that the treatment group receive the treatment at period $T_0 = 0$ and the intervention is phased into the control group after $T_1 = T$ periods. Consider first a situation in which the outcomes Y_{ijt} are observed $(K+1)$ times at $t = 0, T, \cdots, KT$ ("calendar time"). Following the event-study notation, denote "relative time" $K_{jt} = t - T_j$. This denotes the amount of time group j has been exposed to the intervention. If we specify the data-generating process to be

$$Y_{ijt} = \alpha_t + \sum_{k=0}^{K} \tau_k \mathbf{1}\{K_{jt} = kT\} + \epsilon_{ijt} \tag{4}$$

We run the regression of the form

$$Y_{ijt} = \alpha_t + \sum_{k=0}^{K} \hat{\tau}_k \mathbf{1}\{K_{jt} = kT\} + \epsilon_{ijt} \tag{5}$$

Proposition With standard assumptions, the variance of $\hat{\tau}_k$ is

$$\mathrm{Var}(\hat{\tau}_k) = \frac{2}{N}(k+1)\sigma^2 \tag{6}$$

where σ^2 is the residual standard error of the regression.
Proof With standard assumptions (including homoskedasticity)

$$\mathrm{Var}(\hat{\tau}_k) = \sigma^2 \Omega_{kk} \tag{7}$$

where σ is the residual standard error of the regression and Ω_{kk} is the k-th diagonal term in matrix Ω where

$$\Omega = [X^T X]^{-1} \tag{8}$$

where X is a matrix of all the independent variations, including the time fixed effects. Given the setup, we can explicitly express the dummies (with relative time dummies on the left and calendar time dummies on the right) as the block matrix

$$X = \begin{bmatrix} C_K & I_{K+1} \\ \vdots & \vdots \\ C_K & I_{K+1} \\ I_{K+1} & I_{K+1} \\ \vdots & \vdots \\ I_{K+1} & I_{K+1} \end{bmatrix} \tag{9}$$

where I_{K+1} is the identify matrix with dimension $K+1$ and

$$C_K = \begin{bmatrix} \mathbf{0} & 0 \\ I_K & \mathbf{0} \end{bmatrix} \tag{10}$$

where I_K is the identify matrix with dimension K. The X matrix has the dimension $2N \times 2(K+1)$, with the control units stacked on the top and the treatment units stacked on the bottom. Each block represents one unit.

Some block matrix algebra give

$$X^T = \begin{bmatrix} B_K & \cdots & B_K & I_{K+1} & \cdots & I_{K+1} \\ I_{K+1} & \cdots & I_{K+1} & I_{K+1} & \cdots & I_{K+1} \end{bmatrix} \tag{11}$$

where

$$B_K = C_K^T = \begin{bmatrix} \mathbf{0} & I_K \\ 0 & \mathbf{0} \end{bmatrix} \tag{12}$$

So we get

$$X^T X = \begin{bmatrix} N \times (A_K + I_{K+1}) & N \times (B_K + I_{K+1}) \\ N \times (C_K + I_{K+1}) & 2N \times I_{K+1} \end{bmatrix} \tag{13}$$

$$= N \times \begin{bmatrix} A_K + I_{K+1} & B_K + I_{K+1} \\ C_K + I_{K+1} & 2I_{K+1} \end{bmatrix} \tag{14}$$

which has the dimension $2(K+1) \times 2(K+1)$, where

$$A_K = B_K C_K = \begin{bmatrix} I_K & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \tag{15}$$

Denote the inverse of the block matrix as

$$\Omega = [X^T X]^{-1} = \begin{bmatrix} \Omega_1 & \Omega_2 \\ \Omega_3 & \Omega_4 \end{bmatrix} \tag{16}$$

with Ω_1 being an $(K+1) \times (K+1)$ matrix that is relevant for calculating the variance-covariance matrix for the variables of interest (not the time dummies). The formula for inverting block matrices gives

$$\Omega_1 = \frac{1}{N}[(A_K + I_{K+1}) - (B_K + I_{K+1})[2I_{K+1}]^{-1}(C_K + I_{K+1})]^{-1} \tag{17}$$

$$= \frac{2}{N}[I_{K+1} + A_K - B_K - C_K]^{-1} \tag{18}$$

$$= \frac{2}{N} \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & -1 & 2 & -1 & 0 \\ \cdots & 0 & -1 & 2 & -1 \\ \cdots & 0 & 0 & -1 & 1 \end{bmatrix}^{-1} \tag{19}$$

For this tridiagonal matrix, denoted D, denote the diagonal elements $a_k = 2$ if $k < K$ and $a_k = 1$ if $k = K$, the off-diagonal terms $b_k = c_k = -1$. The diagonal elements of the inverse of the tridiagonal matrix are

$$(D^{-1})_{kk} = \theta_{k-1}\phi_{k+1}/\theta_K \tag{20}$$

where $\theta_{-1} = 1$, $\theta_0 = a_0 = 2$ and

$$\theta_k = a_k \theta_{k-1} - b_{k-1} c_{k-1} \theta_{k-2} \tag{21}$$

and $\phi_{K+1} = 1$ and $\phi_K = a_K = 1$ and

$$\phi_k = a_k \phi_{k+1} - b_k c_k \theta_{k+2} \tag{22}$$

With some iterations, we get $\theta_k = k + 2$ if $k < K$, $\theta_k = 1$ if $k = K$ and $\phi_k = 1$ so

$$\Omega_{kk} = \frac{2}{N}(k+1) \tag{23}$$

thereby proving

$$\mathrm{Var}(\hat{\tau}_k) = \frac{2}{N}(k+1)\sigma^2 \tag{24}$$

References

Aker JC. 2017. Comparing cash and voucher transfers in a humanitarian context: Evidence from the democratic republic of congo. *The World Bank Economic Review* 31:44–70

Alatas V, Cahyadi N, Ekasari E, Harmoun S, Hidayat B, et al. 2011. Program keluarga harapan: main findings from the impact evaluation of indonesia's pilot household conditional cash transfer program. *New York: World Bank*

Alderman H, Konde-Lule J, Sebuliba I, Bundy D, Hall A. 2006. Effect on weight gain of routinely giving albendazole to preschool children during child health days in uganda: cluster randomised controlled trial. *bmj* 333:122

Arifeen SE, Hoque DE, Akter T, Rahman M, Hoque ME, et al. 2009. Effect of the integrated management of childhood illness strategy on childhood mortality and nutrition in a rural area in bangladesh: a cluster randomised trial. *The Lancet* 374:393–403

Arnold C, Conway T, Greenslade M. 2011. Cash transfers literature review.

Asfaw S, Davis B, Dewbre J, Handa S, Winters P. 2014. Cash transfer programme, productive activities and labour supply: Evidence from a randomised experiment in kenya. *The Journal of Development Studies* 50:1172–1196

Baird S, Ferreira FH, Özler B, Woolcock M. 2014. Conditional, unconditional and everything in between: a systematic review of the effects of cash transfer programmes on schooling outcomes. *Journal of Development Effectiveness* 6:1–43

Bastagli F, Hagen-Zanker J, Harman L, Barca V, Sturge G, et al. 2016. Cash transfers: what does the evidence say. *A rigorous review of programme impact and the role of design and implementation features. London: ODI*

Blattman C, Fiala N, Martinez S. 2018. The long term impacts of grants on poverty: 9-year evidence from uganda's youth opportunities program. *Working Paper*

Borusyak K, Jaravel X. 2016. Revisiting event study designs

Bowen A, Ma H, Ou J, Billhimer W, Long T, et al. 2007. A cluster-randomized controlled trial evaluating the effect of a handwashing-promotion program in chinese primary schools. *The American journal of tropical medicine and hygiene* 76:1166–1173

Briceño B, Coville A, Martinez S. 2015. Promoting handwashing and sanitation: evidence from a large-scale randomized trial in rural tanzania. The World Bank

Cameron L, Shah M, Olivia S. 2013. Impact evaluation of a large-scale rural sanitation project in indonesia. The World Bank

Charpak N, Ruiz-Peláez JG, Charpak Y, et al. 2001. A randomized, controlled trial of kangaroo mother care: results of follow-up at 1 year of corrected age. *Pediatrics* 108:1072–1079

Chinbuah MA, Kager PA, Abbey M, Gyapong M, Awini E, et al. 2012. Impact of community management of fever (using antimalarials with or without antibiotics) on childhood mortality: a cluster-randomized controlled trial in ghana. *The American journal of tropical medicine and hygiene* 87:11–20

Chola L, Fadnes LT, Engebretsen IM, Nkonki L, Nankabirwa V, et al. 2015. Cost-effectiveness of peer counselling for the promotion of exclusive breastfeeding in uganda. *PloS one* 10:e0142718

Cohen J, Dupas P. 2010. Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics* :1–45

Conde-Agudelo A, Díaz-Rossello JL. 2016. Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. *The Cochrane Library*

Cooper PJ, Tomlinson M, Swartz L, Landman M, Molteno C, et al. 2009. Improving quality of mother-infant relationship and infant attachment in socioeconomically deprived community in south africa: randomised controlled trial. *Bmj* 338:b974

Crump JA, Otieno PO, Slutsker L, Keswick BH, Rosen DH, et al. 2005. Household based treatment of drinking water with flocculant-disinfectant for preventing diarrhoea in areas with turbid source water in rural western kenya: cluster randomised controlled trial. *Bmj* 331:478

Darmstadt GL, Saha SK, Ahmed ANU, Ahmed S, Chowdhury MA, et al. 2008. Effect of skin barrier therapy on neonatal mortality rates in preterm infants in bangladesh: a randomized, controlled, clinical trial. *Pediatrics* 121:522–529

De La Cruz N, Crookston B, Dearden K, Gray B, Ivins N, et al. 2006. Who sleeps under bednets in ghana? a doer/non-doer analysis of malaria prevention behaviours. *Malaria Journal* 5:61

De-Regil LM, Suchdev PS, Vist GE, Walleser S, Peña-Rosas JP. 2013. Home fortification of foods with multiple micronutrient powders for health and nutrition in children under two years of age. *Evidence-Based Child Health: A Cochrane Review Journal* 8:112–201

de Walque D, Fernald L, Gertler P. 2017. Cash transfers and child and adolescent development. 325–342

Deribew A, Alemseged F, Birhanu Z, Sena L, Tegegn A, et al. 2010. Effect of training on the use of long-lasting insecticide-treated bed nets on the burden of malaria among vulnerable groups, south-west ethiopia: baseline results of a cluster randomized trial. *Malaria journal* 9:121

Duflo E, Dupas P, Kremer M. 2015. Education, hiv, and early fertility: Experimental evidence from kenya. *American Economic Review* 105:2757–97

Dupas P. 2009. What matters (and what does not) in households' decision to invest in malaria prevention? *American Economic Review* 99:224–30

Eilander A, Gera T, Sachdev HS, Transler C, van der Knaap HC, et al. 2009. Multiple micronutrient supplementation for improving cognitive performance in children: systematic review of randomized controlled trials–. *The American journal of clinical nutrition* 91:115–130

El Arifeen S, Mullany LC, Shah R, Mannan I, Rahman SM, et al. 2012. The effect of cord cleansing with chlorhexidine on neonatal mortality in rural bangladesh: a community-based, cluster-randomised trial. *The Lancet* 379:1022–1028

Elder JP, Pequegnat W, Ahmed S, Bachman G, Bullock M, et al. 2014. Caregiver behavior change for child survival and development in low-and middle-income countries: An examination of the evidence. *Journal of Health Communication* 19:25–66

Ertem IO, Atay G, Bingoler BE, Dogan DG, Bayhan A, Sarica D. 2006. Promoting child development at sick-child visits: a controlled trial. *Pediatrics* 118:e124–e131

Farnsworth SK, Böse K, Fajobi O, Souza PP, Peniston A, et al. 2014. Community engagement to enhance child survival and early development in low-and middle-income countries: an evidence review. *Journal of health communication* 19:67–88

Fink G, Masiye F. 2012. Assessing the impact of scaling-up bednet coverage through agricultural loan programmes: evidence from a cluster randomised controlled trial in katete, zambia. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 106:660–667

Fiszbein A, Schady NR. 2009. Conditional cash transfers: reducing present and future poverty. The World Bank

Freeman MC, Greene LE, Dreibelbis R, Saboori S, Muga R, et al. 2012. Assessing the impact of a school-based water treatment, hygiene and sanitation programme on pupil absence in nyanza province, kenya: a cluster-randomized trial. *Tropical Medicine & International Health* 17:380–391

Gardner JM, Walker SP, Powell CA, Grantham-McGregor S. 2003. A randomized controlled trial of a home-visiting intervention on cognition and behavior in term low birth weight infants. *The Journal of pediatrics* 143:634–639

Gertler P, Heckman J, Pinto R, Zanolini A, Vermeersch C, et al. 2014. Labor market returns to an early childhood stimulation intervention in jamaica. *Science* 344:998–1001

Gertler P, Shah M, Alzua ML, Cameron L, Martinez S, Patil S. 2015. How does health promotion work? evidence from the dirty business of eliminating open defecation. Tech. rep., National Bureau of Economic Research

Grantham-McGregor SM, Powell CA, Walker SP, Himes JH. 1991. Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the jamaican study. *The Lancet* 338:1–5

Hall A, Khanh LNB, Bundy D, Dung NQ, Hong TS, Lansdown R. 2006. A randomized trial of six monthly deworming on the growth and educational achievements of vietnamese school children. *Unpublished manuscript*

Hammer J, Spears D. 2016. Village sanitation and child health: effects and external validity in a randomized field experiment in rural india. *Journal of Health Economics* 48:135–148

Hill Z, Kirkwood BR, Edmond K, Organization WH, et al. 2004. Family and community practices that promote child survival, growth and development: a review of the evidence

iLiNS. 2017. ilins project publications

Imdad A, Mayo-Wilson E, Herzer K, Bhutta ZA. 2017. Vitamin a supplementation for preventing morbidity and mortality in children from six months to five years of age. *The Cochrane Library*

Imdad A, Mullany LC, Baqui AH, El Arifeen S, Tielsch JM, et al. 2013. The effect of umbilical cord cleansing with chlorhexidine on omphalitis and neonatal mortality in community settings in developing countries: a meta-analysis. *BMC Public Health* 13:S15

Karlan D, Osei R, Osei-Akoto I, Udry C. 2014. Agricultural decisions after relaxing credit and risk constraints. *The Quarterly Journal of Economics* 129:597–652

Kramer MS, Chalmers B, Hodnett ED, Sevkovskaya Z, Dzikovich I, et al. 2001. Promotion of breastfeeding intervention trial (probit): a randomized trial in the republic of belarus. *Jama* 285:413–420

Kremer M, Leino J, Miguel E, Zwane AP. 2011. Spring cleaning: Rural water impacts, valuation, and property rights institutions. *The Quarterly Journal of Economics* 126:145–205

Krezanoski PJ, Comfort AB, Hamer DH. 2010. Effect of incentives on insecticide-treated bed net use in sub-saharan africa: a cluster randomized trial in madagascar. *Malaria journal* 9:186

Kumar V, Mohanty S, Kumar A, Misra RP, Santosham M, et al. 2008. Effect of community-based behaviour change management on neonatal mortality in shivgarh, uttar pradesh, india: a cluster-randomised controlled trial. *The Lancet* 372:1151–1162

Larsen DA, Grisham T, Slawsky E, Narine L. 2017. An individual-level meta-analysis assessing the impact of community-level sanitation access on child stunting, anemia, and diarrhea: Evidence from dhs and mics surveys. *PLoS neglected tropical diseases* 11:e0005591

le Roux IM, Tomlinson M, Harwood JM, Oconnor MJ, Worthman CM, et al. 2013. Outcomes of home visits for pregnant mothers and their infants: a cluster randomised controlled trial. *AIDS (London, England)* 27:1461

Levin C, Brouwer E. 2014. Saving brains: Literature review of reproductive, neonatal, child and maternal health and nutrition interventions to mitigate basic risk factors to promote child development

Lewycka S, Mwansambo C, Rosato M, Kazembe P, Phiri T, et al. 2013. Effect of women's groups and volunteer peer counselling on rates of mortality, morbidity, and health behaviours in mothers and children in rural malawi (maimwana): a factorial, cluster-randomised controlled trial. *The Lancet* 381:1721–1735

Luby SP, Agboatwalla M, Feikin DR, Painter J, Billhimer W, et al. 2005. Effect of handwashing on child health: a randomised controlled trial. *The Lancet* 366:225–233

Marsh V, Mutemi W, Some E, Haaland A, Snow R. 1996. Evaluating the community education programme of an insecticide-treated bed net trial on the kenyan coast. *Health policy and planning* 11:280–291

Martinez S, Naudeau S, Pereira V. 2017. Preschool and child development under extreme poverty: evidence from a randomized experiment in rural mozambique. The World Bank

Meager R. 2017. Vitamin a supplements often substantially reduce child mortality, but their impact is heterogeneous: Resolving a controversy in meta-analysis. *Working paper*

Merttens F, Hurrell A, Marzi M, Attah R, Farhat M, et al. 2013. Kenya hunger safety net programme monitoring and evaluation component. *Impact Evluation Final Report. Oxford Policy Management*

Miguel E, Kremer M. 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72:159–217

Mo D, Zhang L, Yi H, Luo R, Rozelle S, Brinton C. 2013. School dropouts and conditional cash transfers: Evidence from a randomised controlled trial in rural china's junior high schools. *The Journal of Development Studies* 49:190–207

Molina Millán T, Barham TCJ, Macours K, Maluccio JA, Stampini M. 2018. Long-term impacts of conditional cash transfers in latin america: Review of the evidence. Tech. rep.

Mullany LC, Darmstadt GL, Khatry SK, Katz J, LeClerq SC, et al. 2006. Topical applications of chlorhexidine to the umbilical cord for prevention of omphalitis and neonatal mortality in southern nepal: a community-based, cluster-randomised trial. *The Lancet* 367:910–918

Müller O, De Allegri M, Becher H, Tiendrebogo J, Beiersmann C, et al. 2008. Distribution systems of insecticide-treated bed nets for malaria control in rural burkina faso: cluster-randomized controlled trial. *PLoS One* 3:e3182

Owusu-Addo E, Renzaho AM, Smith BJ. 2018. Evaluation of cash transfer programs in sub-saharan africa: a methodological review. *Evaluation and program planning* 68:47–56

Pandey P, Sehgal AR, Riboud M, Levine D, Goyal M. 2007. Informing resource-poor populations and the delivery of entitled health and social services in rural india: a cluster randomized controlled trial. *Jama* 298:1867–1875

Patil SR, Arnold BF, Salvatore AL, Briceno B, Ganguly S, et al. 2014. The effect of india's total sanitation campaign on defecation behaviors and child health in rural madhya pradesh: a cluster randomized controlled trial. *PLoS medicine* 11:e1001709

Persson LÅ, Nga NT, Målqvist M, Hoa DTP, Eriksson L, et al. 2013. Effect of facilitation of local maternal-and-newborn stakeholder groups on neonatal mortality: cluster-randomized controlled trial. *PLoS medicine* 10:e1001445

Polec LA, Petkovic J, Welch V, Ueffing E, Ghogomu ET, et al. 2015. Strategies to increase the ownership and use of insecticide-treated bednets to prevent malaria. *Cochrane Database of Systematic Reviews*

Reller ME, Mendoza CE, Lopez MB, Alvarez M, Hoekstra RM, et al. 2003. A randomized controlled trial of household-based flocculant-disinfectant drinking water treatment for diarrhea prevention in rural guatemala. *The American Journal of Tropical Medicine and Hygiene* 69:411–419

Robertson L, Mushati P, Eaton JW, Dumba L, Mavise G, et al. 2013. Effects of unconditional and conditional cash transfers on child health and development in zimbabwe: a cluster-randomised trial. *The Lancet* 381:1283–1292

Roy SS, Mahapatra R, Rath S, Bajpai A, Singh V, et al. 2013. Improved neonatal survival after participatory learning and action with women's groups: a prospective study in rural eastern india. *Bulletin of the World Health Organization* 91:426–433B

Saavedra JE, García S, et al. 2012. Impacts of conditional cash transfer programs on educational outcomes in developing countries: a meta-analysis. *RAND Labor and Population Working Paper Series, WR-921-1*

Singh M, Brown G, Rogerson SJ. 2013. Ownership and use of insecticide-treated nets during pregnancy in sub-saharan africa: a review. *Malaria journal* 12:268

Sinha B, Chowdhury R, Sankar MJ, Martines J, Taneja S, et al. 2015. Interventions to improve breastfeeding outcomes: a systematic review and meta-analysis. *Acta Paediatrica* 104:114–134

Soofi S, Cousens S, Imdad A, Bhutto N, Ali N, Bhutta ZA. 2012. Topical application of chlorhexidine to neonatal umbilical cords for prevention of omphalitis and neonatal mortality in a rural district of pakistan: a community-based, cluster-randomised trial. *The Lancet* 379:1029–1036

Speich B, Croll D, Fürst T, Utzinger J, Keiser J. 2016. Effect of sanitation and water treatment on intestinal protozoa infection: a systematic review and meta-analysis. *The Lancet Infectious Diseases* 16:87–99

Strunz EC, Addiss DG, Stocks ME, Ogden S, Utzinger J, Freeman MC. 2014. Water, sanitation, hygiene, and soil-transmitted helminth infection: a systematic review and meta-analysis. *PLoS medicine* 11:e1001620

Tahir NM, Al-Sadat N. 2013. Does telephone lactation counselling improve breastfeeding practices?: A randomised controlled trial. *International journal of nursing studies* 50:16–25

Tarozzi A, Mahajan A, Blackburn B, Kopf D, Krishnan L, Yoong J. 2014. Micro-loans, insecticide-treated bednets, and malaria: evidence from a randomized controlled trial in orissa, india. *American Economic Review* 104:1909–41

Tripathy P, Nair N, Barnett S, Mahapatra R, Borghi J, et al. 2010. Effect of a participatory intervention with women's groups on birth outcomes and maternal depression in jharkhand and orissa, india: a cluster-randomised controlled trial. *The Lancet* 375:1182–1192

Vazir S, Engle P, Balakrishna N, Griffiths PL, Johnson SL, et al. 2013. Cluster-randomized trial on complementary and responsive feeding education to caregivers found improved dietary intake, growth and development among rural indian toddlers. *Maternal & child nutrition* 9:99–117

Walker SP, Powell CA, Grantham-McGregor SM, Himes JH, Chang SM. 1991. Nutritional supplementation, psychosocial stimulation, and growth of stunted children: the jamaican study. *The American journal of clinical nutrition* 54:642–648

Yousafzai AK, Aboud F. 2014. Review of implementation processes for integrated nutrition and psychosocial stimulation interventions. *Annals of the New York Academy of Sciences* 1308:33–45

Zwane AP, Kremer M. 2007. What works in fighting diarrheal diseases in developing countries? a critical review. *The World Bank Research Observer* 22:1–24