Title
Essays in Development Economics

Author
Huang, Yue

Publication Date
2021

Peer reviewed|Thesis/dissertation

Essays in Development Economics

by

Yue Huang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Agricultural and Resource Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Edward Miguel, Co-chair
Professor Marco Gonzalez-Navarro, Co-chair
Professor Solomon Hsiang
Professor Jeremy Magruder

Spring 2021

Essays in Development Economics

Abstract

Essays in Development Economics

by

Yue Huang

Doctor of Philosophy in Agricultural and Resource Economics

University of California, Berkeley

Professor Edward Miguel, Co-chair

Professor Marco Gonzalez-Navarro, Co-chair

Big data promises to bring new opportunities of cost saving and data-driven decision making into the field of international development. This dissertation illustrates different ways of utilizing the ever growing repository of satellite imagery and crowd-sourced data to evaluate effectiveness of development programs, while achieving cost saving and timely delivery of insights. Chapter 2 leverages high-resolution daytime satellite imagery and deep learning methods to evaluate development aid programs entirely remotely, using an unconditional cash transfer program in western Kenya as a proof of concept. Chapter 3 shows how we compile crowd-sourced policy data to deliver timely insights on the effectiveness of international COVID-19 containment policies. Chapter 4 explores a new idea to quickly and relatively inexpensively gather more evidence on the long-term impacts of development programs—instead of implementing a new program and tracking the participants for decades to come, we revisit existing randomized control trials that were conducted in the past few decades and identify those that can be followed up for evaluation.

Contents

Acknowledgments

This work would not have been possible without the continued support and fantastic advising of my PhD advisors. I am deeply grateful to Marco Gonzalez-Navarro for always believing in me. In the most difficult times of my PhD years, Marco reminded me that my ideas don't have to be perfect from the beginning, and that I needed to start somewhere and let my projects evolve over time. Marco has spent countless hours with me iterating over numerous versions of figures, tables, and paper drafts. In doing that, he taught me the value of patience and perseverance, in academic research and in life. Edward Miguel has been my role model and has taught me so much about asking important economic questions, answering them with rigor, and communicating with clarity and transparency. Ted provided me with the first opportunity to work on research projects in development economics, connected me with people and resources that I needed to pursue my ideas, and would always share incredible insights that guided my projects in the right direction. I am indebted to Solomon Hsiang for helping me grow into the person that I am today, and for constantly inspiring me with his intellectual leadership, and the audacity and fearlessness to tackle big and important policy questions. Sol pushed me to work independently, to stop seeking validation from others, to take ownership over my own work, and to always think of the practical policy implications—the "so what" questions—for my research projects. I am thankful to Jeremy Magruder for advising me, and for teaching the development economics course—a fascinating class that eventually inspired me to pursue my PhD in the field of international development.

I thank the entire development economics community at UC Berkeley, the Global Policy Lab, and the 2016 Agricultural and Resource Economics (ARE) cohort for giving me feedback on my early-stage work and supporting me emotionally through the PhD journey. I have benefited tremendously from working with Michael Anderson, Craig McIntosh, Prashant Bharadwaj, Adrien Bouguen, Natasha Beale, Michael Kremer, Julius Ruschenpohler, Ellen Bruno, Trinetta Chong, Hannah Druckenmiller, Anna Tompsett, Andreas Madestam, Nicklas Nordfors, Daniel Allen, Sebastien Annan-Phan, Kendon Bell, Ian Bolliger, Andrew Hultgren, Emma Krasovich, Peiley Lau, Jaecheol Lee, Esther Rolf, Jeanette Tseng, Tiffany Wu, and from interactions with Ben Faber, Marshall Burke, Joshua Blumenstock, Supreet Kaur, Ethan Ligon, Elisabeth Sadoulet, Alain De Janvry, Aprajit Mahajan, and many others. My amazing cohort—Wei Lin, Shelley He, Carly Trachtman, Jay Sayre, Molly Sears, James Sears, Daniel Kannell, Katie Wright, Robert Pickmans, Leopold Biardeau, and Hannah Druckenmiller, and the Global Policy Lab community have been supporting and inspiring me throughout the years.

Finally, I am grateful to my friend and partner, Max Jiang, for being my anchor and teaching me many things, including how to exit the Vim editor. I thank my parents, Jiaofang Shi and Shuhao Huang, who have worked hard all their lives so that I could grow up to have more opportunities than they had.

Chapter 1

Introduction

"Why should the financial services industry, where mere dollars are at stake, be using more advanced technologies than the aid industry, where human life is at stake?"

— Sendhil Mullainathan (Mullainathan, 2016)

Big data promise to bring new opportunities of cost saving and data-driven decision making into the field of international development. The total volume of data in the world is increasing by 40% annually (UN, 2021). Most of these data are inexpensively and passively collected from satellites, mobile phones, credit card transactions, crowd-sourcing platforms, and so on. The increasing availability of these data—together with the emergence of machine learning algorithms which would help us make sense of them—has brought changes to many disciplines, including development economics. Novel data sources can reduce the costs and time of measuring living standards and tracking progress towards poverty alleviation goals (Jean et al., 2016a; Blumenstock, 2016; Yeh et al., 2020; Watmough et al., 2019; Engstrom et al., 2017; Babenko et al., 2017), help target development programs more effectively (Blumenstock, 2020; Aiken et al., 2020), provide crucial input for the operations and logistics team (Maxar, 2017), and facilitate evaluation of program effectiveness.

This dissertation illustrates different ways of utilizing this ever growing repository of data for research in development economics, with a particular focus on achieving cost saving and timely delivery of insights.

Chapter 2 explores the potential of using high-resolution daytime satellite imagery and deep learning methods to reduce the costs of program evaluation in international development. I provide the first evidence that we can leverage these technologies to evaluate anti-poverty programs without incurring additional data collection costs. As a proof of concept, I evaluate an unconditional cash transfer program in rural Kenya, and show statistically significant and economically sizeable increases in remotely-sensed housing quality metrics. With an Engel curve approach derived from micro-economic theory, I infer the program effects on household wealth, and obtain consistent results with costly field surveys. On the contrary, the program effects cannot be detected with nighttime light intensity, the most widely used remotely sensed proxy of economic development.

Chapter 3, coauthored with Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Andrew Hultgren, Emma Krasovich, Peiley Lau, Jaecheol Lee, Esther Rolf, Jeanette Tseng and Tiffany Wu, shows how we compile crowd-sourced policy data to deliver timely insights on the effectiveness of international COVID-19 containment policies. Governments around the world are responding to the novel coronavirus (COVID-19) pandemic (Wu et al., 2020a) with unprecedented policies designed to slow the growth rate of infections. Many actions, such as closing schools and restricting populations to their homes, impose large and visible costs on society, but their benefits cannot be directly observed and are currently understood only through process-based simulations (Ferguson et al., 2020; Chinazzi et al., 2020; Kraemer et al., 2020). Here, we compile new data on 1,717 local, regional, and national non-pharmaceutical interventions deployed in the ongoing pandemic across localities in China, South Korea, Italy, Iran, France, and the United States (US). We then apply reduced-form econometric methods, commonly used to measure the effect of policies on economic growth (Greene, 2003; Romer and Romer, 2010) to empirically evaluate the effect that these anti-contagion policies have had on the growth rate of infections. In the absence of policy actions, we estimate that early infections of COVID-19 exhibit exponential growth rates of roughly 38% per day. We find that anti-contagion policies have significantly and substantially slowed this growth. Some policies have different impacts on different populations, but we obtain consistent evidence that the policy packages now deployed are achieving large, beneficial, and measurable health outcomes. We estimate that across these six countries, interventions prevented or delayed on the order of 62 million confirmed cases, corresponding to averting roughly 530 million total infections. These findings may help inform whether or when these policies should be deployed, intensified, or lifted, and they can support decision-making in the other 180+ countries where COVID-19 has been reported (WHO, 2020).

Chapter 4, coauthored with Adrien Bouguen, Michael Kremer and Edward Miguel, explores the potential of a new idea to quickly and relatively inexpensively gather more evidence on the long-term impacts of development programs. Instead of implementing a new program and tracking the participants for decades to come, we look into the possibility of revisiting existing randomized control trials that were conducted in the past few decades and following up with the participants to learn about the programs' long-term impacts. We start by assessing evidence from randomized control trials (RCTs) on long-run economic productivity and living standards in poor countries, and 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.

Taken together, these applications demonstrate the promise of big data to bring positive change to the field of international development. These methods, however, are not without their limitations. Satellite imagery, for example, provides incredibly rich observations for certain aspects of people's financial well-being—their physical assets, agricultural productivity, local infrastructure, etc.—but not others. Without sufficient caution in interpretation, relying solely on these inexpensive measurements for program evaluation or aid targeting may lead to unintended biases (Blumenstock, 2018).

Chapter 2

Using Satellite Imagery and Deep Learning to Evaluate Anti-Poverty Programs

2.1 Introduction

Satellite imagery and deep learning have proven to be effective tools for mapping poverty in low-income countries (Jean et al., 2016a; Blumenstock, 2016; Engstrom et al., 2017; Babenko et al., 2017; Watmough et al., 2019; Yeh et al., 2020). These technologies promise to disrupt the traditional paradigm of international development that is heavily reliant on in-person field surveys. They help aid organizations plan vaccination and electrification campaigns (Facebook, 2019), and identify the most vulnerable communities when sending relief (Aiken et al., 2020; Blumenstock, 2020; Yeh et al., 2020). In this paper, I use a proof-of-concept experiment (Egger et al., 2019) to show that we can similarly leverage these technologies to evaluate the effectiveness of anti-poverty programs, without coordinating costly and logistically challenging field work. With high-resolution daytime satellite imagery and a state-of-the-art deep learning model, I can inexpensively and accurately measure housing quality,

I am extremely grateful to my advisors Marco Gonzalez-Navarro, Edward Miguel and Solomon Hsiang for their continued support and fantastic advising. I also benefited tremendously from suggestions and comments from Jeremy Magruder, Ben Faber, Marshall Burke, Joshua Blumenstock, Supreet Kaur, Ethan Ligon, Elisabeth Sadoulet, Alain De Janvry, Aprajit Mahajan, and the participants in the AGU Fall Meeting 2019 (session: GC34C - Advances in Remote Sensing, Machine Learning, and Economics to Improve Risk Management and Evaluate Impacts in Socioenvironmental Systems), the UC Berkeley Trade Lunch, the UC Berkeley Development Workshop, the UC Berkeley Development Lunch, and the UC Berkeley Good Data Seminar. I thank Edward Miguel, Michael Walker, Dennis Egger, Johannes Haushofer, and Paul Niehaus, and the rest of the GiveDirectly team, for generously sharing the dataset with me and responding to my various inquiries.

a proxy of economic well-being. When combined with geo-coded program implementation records, I can directly observe the program effects on housing quality, and indirectly estimate the effects on economic well-being.

Rigorous program evaluation based on in-person household surveys has formed the basis of the modern approach of fighting against global poverty (Deaton, 1997; Banerjee and Duflo, 2011). Household surveys—typically comprehensive questionnaires containing hundreds of questions that touch every aspect of people's financial lives—help generate crucial insights into the effectiveness of different anti-poverty programs, and provide input for evidence-based policy making. However, they can sometimes be prohibitively expensive to conduct. Studies funded by major donors cost 0.5 million US dollars on average (according to the World Bank) and up to 1.5 million US dollars (according to USAID) (Pamies-Sumner, 2015). Unanticipated events, such as political unrest and pandemics, often disrupt field surveys, leaving crucial data missing (Brune et al., 2020).

The most widely used remotely sensed measure of economic development is nighttime light intensity (hereafter "night light"), which captures the amount of light emitted from Earth at night, and is highly correlated with Gross Domestic Product (GDP), as well as subnational economic development (Henderson et al., 2012; Chen and Nordhaus, 2011; Michalopoulos and Papaioannou, 2014). However, the night light data show poor sensitivity in less developed and rural areas (Jean et al., 2016a), presumably because of low electrification rates—for example, from 1992 to 2008, 99.73% of pixels were completely unlit in Madagascar, 99.47% in Mozambique, and this is representative of low-income countries (Henderson et al., 2012). This makes the data less useful for studying the very target of many international development programs—people living under the poverty line. Additionally, the low spatial granularity of night light prevents it from being used to evaluate programs that generate fine spatial variations, including most of the randomized controlled trials, the gold standard for program evaluation.

Housing quality promises to be a better proxy for economic development than night light. Expenditure on housing accounts for a sizable 10–20% of people's total expenditure (OECD, 2014). It is more likely to reflect private ownership of assets than night light, which is strongly correlated with public good provision (e.g., street lights). In many rural, low-income contexts, people are tied to their lands and don't migrate often, making it possible to track changes over time. They also tend to frequently build and/or upgrade houses, making housing quality a sensitive proxy of economic well-being, even in poor and rural communities with low electrification rates. With modern deep learning techniques, we can measure building footprint with high accuracy at scale (Lindenbaum, 2017). The visual appearance of buildings in satellite imagery (for example, roof color, roof reflectance, building footprint, and the geometric alignment between neighboring buildings) contains rich information about the quality of housing (Marx et al., 2019; Michaels et al., 2017; Kohler et al., 2017). Compared to "black-box" machine learning predictions of economic well-being, explicitly using housing quality as a proxy makes this approach less susceptible to biases introduced by unstable relationships between satellite observations and economic well-being, an important concern that manifests itself in the subsequent results.

2.2 Results

As a proof of concept, I evaluate a randomized controlled trial conducted in 2014–2017 in 653 villages in rural Kenya (Egger et al., 2019). GiveDirectly, a US charity, implemented a program to send unconditional cash transfers to rural households via mobile money, if they met the eligibility criteria of living under a thatched roof (a low quality roof material, often indicating poor living conditions). Each recipient household would receive $1,000—equivalent to about 75% of their annual household expenditure—in lump sum, and could spend it however they want to. To evaluate the effectiveness of the program, GiveDirectly randomly selected about half of the 653 villages as the treatment group, where eligible households (about 1/3 of the population) received transfers, and used the rest as the control group. (In effect, the trial employed a more sophisticated two-tier randomization design, which is not directly relevant to the statistical analysis in this study.) The authors conducted extensive household surveys after the distribution of the transfers, and comprehensively measured program impacts.

Mapping Treatment Intensity and Housing Quality. To evaluate program impacts, I first construct a map that shows the intensity of the policy or development aid program (hereafter "treatment") in different geographical units (for example, raster grid cells). This can be derived from spatially explicit program implementation records, which document where the program was administered. Because of the high resolution of housing quality metrics, one can study programs that induced extremely fine spatial variation—for example, household-randomized trials. Importantly, the variation in treatment intensity has to be either random (if induced by an experiment) or as good as random (in a natural experiment setting), as is the case for any credible applied econometric study.

For the GiveDirectly experiment, I construct the treatment intensity map from a local census in 2014–2015, which surveyed all the 65,385 households living in the study area (Egger et al., 2019). The census data record each household's geo-location, and whether they belong to the treatment (T), control (C), or out-of-sample (O) group (Figure 2.1a). Among the three groups, only the treatment households eventually received the cash transfer from GiveDirectly. The control households were randomized into not receiving the transfer, whereas the out-of-sample households were never eligible to participate in the program. I lay out a regular grid, and count the number of treatment households in each grid cell (Figure 2.1b). As every transfer was roughly USD 1,000, this variable can be interpreted as the amount of cash infusion (in $1,000) into a given grid cell, and is my preferred measure of treatment intensity (Figure 2.1c).

Then, I measure housing quality in daytime satellite images with deep learning techniques. The input images are from Google Static Maps (Google Static Maps, 2020). They are taken after the GiveDirectly intervention, have a spatial resolution of about 30cm per pixel, and contain only the RGB (red, green, blue) bands (Figure 2.1d).

To segment buildings, I train a state-of-the-art deep learning model, Mask R-CNN (He et al., 2017), on large, publicly available datasets such as COCO (Common Objects in Con-

text) (COCO, 2020) and Open AI Tanzania (Open AI Tanzania, 2020), as well as a small annotated dataset, which are randomly sampled from all the input images (see Supplementary Materials A.2 for details on model training). The model predictions are highly accurate, both quantitatively (Supplementary Figure A.1) and qualitatively (Supplementary Figure A.2). The model generalizes well to other countries, such as Mexico, where the number of houses identified in the deep learning predictions is highly correlated with the census population count (Supplementary Figure A.8 and Supplementary Materials A.3). After post-processing, each predicted instance of buildings is represented by a polygon and a "representative" roof color (Figure 2.1e). The Mask R-CNN model conducts instance segmentation (as opposed to semantic segmentation), meaning that it is able to identify every building instance separately, even if they are adjacent to each other. As such, I can measure housing quality for each household, and evaluate program impacts at an unprecedentedly high spatial granularity.

I extract two quality metrics for each building: the size of building footprint, and the type of roof material. The roofs are classified into three types: tin roof, thatched roof, and painted roof, based on their color profiles (Supplementary Figure A.3). Compared to tin roofs, thatched roofs are generally of lower quality—they are cheaper, less durable and require frequent repairs and replacements (Haushofer and Shapiro, 2016; Egger et al., 2019). (Painted roofs are relatively uncommon in the study area.) In prior work, roof reflectance and roof color have been shown to be good proxies of housing quality (Marx et al., 2019; Michaels et al., 2017). As such, I aggregate the total building footprint to measure all housing assets (Figure 2.1f, Building Footprint), and the footprint of tin-roof buildings to measure high-quality housing assets (Figure 2.1f, Tin-roof Area), in each grid cell. To obtain night light data for systematic comparison, I download and resample the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images in 2019 (Google Earth Engine, 2020; Elvidge et al., 2017).

The maps of treatment intensity and remotely sensed outcomes for the GiveDirectly experiment are shown in Figure 2.2. For visual display purposes, I plot the maps with a spatial resolution of 0.005° (roughly 500 meters), which is lower than the resolution used in the subsequent statistical analysis. The experiment generated substantial variation in treatment intensity, as expected (Figure 2.2a). The night light data demonstrate little variation in this rural, sparsely populated area, except in a few spots close to local towns (Figure 2.2d), whereas both of the housing quality measures capture richer variation in the entire area (Figure 2.2b, c).

Estimating the Program Effects on Housing Quality. I regress the remotely sensed outcomes on treatment intensity to estimate the causal effects of the GiveDirectly cash transfer. I choose a spatial resolution of 0.001° (approximately 100m), such that most of the grid cells contain 0–5 households. I exploit only the experimentally-induced random variation in treatment intensity for identification, and account for pre-determined differences in program eligibility. Intuitively, imagine two grid cells, one containing a household who

received the transfer, and the other containing a household who was eligible to get the transfer but did not because they were randomized into the control group. With a valid randomization (Egger et al., 2019), the differences in outcomes between the two can be attributed to the cash transfer. I plot the causal effects on night light and housing quality as cash infusion intensity increases (Figure 2.3, in color), without making assumptions on the structure of the effects. The results suggest that the effects grow linearly with the amount of cash infusion. I therefore also report an "average" effect, estimated with the assumption that each $1,000 transfer generated an effect of the same magnitude (Figure 2.3, panel subtitles). I demonstrate the validity of the empirical strategy further by running 100 placebo simulations—I artificially generate placebo cash transfers that did not actually take place but is consistent with the original randomization design, and estimate their treatment effects (Figure 2.3, in gray). The resulting estimates are reassuringly centered around zero and not biased.

I observe statistically significant and economically sizable effects on housing quality, on both the extensive margin (larger building footprint) (Figure 2.3a), and the intensive margin (higher quality roofs) (Figure 2.3b). On average, a $1,000 cash transfer significantly increased building footprint by 7.9 square meters (85.0 square feet), and tin-roof area by 13.6 square meters (146.4 square feet). These increases indicate that households may have built new structures—either primary residences or auxiliary structures, such as kitchens and sheds, expanded their existing structures, and/or upgraded their thatched roofs to tin roofs, an improvement that people commonly used the transfer for (Haushofer and Shapiro, 2016). These estimates are consistent with the results from extensive field surveys, which also documented large increases in housing asset values (Egger et al., 2019).

On the other hand, I do not observe any program effects on night light (Figure 2.3c), despite the fact that the cash transfer had large positive impacts on many aspects of the recipient households' economic well-being—food expenditure, consumer durable spending, asset holding, and housing values (Egger et al., 2019). The estimated effect is not statistically significant, small in magnitude, and actually slightly negative. This may be because of low demand for electrification (Lee et al., 2020), or the poor sensitivity of night light in low-income, rural regions (Jean et al., 2016a).

Recovering the Program Effects on Economic Well-being with Engel Curves. I recover the program effects on household economic well-being with a canonical micro-economic concept, the Engel curve. Engel curves describe how household expenditure on certain goods or services depends on households' economic well-being. For example, poorer families spend a larger share of their expenditure on food. Therefore, if we cannot measure economic well-being directly (e.g., due to price changes), we can measure how much of people's expenditure is spent on food, to infer their economic well-being (Elbers et al., 2003; Tarozzi and Deaton, 2009; Young, 2012; Atkin et al., 2020). I adapt this framework to housing quality—someone who lives in a larger house are likely to be wealthier (Figure 2.4a). If we know that someone's house size increased, then we could infer that their wealth

also grew—as if they were moving up on the Engel curve. Mathematically, the slope of the Engel curve represents the ratio between the change in house size and the change in wealth. I divide the change in the house size (Figure 2.3) by the slope of the Engel curve (Figure 2.4a) to infer the corresponding change in wealth (Figure 2.4b). Importantly, the validity of this approach depends on the assumption that the Engel curve does not shift in response to the treatment.

The Engel curves can be derived from any geo-coded consumption and expenditure survey, as long as the surveyed households are—or can be re-weighted to be—representative of the sample in the previous treatment effect estimation step. Notably, the sample does not necessarily have to include any one who has received the treatment, opening up the possibilities of using existing data sources (such as the Living Standards Measurement Study) to estimate Engel curves. In this study, I derive the Engel curves from an endline survey of the GiveDirectly trial participants between May 2016 and June 2017, which includes 5,543 geo-coded households who were eligible for the transfer. Of these households, only those assigned to the control group were used for the estimation. In Figure 2.4a, I show the relationship between survey-based measures of economic well-being (x-axis) and remotely sensed night light or housing quality measures (y-axis). The Engel curves are estimated with a linear regression (dotted lines). The non-linear fit with LOESS (solid lines) shows only small deviations from the linear regression line, and I cannot reject the null hypothesis that the Engel curves for total assets are linear (see Methods for details). The Engel curves are also roughly monotonically increasing, validating the choice of these variables as wealth proxies.

I scale the program effects on each remotely sensed outcome to estimate the program effects on household wealth, measured by aggregating the values of a variety of assets. In Figure 2.4b, I compare the satellite-derived estimates against the survey-based estimates, which are computed from rich endline household survey data and taken from Table 1, Column 1 in the original paper (Egger et al., 2019). The estimate based on building footprint (USD 466 PPP) is both informative and consistent with the survey estimate (USD 556 PPP). The estimate based on night light is slightly negative and imprecise, bounding the treatment effect at over USD 1,800 PPP. For reference, the entire GiveDirectly cash transfer is worth USD 1,871 PPP (USD 1,000 nominal), so this estimate hardly provides any new information. The estimate based on tin-roof area is biased. The results are qualitatively similar when I distinguish between housing asset (Supplementary Figure A.4) and non-housing asset (Supplementary Figure A.5), or when I use annual consumption expenditure as the alternative measure of economic well-being (Supplementary Figure A.6).

The bias in the satellite-derived estimate based on tin-roof area may be due to the violation of a key assumption, that the Engel curve cannot change directly in response to the treatment. For example, if households participate in a program that directly gives them food, we can no longer look at their food consumption to infer the program effects on economic well-being, because the relationship between the two will be altered. In this case, only households that lived in thatched-roof houses were eligible for the GiveDirectly transfers. Households' usual consumption patterns of high-quality tin roofs might have been affected by this eligibility criteria. The treatment households owned more tin-roof buildings,

compared to control households with the same amount of wealth (Supplementary Figure A.7). They may have interpreted this as a "labelled" cash transfer (Benhassine et al., 2015a). Roof upgrading may have become a more salient investment because of the targeting criteria used. The lump sum nature of the transfer may have led recipient households to spend more on lumpy purchases than they normally would. Any of these factors could cause the tin-roof area estimates in Figure 2.4b to be inflated.

These results highlight the importance of using interpretable proxies when evaluating programs with machine learning predictions. An emerging literature is making great progress in mapping poverty with satellite imagery and machine learning with a high spatial granularity at scale (Jean et al., 2016a; Yeh et al., 2020; Blumenstock, 2016; Watmough et al., 2019; Aiken et al., 2020; Blumenstock, 2020; Engstrom et al., 2017; Babenko et al., 2017). Typically, a machine learning model first learns the mapping between the input satellite images and the ground truth labels of wealth or consumption expenditure, assembled from geo-coded household surveys. Then, the model generates predicted poverty maps for every region in the sample, including those with no survey coverage. The model implicitly combines and executes two tasks: (1) extracting semantically meaningful observations of, say, housing quality, agricultural productivity, or infrastructure, from raw satellite images; and (2) inferring economic well-being from observing the consumption patterns of these private or public goods (equivalent to the Engel curve analysis in this study). While the flexibility of the machine learning models helps improve predictive performance, the difficulty in interpretation makes it almost impossible to know or constrain what private or public goods are identified and utilized by the model. Since black-box machine learning models utilize as much information as possible from the input satellite images, it is very likely that the Engel curves of at least some of the observed goods will change (similarly to the tin-roof area variable shown in Figure 2.4b), introducing biases in estimated program effects. In this study, I disentangle the two tasks, so that the first task can be framed as a traditional object detection or segmentation task, allowing me to leverage extensive research in computer science; and the second task becomes more transparent, explicit, and the assumptions testable (for example, with Supplementary Figure A.7).

2.3 Discussion

Program evaluation based on satellite imagery is inexpensive, but not without limitations. Remotely sensed variables can be context specific, and less directly interpretable than explicit survey measures. For example, the assumption that most buildings are one storey high is appropriate for the GiveDirectly study area (and most other rural areas), but may quickly become questionable as we move to an urbanized context. The fact that wealth is generally associated with bigger houses is also mostly appropriate, but can fail in highly dense urban areas, where land supply is restricted and housing prices are high. The use of imputed types of roof materials to quantify roof quality, and thus housing quality, is valid in low-income contexts, but not necessarily in middle- or high-income contexts, where most roofs are made

of durable materials, and roof colors are more of a reflection of personal taste and culture, rather than wealth or social status.

Another fundamental limitation to evaluating programs based on satellite imagery is that the program would have to generate impacts on housing, or other observable variables such as agricultural productivity and infrastructure. This is less plausible for certain programs targeted at addressing other development challenges, or less intensive programs. For example, it seems unlikely that the benefits of vaccination campaigns, or teacher training programs could be observed from space. We also cannot observe the movement of people. Migration rates are very low in the GiveDirectly study area (Egger et al., 2019), but this could present major challenges if used to study other programs (for example, education-related interventions) that naturally impacts mobility.

2.4 Methods

Constructing the Treatment Intensity Map. To construct the treatment intensity map, I utilize data from a baseline census, which was conducted by the authors of the original paper in 2014–2015. The census identified all 65,385 households (roughly 280,000 people) residing in 653 villages in the study area, recorded their GPS coordinates, whether each household was eligible for the GiveDirectly cash transfer, and whether they had been randomized into the treatment or control group (Egger et al., 2019). To address the measurement errors of the GPS collection devices, I discard 58 outliers (living more than 2 kilometers away from the village centers) and impute those and other 4 missing GPS coordinates with village center coordinates. Then, I convert these household records into a raster map. I lay out a regular grid, and count, in each grid cell, the number of households that ultimately received the GiveDirectly cash transfer (see Figure 2.1 and Figure 2.2a). Grid cells containing no eligible households are excluded. To account for pre-determined policy intensity differences, I record (and later control for) the number of households that were eligible for the cash transfer, regardless of whether they had been randomized into the treatment or control group.

Obtaining High-resolution Daytime Satellite Images. I utilize high-resolution daytime satellite images from Google Static Maps (Google Static Maps, 2020). These images have a spatial resolution of about 30cm per pixel (at equator), and contain only the RGB (red, green, blue) bands (see Figure 2.1d and Supplementary Figure A.2 for examples). These images come from a variety of commercial providers such as Maxar (formerly DigitalGlobe) and Airbus, and have been seamlessly mosaicked together. They have also been geo-referenced and pre-processed to remove clouds and address other data quality issues. Google does not provide the exact timestamps for these images, but I estimate that they were taken in 2019, most likely on Dec 30, 2019. The dates for retrieving these images from the Google Static Maps API are between Feb 19 and Feb 21, 2020, and the Google Earth Pro imagery archive reflects that the closest available images in the study area were from

Dec 30, 2019. Multiple other satellite images taken in February, March, July, August and September 2019 are also available in the study area, indicating that the images used in this study are most certainly from 2019.

Extracting Housing Quality Metrics with Mask R-CNN. I first leverage a state-of-the-art deep learning model, Mask R-CNN (He et al., 2017) to segment buildings—that is, to detect each building and the pixels that they occupy—in the Google Static Map satellite images. I then convert the pixel-wise predictions to polygons, and extract housing quality metrics related to the size of the building and the roof materials from each polygon (see Figure 2.1e and Supplementary Figure A.2 for examples).

Loosely speaking, the Mask R-CNN model operates as follows. First, the model proposes a large number of "regions of interest", each of which potentially contains a building. Then, the model uses convolutional filters to identify patterns within the proposed region that are indicative of the presence of buildings, such as the sharp edges, the highly reflective roofs, and the building shadows. Finally, the model predicts whether each proposed region contains a building, as well as whether each pixel is occupied by the building.

I train the Mask R-CNN model with a multi-step process and a transfer learning framework, as described in greater detail in Supplementary Materials A.2. Publicly available building footprint datasets in rural and low-income regions are rare, and they often differ substantially in spatial resolution, sensor instrument, and landscape from inference images (that is, the target images that the model will make predictions for). Relying solely on publicly available training data is therefore insufficient for achieving satisfactory predictive performance. I curate a set of in-sample annotations by randomly sampling 120 images from all the Google Static Map images in the study area, and manually creating high-quality building footprint annotations for them. I pre-train the Mask R-CNN model on large, publicly available datasets such as COCO (Common Objects in Context) and Open AI Tanzania, and fine-tune them on this set of in-sample annotations.

The model predictions are highly accurate. The overall F1 score (a standard performance metric for instance segmentation) on a random subset of inference images is 0.79 (Supplementary Figure A.1). The F1 score is the harmonic mean of precision (the proportion of model-identified buildings that are actual buildings) and recall (the proportion of actual buildings that are correctly identified by the model). Here, a building is deemed to be correctly identified if the predicted pixel mask and the ground truth pixel mask have sufficient overlap (more precisely, if the intersection of the two masks is more than 50% of the union of the two masks). As a reference point, the top winner in the 2nd SpaceNet building footprint extraction competition reported an F1 score of 0.69 (Lindenbaum, 2017). This demonstrates that the Mask R-CNN model used in this study performs well, although building footprint segmentation in rural, less complex scenes is generally easier than in modern cities so these metrics are not directly comparable.

I post-process the model-predicted pixel masks by converting them to polygons, and simplifying the polygons with the Douglas-Peucker algorithm with a pixel tolerance of 3.

For each polygon, I compute two housing quality metrics: building footprint and type of roof materials. I then lay out a regular grid, assign each building to grid cells based on the centroids of the polygons, and aggregate to obtain two metrics at the pixel level: building footprint (Figure 2.2b) and tin-roof area (Figure 2.2c).

First, I measure the size of each building polygon and convert it to square meters. I correct for area distortion, which is induced by the Web Mercator projection system that the Google Static Map uses. This metric may appear larger than what one expects for the size of homes in a low-income context (Figure 2.4), because (1) it represents the footprint of the entire building, which is typically larger than the size of the livable area; and (2) it accounts for both residential and non-residential structures, since the model is not able to distinguish between the two.

Second, I estimate the types of roof materials based on the colors of the roofs, and compute the footprint of tin-roof buildings in each grid cell. For each building, I take all the pixels associated with the given building instance, and assign a "representative" roof color by computing the average values in the RGB (Red, Green, Blue) channels. Since the Euclidean distances between color vectors in the RGB color space does not reflect perceptual differences, I project all the RGB color vectors to the CIELAB color space, and cluster these roof color vectors into 8 groups by running the K-means clustering algorithm. I further classify these 8 groups into three types of roof materials: tin roof, thatched roof, and painted roof (Supplementary Figure A.3), and compute the total footprint of tin-roof buildings.

Obtaining the Night Light Data. To measure nighttime luminosity, I use the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images hosted on Google Earth Engine (Google Earth Engine, 2020; Elvidge et al., 2017). The VIIRS-DNB data product excludes areas impacted by cloud cover and correct for stray light (Mills et al., 2013). However, it has not been filtered to screen out lights from aurora, fires, boats, and other temporal lights, and lights are not separated from background (non-light) values (Google Earth Engine, 2020). This data product has a native spatial resolution of 15 arc seconds (approximately 463 meters at the equator), and I resample the data by conducting nearest neighbor interpolation when necessary. I average over all the monthly observations in 2019 and construct a single cross sectional observation, to reduce seasonality effects and for consistency with the daytime satellite imagery (Figure 2.2d). The VIIRS-DNB data product is considered superior to the more widely used night light data, DMSP-OLS (the United States Air Force Defense Meteorological Satellite Program, Operational Linescan System) because it preserves finer spatial details, has a lower detection limit and displays no saturation on bright lights (Elvidge et al., 2013). This ensures that I conduct a fair comparison with the most modern and high-quality night light data product.

Estimating the Program Effects on Housing Quality. The main econometric specification for Figure 2.3 is as follows

$$y_i = \sum_{k \in K} \tau_k \mathbf{1}\{x_i = k\} + \sum_{m \in M} \beta_m \mathbf{1}\{e_i = m\} + \epsilon_i \tag{2.1}$$

where each observation i represents a $0.001^\circ \times 0.001^\circ$ grid cell (approximately 100m×100m); τ_k represents the estimate of interest: the treatment effects of the unconditional cash transfer on remotely sensed outcomes; x_i denotes the number of recipient households per grid cell (equivalent to the amount of cash infusion in \$1000); e_i denotes the number of eligible households per grid cell, with $m \in M = \{0, 1, 2, 3, \cdots\}$; and y_i denotes remotely sensed outcomes: night light, building footprint, and tin-roof area. To account for pre-existing differences in population density or wealth, which may cause non-random variation in treatment intensity, I flexibly control for the number of eligible households per grid cell, and exclude grid cells with no eligible households. Because the grid cells are fairly small and the number of observations for $k > 2$ is small, I bin the number of recipient households into four bins $k \in K = \{0, 1, 2, 2+\}$, to preserve statistical power. Standard errors are calculated à la Conley, with a uniform kernel and a 3km cutoff (Conley, 1999, 2008; Hsiang, 2010; Burlig and Woerman, 2016). To reduce the effects of outliers (due to sensor malfunctioning or machine learning model prediction errors), I winsorize all remotely sensed variables at the 97.5 percentile.

I run 100 placebo simulations to further demonstrate the validity of the main specification. In each simulation, I randomly assign half of the 68 groups of villages to the high-saturation group, and the other half to the low-saturation group. In the high-saturation groups, I randomly assign 2/3 of the villages to the treatment group (and the rest to the control group); whereas in the low-saturation group, I assign only 1/3 of the villages to the treatment group (and the rest to the control group). This mimics the two-tier randomization scheme of the original trial (Egger et al., 2019). Using these simulated placebo treatment status variables, I estimate the placebo treatment effects with the econometric specification described in Equation 2.1.

To compute a single pooled treatment effect, I make an assumption of linear treatment effects—every transfer of \$1000 has an effect of the same magnitude, regardless of the treatment intensity in that geographical area. The resulting econometric specification is as follows

$$y_i = \tau x_i + \sum_{m \in M} \beta_m \mathbf{1}\{e_i = m\} + \epsilon_i \tag{2.2}$$

where τ is the "average" treatment effect, and all else remain the same as in Equation 2.1.

Estimating the Engel Curves. An Engel curve describes how household expenditure on a particular good varies with income—a relationship that can be used to infer households' economic well-being from the consumption patterns of a limited subset of goods (Elbers

et al., 2003; Tarozzi and Deaton, 2009; Young, 2012; Atkin et al., 2020). The mathematical formulation is

$$Q_{hp} = F_p(W_h) + \epsilon_{hp} \tag{2.3}$$

where household h with W_h wealth (or other measures of economic well-being) would consume Q_{hp} quantities of a normal good p, and $F_p(\cdot)$ represents the Engel curve for product p in the population. With a linearity assumption, this can be simplified to be

$$Q_{hp} = \alpha_p + \beta_p W_h + \epsilon_{hp} \tag{2.4}$$

where α_p is the intercept and β_p is the slope of a linear Engel curve.

In this study, I estimate the Engel curves—the relationships between remotely sensed metrics and survey-based measures of economic well-being—based on the endline survey of the original GiveDirectly trial, which includes a representative set of 5,543 geo-coded households who were eligible for the transfer. The households participated in a comprehensive consumption and expenditure survey between May 2016 and June 2017, after the distribution of cash transfers. From the surveys, I observe annualized household consumption expenditure, and asset values. Household consumption expenditure is the annualized sum of total food consumption in the last 7 days, frequent purchases in the last month, and infrequent purchases over the last 12 months. Household assets include housing and non-housing assets, but not land values. Housing asset values are measured as the respondent's self-reported cost to build a home like theirs. Non-housing assets include livestock, transportation (bicycles, motorcycles, and cars), electronics, farm tools, furniture, other home goods, and lending or borrowing from formal or informal sources. I do not study land values because they are difficult to value given thin local markets (Egger et al., 2019).

I perform heuristic matching between the buildings and the household survey GPS coordinates, to link variables in the survey with remotely sensed variables. First, I take the baseline census data, which geo-coded every single household who lived in the study area, and assign every building in the satellite images to its closest census GPS coordinate, if the distance between the two was within 250m. This ensures that every building is matched to at most one household. Second, I match GPS coordinates from the survey with GPS coordinates from the census. While the same household supposedly had the same geo-location, these two often differed because of the measurement errors of the GPS collection devices, and because the coordinates might be recorded anywhere on the participants' plots and not necessarily in their primary residence. I similarly assign each survey GPS coordinate to its closest census GPS coordinate, if the distance between the two was within 250m. In cases of multiple surveys being assigned to the same census coordinate, I keep the closest survey. The final sample contains only census observations that are matched with both buildings in the satellite images and survey records, and consists of 4,594 households.

The Engel curves are estimated with only the control group (Figure 2.4a and Supplementary Figure A.4a, A.5a and A.6a). They are estimated both non-linearly with LOESS (see Equation 2.3 and the solid lines in Figure 2.4a) and linearly (see Equation 2.4 and

the dotted lines in Figure 2.4a). When fitting LOESS, I allow for locally-fitted quadratic polynomials, and use 75% of the data points for each fit. I test for the non-linearity of the Engel curves in a separate procedure. I first run a linear regression, take the residuals, and fit the residuals with a natural (cubic) spline with 5 knots. I then conduct an F-test on the coefficients of the natural spline basis, and reject the null hypothesis (linearity) if these coefficients are jointly significant. To minimize the influence of outliers, I winsorize consumption expenditure, housing asset values, non-housing asset values and overall asset values at the 97.5 percentile of the full (eligible and non-eligible) sample. Additionally, I winsorize non-housing asset values and overall asset values at the 2.5 percentile, as outliers with a large amount of debt exist and could potentially drive the results otherwise. I also winsorize all the remotely sensed variables at the 97.5 percentile.

Recovering the Program Effects on Economic Well-being. I adapt a prior mathematical formulation that uses the Engel curve to infer changes in economic well-being (Young, 2012). Suppose that one is interested in studying the effect of a plausibly exogenous treatment Z on, say, wealth W (denoted $\hat{\tau}_W$), but can only inexpensively observe its effect on the consumption of product p (denoted $\hat{\tau}_{Q_p}$). Recall that $\hat{\beta}_p$ is the estimated slope of the linear Engel curve in Equation 2.4, then

$$\hat{\tau}_W = \hat{\tau}_{Q_p} / \hat{\beta}_p \tag{2.5}$$

Using a formula for propagation of error (or the multivariate Delta method), one can derive the standard error for $\hat{\tau}_W$ as follows. This derivation is based on prior work (Young, 2012), but additionally accounts for the precision of the slope of the Engel curve.

$$\left(\frac{\hat{\sigma}(\hat{\tau}_W)}{\hat{\tau}_W} \right)^2 = \left(\frac{\hat{\sigma}(\hat{\tau}_{Q_p})}{\hat{\tau}_{Q_p}} \right)^2 + \left(\frac{\hat{\sigma}(\hat{\beta}_p)}{\hat{\beta}_p} \right)^2 \tag{2.6}$$

A key assumption of this approach is that $\hat{\beta}_p$ does not depend on Z—that is, the Engel curve does not change in direct response to the treatment—also termed the conditional independence assumption (Tarozzi and Deaton, 2009).

I estimate the treatment effects on wealth (or other measures of economic well-being) according to Equation 2.5 and Equation 2.6, with the treatment effect estimates for remotely sensed variables, and the slopes of the Engel curves. I compare the satellite-derived estimates against the survey-based estimates, taken from Table 1, Column 1 in the original paper (Egger et al., 2019), which were based on the endline household survey data (Figure 2.4b).

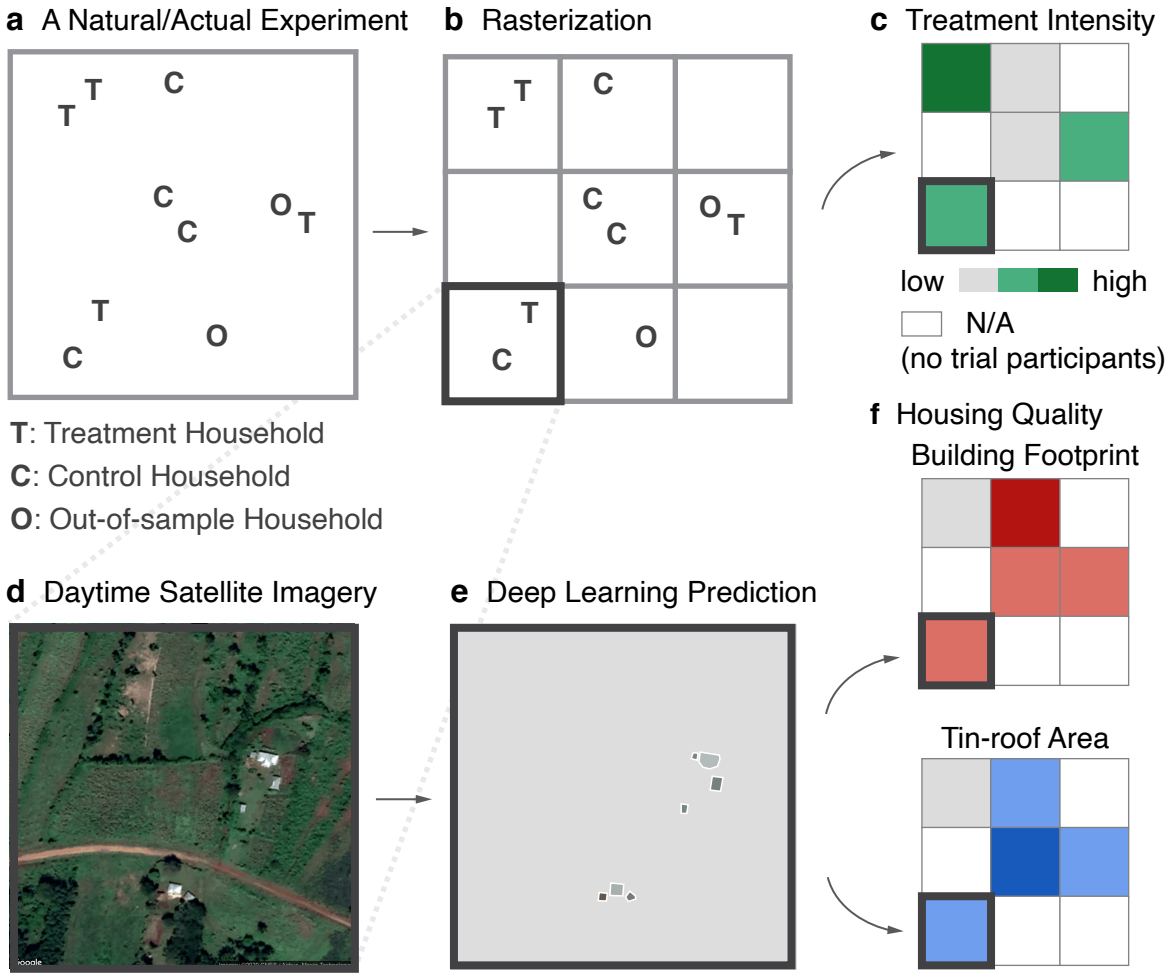

Figure 2.1: **Constructing maps of treatment intensity and remotely sensed outcomes from program implementation records and satellite imagery.** **a** An illustration of geocoded program implementation records. **b** Placing a regular grid over **a** and measuring the intensity of the treatment in each grid cell. **c** Constructed raster of treatment intensity. **d** An example daytime satellite image from Google Static Maps. **e** Example deep learning predictions on **d**. Each building is outlined in white and filled with the "representative" roof color. **f** Constructed rasters of remotely sensed housing quality outcomes. In **c** and **f**, grid cells without trial participants are omitted.

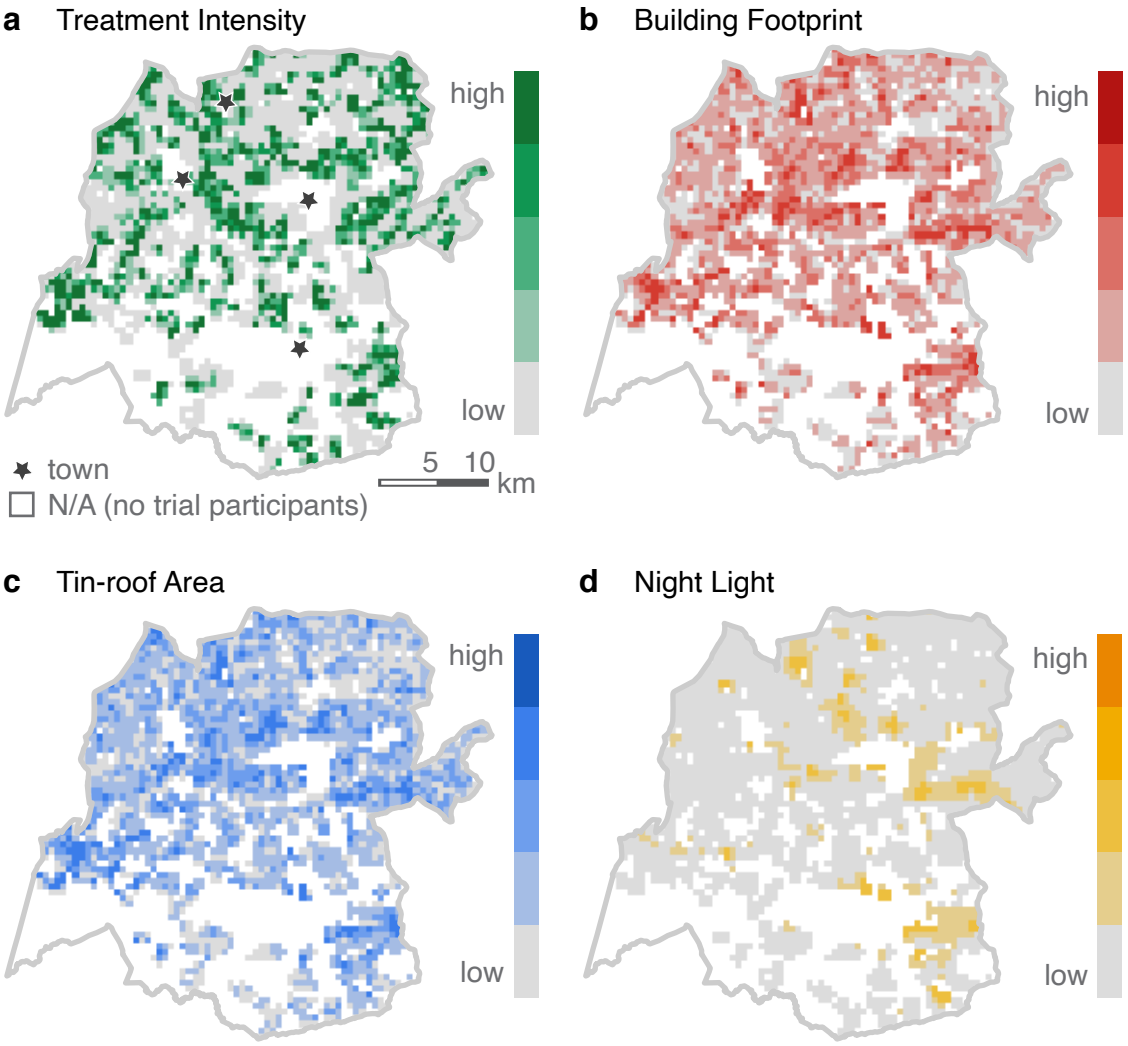

Figure 2.2: **Mapping treatment intensity and remotely sensed outcomes in the GiveDirectly study area in 2019.** **a** Treatment intensity represents the number of households who received a $1,000 cash transfer from GiveDirectly. **b** Building footprint measures the total area covered by any building. **c** Tin-roof area measures the total footprint of buildings with roofs made of tin, a high quality construction material. **d** Night light is the average radiance in the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB). In all the panels, the gray lines outline the GiveDirectly study area in Siaya, Kenya. Grid cells without trial participants are omitted.

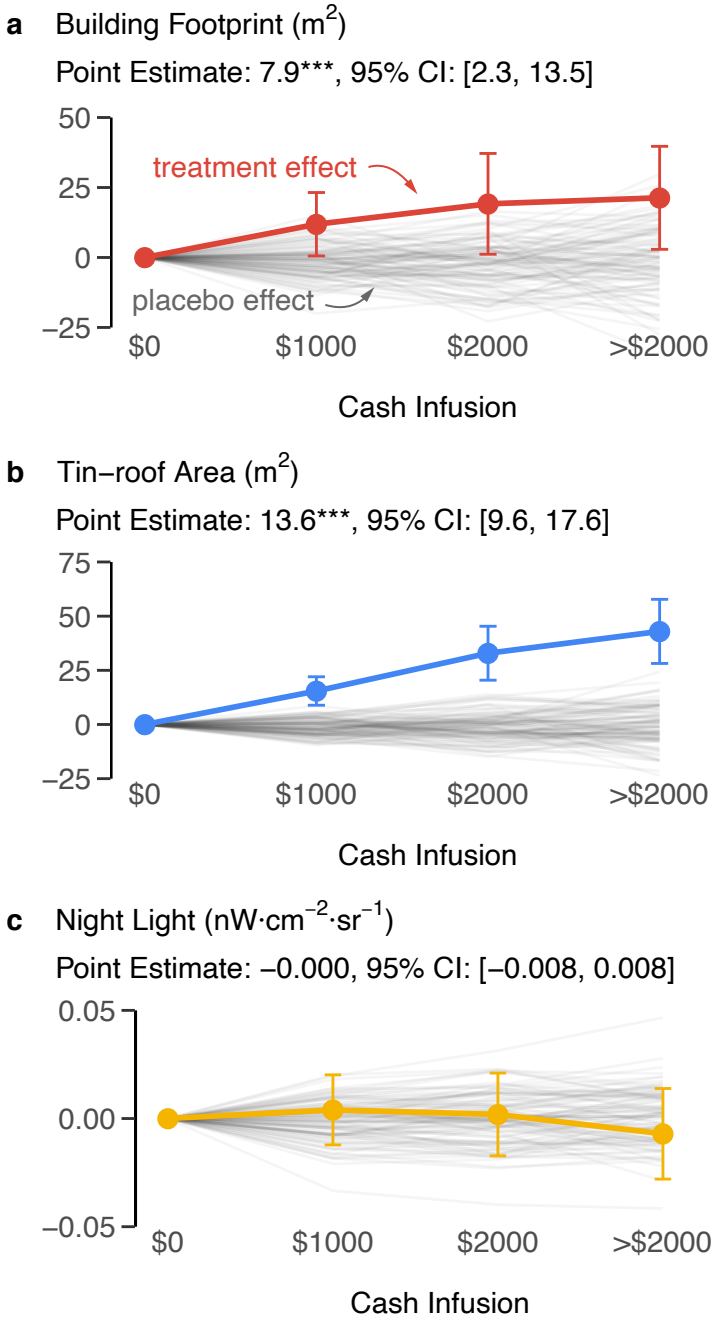

Figure 2.3: **Housing quality increased in response to the GiveDirectly cash transfer, but night light remained unchanged.** The treatment effects of the cash transfers on building footprint (**a**), tin-roof area (**b**), and night light (**c**) are shown in color. The dots represent the point estimates, and the error bars represent the 95% confidence intervals. Gray lines show the estimated effects of the placebo cash infusions from 100 simulations. The average treatment effects of a $1,000 transfer (and their 95% confidence intervals), estimated based on a constant effect assumption, are reported in the panel subtitles. *** indicates statistical significance at the 1% level.

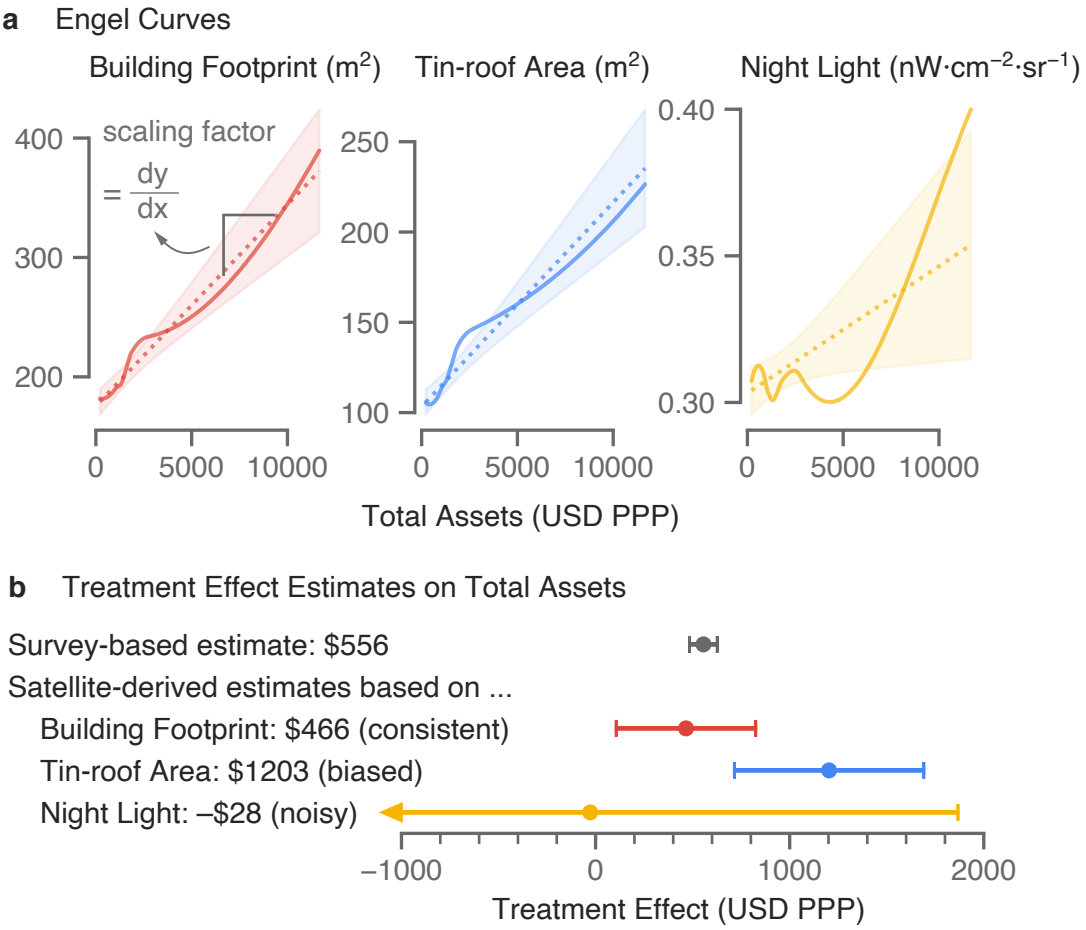

Figure 2.4: **The treatment effect on household assets can be correctly recovered by scaling the effect on building footprint.** **a** The Engel curves of building footprint, tin-roof area, and night light, estimated with LOESS (solid line) or a linear regression (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the survey-based versus satellite-derived treatment effects. The dots show the point estimates. The error bars show the 95% confidence intervals, with the arrow(s) marking upper/lower bounds that are out of range (if any).

Chapter 3

The Effect of Large-Scale Anti-Contagion Policies on the COVID-19 Pandemic

3.1 Introduction

The COVID-19 pandemic is forcing societies worldwide to make consequential policy decisions with limited information. After containment of the initial outbreak failed, attention

The materials in this chapter have been published as "The effect of large-scale anti-contagion policies on the COVID-19 pandemic" in *Nature* (Hsiang et al., 2020). It was coauthored with Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Andrew Hultgren, Emma Krasovich, Peiley Lau, Jaecheol Lee, Esther Rolf, Jeanette Tseng and Tiffany Wu. The published version, including the Extended Data figures, can be found online at `https://www.nature.com/articles/s41586-020-2404-8`. The Supplementary Information can be found at `https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-020-2404-8/MediaObjects/41586_2020_2404_MOESM1_ESM.pdf`. The datasets generated during and/or analysed during the current study are available at `https://github.com/bolliger32/gpl-covid` (DOI: 10.5281/zenodo.3832367). For easier replication, we have also created a CodeOcean "capsule", which contains a pre-built computing environment in addition to the source code and data. This is available at `https://codeocean.com/capsule/1887579/tree/v1` (DOI: 10.24433/CO.6625287.v2). Future updates and/or extensions to data or code will be listed at `http://www.globalpolicy.science/covid19`.

We thank B. Chen for her role in initiating this work and A. Feller for his feedback. S.A.-P., E.K., P.L. and J.T. are supported by a gift from the Tuaropaki Trust. T.C. is supported by an AI for Earth grant from National Geographic and Microsoft. D.A., A.H. and I.B. are supported through joint collaborations with the Climate Impact Lab. K.B. is supported by the Royal Society Te Aparangi Rutherford Postdoctoral Fellowship. H.D. and E.R. are supported by the National Science Foundation Graduate Research Fellowship under grants DGE 1106400 and 1752814, respectively. Opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of supporting organizations.

turned to implementing non-pharmaceutical interventions designed to slow contagion of the virus. In general, these policies aim to decrease virus transmission by reducing contact among individuals within or between populations, such as by closing restaurants or restricting travel, thereby slowing the spread of COVID-19 to a manageable rate. These large-scale anti-contagion policies are informed by epidemiological simulations (Kraemer et al., 2020; Li et al., 2020b; Ferguson et al., 2020; Tang et al., 2020) and a small number of natural experiments in past epidemics (Hatchett et al., 2007). However, the actual effects of these policies on infection rates in the ongoing pandemic are unknown. Because the modern world has never confronted this pathogen, nor deployed anti-contagion policies of such scale and scope, it is crucial that direct measurements of policy impacts be used alongside numerical simulations in current decision-making.

Societies around the world are weighing whether the health benefits of anti-contagion policies are worth their social and economic costs. Many of these costs are plainly seen; for example, business restrictions increase unemployment and school closures impact educational outcomes. It is therefore not surprising that some populations have hesitated before implementing such dramatic policies, especially when their costs are visible while their health benefits – infections and deaths that would have occurred but instead were avoided or delayed – are unseen. Our objective is to measure the direct health benefits of these policies; specifically, how much these policies slowed the growth rate of infections. To do this, we compare the growth rate of infections within hundreds of sub-national regions before and after each of these policies is implemented locally. Intuitively, each administrative unit observed just prior to a policy deployment serves as the "control" for the same unit in the days after it receives a policy "treatment" (see Supplementary Information for accounts of these deployments). Our hope is to learn from the recent experience of six countries where early spread of the virus triggered large-scale policy actions, in part so that societies and decision-makers in the remaining 180+ countries can access this information.

Here we directly estimate the effects of 1,717 local, regional, and national policies on the growth rate of infections across localities within China, France, Iran, Italy, South Korea, and the US (see Figure 3.1 and Supplementary Table 1). We compile subnational data on daily infection rates, changes in case definitions, and the timing of policy deployments, including (1) travel restrictions, (2) social distancing through cancellations of events and suspensions of educational/commercial/religious activities, (3) quarantines and lockdowns, and (4) additional policies such as emergency declarations and expansions of paid sick leave, from the earliest available dates to April 6, 2020 (see Supplementary Notes, also Extended Data Fig. 1). During this period, populations remained almost entirely susceptible to COVID-19, causing the natural spread of infections to exhibit almost perfect exponential growth (Ma, 2020; Muniz-Rodriguez et al., 2020). The rate of this exponential growth could change daily, determined by epidemiological factors, such as disease infectivity, as well as policies that alter behavior (Chowell et al., 2016; Ma, 2020; Tang et al., 2020). Because policies were deployed while the epidemic unfolded, we can estimate their effects empirically. To do this, we examine how the daily growth rate of infections in each locality changes in response to the collection of ongoing policies applied to that locality on that day.

3.2 Results

We employ well-established "reduced-form" econometric techniques (Angrist and Pischke, 2008; Greene, 2003) commonly used to measure the effects of events (Romer and Romer, 2010; Burke et al., 2015) on economic growth rates. Similar to early COVID-19 infections, economic output generally increases exponentially with a variable rate that can be affected by policies and other conditions. Here, this technique aims to measure the total magnitude of the effect of *changes in policy*, without requiring explicit prior information about fundamental epidemiological parameters or mechanisms, many of which remain uncertain in the current pandemic. Rather, the collective influence of these factors is empirically recovered from the data without modeling their individual effects explicitly (see Methods). Prior work on influenza (Kandula et al., 2018), for example, has shown that such statistical approaches can provide important complementary information to process-based models.

To construct the dependent variable, we transform location-specific, subnational time-series data on infections into first-differences of their natural logarithm, which is the *per-day growth rate of infections* (see Methods). We use data from first- or second-level administrative units and data on active or cumulative cases, depending on availability (see Supplementary Information). We employ widely-used panel regression models (Angrist and Pischke, 2008; Greene, 2003) to estimate how the daily growth rate of infections changes over time within a location when different combinations of large-scale policies are enacted (see Methods). Our econometric approach accounts for differences in the baseline growth rate of infections across subnational locations, which may be affected by time-invariant characteristics, such as demographics, socio-economic status, culture, and health systems; it accounts for systematic patterns in growth rates within countries unrelated to policy, such as the effect of the work-week; it is robust to systematic under-surveillance specific to each subnational unit; and it accounts for changes in procedures to diagnose positive cases (see Methods and Supplementary Information).

We estimate that in the absence of policy, early infection rates of COVID-19 grow 43% per day on average across these six countries (Standard Error [SE]= 5%), implying a doubling time of approximately 2 days. Country-specific estimates range from 34% per day in the US (SE= 7%) to 68% per day in Iran (SE= 9%). We cannot determine if the high estimate for Iran results from true epidemiological differences, data quality issues (see Methods), the concurrence of the initial outbreak with a major religious holiday and pilgrimage (see Supplementary Notes), or sampling variability. Excluding Iran, the average growth rate is 38% per day (SE= 5%). Growth rates in all five other countries are independently estimated to be very near this value (Figure 3.2a). These estimated values differ from observed average growth rates because the latter are confounded by the effects of policy. These growth rates are not driven by the expansion of testing or increasing rates of case detection (see Methods and Extended Data Fig. 2) nor by data from individual regions (Extended Data Fig. 3).

Some prior analyses of pre-intervention infections in Wuhan suggest slower growth rates (doubling every 5–7 days) (Wu et al., 2020b; Li et al., 2020a) using data collected before national standards for diagnosis and case definitions were first issued by the Chinese govern-

ment on January 15, 2020 (Tsang et al., 2020). However, case data in Wuhan from before this date contain multiple irregularities: the cumulative case count decreased on January 9; no new cases were reported during January 9-15; and there were concerns that information about the outbreak was suppressed (BBC News, 2020) (see Supplementary Table 2). When we remove these problematic data, utilizing a shorter but more reliable pre-intervention time series from Wuhan (January 16–21), we recover a growth rate of 43% per day (SE= 3%, doubling every 2 days) consistent with results from all other countries except Iran (Figure 3.2a, Supplementary Table 3).

During the early stages of an epidemic, a large proportion of the population remains susceptible to the virus, and if the spread of the virus is left uninhibited by policy or behavioral change, exponential growth continues until the fraction of the susceptible population declines meaningfully (Fisman et al., 2014; Maier and Brockmann, 2020; Chowell et al., 2016; Ma, 2020). After correcting for estimated rates of case-detection (Russell et al., 2020), we compute that the minimum susceptible fraction across administrative units in our sample is 72% of the total population (Cremona, Italy) and 87% of units would likely be in a regime of uninhibited exponential growth ($> 95\%$ susceptible) if policies were removed on the last date of our sample.

Consistent with predictions from epidemiological models (Bootsma and Ferguson, 2007; Ferguson et al., 2020; Hatchett et al., 2007), we find that the combined effect of policies within each country reduces the growth rate of infections by a substantial and statistically significant amount (Figure 3.2b, Supplementary Table 3). For example, a locality in France with a baseline growth rate of 0.33 (national average) that fully deployed all policy actions used in France would be expected to lower its daily growth rate by -0.17 to a growth rate of 0.16. In general, the estimated total effects of policy packages are large enough that they can in principle offset a large fraction of, or even eliminate, the baseline growth rate of infections—although in several countries, many localities have not deployed the full set of policies. Overall, the estimated effects of all policies combined are generally insensitive to withholding regional (i.e. state- or province-level) blocks of data from the sample (Extended Data Fig. 3).

In China, only three policies were enacted across 116 cities early in a seven week period, providing us with sufficient data to empirically estimate how the effects of these policies evolved over time without making assumptions about the timing of these effects (see Methods and Fig. 3.2b). We estimate that the combined effect of these policies reduced the growth rate of infections by -0.026 (SE= 0.046) in the first week following their deployment, increasing substantially in the second week to -0.20 (SE= 0.049), and essentially stabilizing in the third week near -0.28 (SE= 0.047). In other countries, we lack sufficient data to estimate these temporal dynamics explicitly and only report the average pooled effect of policies across all days following their deployment (see Methods). If other countries have transient responses similar to China, we would expect effects in the first week following deployment to be smaller in magnitude than the average effect we report. In Extended Data Fig. 5a and Supplementary Methods Section 3, we explore how our estimates would change if we impose the assumption that policies cannot affect infection growth rates until after a fixed number

of days; however, we do not find evidence this improves model fit.

The estimates above (Figure 3.2b) capture the superposition of all policies deployed in each country, i.e., they represent the average effect of policies that we would expect to observe if all policies enacted anywhere in each country were implemented simultaneously in a single region of that country. We also estimate the effects of individual policies or clusters of policies (Figure 3.2c) that are grouped based on either their similarity in goal (e.g., library and museum closures) or timing (e.g., policies deployed simultaneously). Our estimates for these individual effects tend to be statistically noisier than the estimates for all policies combined. Some estimates for the same policy differ between countries, perhaps because policies are not implemented identically or because populations behave differently. Nonetheless, 22 out of 29 point estimates indicate that individual policies are likely contributing to reducing the growth rate of infections. Seven policies (one in South Korea, two in Italy, and four in the US) have point estimates that are positive, six of which are small in magnitude (< 0.1) and not statistically different from zero (5% level). Consistent with greater overall uncertainty in these dis-aggregated estimates, some in China, South Korea, Italy, and France are somewhat more sensitive to withholding regional blocks of data (Extended Data Fig. 4), but remain broadly robust to assuming a constant delayed effect of all policies (Extended Data Fig. 5b).

Based on these results, we find that the deployment of anti-contagion policies in all six countries significantly and substantially slowed the pandemic. We combine the estimates above with our data on the timing of the 1,717 policy deployments to estimate the total effect of all policies across the dates in our sample. To do this, we use our estimates to predict the growth rate of infections in each locality on each day, given the actual policies in effect at that location on that date (Figure 3.3, blue markers). We then use the same model to predict what counterfactual growth rates would be on that date if the effects of all policies were removed (Figure 3.3, red markers), which we call the "no-policy scenario." The difference between these two predictions is our estimated effect that all deployed policies had on the growth rate of infections. During our sample, we estimate that all policies combined slowed the average growth rate of infections by -0.252 per day (SE$= 0.045$, p< 0.001) in China, -0.248 (SE$= 0.089$, p< 0.01) in South Korea, -0.24 (SE$= 0.068$, p< 0.001) in Italy, -0.355 (SE$= 0.063$, p< 0.001) in Iran, -0.123 (SE$= 0.019$, p< 0.001) in France and -0.084 (SE$= 0.03$, p< 0.01) in the US. These results are robust to modeling the effects of policies without grouping them (Extended Data Fig. 6a and Supplementary Table 4) or assuming a delayed effect of policy on infection growth rates (Supplementary Table 5).

The number of COVID-19 infections on a date depends on the growth rate of infections on all prior days. Thus, persistent reductions in growth rates have a compounding effect on infections, until growth is slowed by a shrinking susceptible population. To provide a sense of scale for our results, we integrate the growth rate of infections in each locality from Figure 3.3 to estimate cumulative infections, both with actual anti-contagion policies and in the no-policy counterfactual scenario. To account for the declining susceptible population in each administrative unit, we couple our econometric estimates of the effects of policies with a Susceptible-Infected-Removed (SIR) model (Ma, 2020; Chowell et al., 2016) that adjusts the susceptible population in each administrative unit based on estimated case-detection rates

(Russell et al., 2020; Meyerowitz-Katz and Merone, 2020) (see Methods). This allows us to extend our projections beyond the initial exponential growth phase of infections, a threshold that many localities cross in our no-policy scenario.

Our results suggest that ongoing anti-contagion policies have already substantially reduced the number of COVID-19 infections observed in the world today (Figure 3.4). Our central estimates suggest that there would be roughly 37 million more cumulative confirmed cases (corresponding to 285 million more total infections, including the confirmed cases) in China, 11.5 million more confirmed cases in South Korea (38 million total infections), 2.1 million more confirmed cases in Italy (49 million total infections), 5 million more confirmed cases in Iran (54 million total infections), 1.4 million more confirmed cases in France (45 million total infections), and 4.8 million more confirmed cases (60 million total infections) in the US had these countries never enacted any anti-contagion policies since the start of the pandemic. The magnitudes of these impacts partially reflect the timing, intensity, and extent of policy deployment (e.g., how many localities deployed policies), and the duration for which they have been applied. Several of these estimates are subject to large statistical uncertainties (see intervals in Figure 3.4). Sensitivity tests (Extended Data Fig. 7) that assume a range of plausible alternative parameter values relating to disease dynamics, such as incorporating a Susceptible-Exposed-Infected-Removed (SEIR) model, suggest that interventions may have reduced the severity of the outbreak by a total of 55–66 million confirmed cases over the dates in our sample (central estimates). Sensitivity tests varying the assumed infection-fatality ratio (Supplementary Table 6) suggest a corresponding range of 46–77 million confirmed cases (490–580 million total infections).

3.3 Discussion

Our empirical results indicate that large-scale anti-contagion policies are slowing the COVID-19 pandemic. Because infection rates in the countries we study would have initially followed rapid exponential growth had no policies been applied, our results suggest that these policies have provided large health benefits. For example, we estimate that there would be roughly 465× the observed number of confirmed cases in China, 17× in Italy, and 14× in the US by the end of our sample if large-scale anti-contagion policies had not been deployed. Consistent with process-based simulations of COVID-19 infections (Kraemer et al., 2020; Tang et al., 2020; Ferguson et al., 2020; Maier and Brockmann, 2020; Li et al., 2020b; Kucharski et al., 2020), our analysis of existing policies indicates that seemingly small delays in policy deployment likely produced dramatically different health outcomes.

While the limitations of available data pose challenges to our analysis, our aim is to use what data exist to estimate the first-order impacts of unprecedented policy actions in an ongoing global crisis. As more data become available, related findings will become more precise and may capture more complex interactions. Furthermore, this analysis does not account for interactions between populations in nearby localities (Chowell et al., 2016), nor mobility networks (Li et al., 2020b; Tang et al., 2020; Chinazzi et al., 2020; Kraemer et al.,

2020). Nonetheless, we hope these results can support critical decision-making, both in the countries we study and in the other 180+ countries where COVID-19 infections have been reported (WHO, 2020).

A key advantage of our reduced-form "top down" statistical approach is that it captures the real-world behavior of affected populations without requiring that we explicitly model underlying mechanisms and processes. This is useful in the current pandemic where many process-related parameters remain uncertain. However, our results cannot and should not be interpreted as a substitute for "bottom up" process-based epidemiological models specifically designed to provide guidance in public health crises. Rather, our results complement existing models, for example, by helping to calibrate key model parameters. We believe both forward-looking simulations and backward-looking empirical evaluations should be used to inform decision-making.

Our analysis measures changes in local infection growth rates associated with changes in anti-contagion policies. A necessary condition for this association to be interpreted as the plausibly causal effect of these policies is that the timing of policy deployment is independent of infection growth rates (Angrist and Pischke, 2008). This assumption is supported by established epidemiological theory (Chowell et al., 2016; Ma, 2020; Anderson and May, 1992) and evidence (Nishiura et al., 2010; WHO Ebola Response Team, 2014), which indicate that infections in the absence of policy will grow exponentially early in the epidemic, implying that pre-policy infection growth rates should be constant over time and therefore uncorrelated with the timing of policy deployment. Further, scientific guidance to decision-makers early in the current epidemic explicitly projected constant growth rates in the absence of anti-contagion measures, limiting the possibility that anticipated changes in natural growth rates affected decision-making (Ferguson et al., 2020; Flaxman et al., 2020; Lourenço et al., 2020; Maier and Brockmann, 2020). In practice, policies tended to be deployed in response to high total numbers of cases (e.g. in France) (Ministère des Solidarités et de la Santé, 2020), in response to outbreaks in other regions (e.g. in China, South Korea, and Iran) (Tian et al., 2020), after delays due to political constraints (e.g. in the US and Italy), and often with timing that coincided with arbitrary events, like weekends or holidays (see Supplementary Notes for detailed chronologies).

Our analysis accounts for documented changes in COVID-19 testing procedures and availability, as well as differences in case-detection across locations; however, unobserved trends in case-detection could affect our results (see Methods). We analyze estimated case-detection trends (Russell et al., 2020) (Extended Data Fig. 2), finding that this potential bias is small, possibly elevating our estimated no-policy growth rates by 0.022 (7%) on average.

It is also possible that changing public knowledge during the period of our study affects our results. If individuals alter behavior in response to new information unrelated to anti-contagion policies, such as seeking out online resources, this could alter the growth rate of infections and thus affect our estimates. If increasing availability of information reduces infection growth rates, it would cause us to overstate the effectiveness of anti-contagion policies. We note, however, that if public knowledge is increasing in response to policy actions, such as through news reports, then it should be considered a pathway through

which policies alter infection growth, not a form of bias. Investigating these potential effects is beyond the scope of this analysis, but it is an important topic for future investigations.

Finally, our analysis focuses on confirmed infections, but other outcomes, such as hospitalizations or deaths, are also of policy interest. Future work on these outcomes may require additional modeling approaches because they are relatively more context- and state-dependent. Nonetheless, we experimentally implement our approach on the daily growth rate of hospitalizations in France, where hospitalization data is available at the granularity of this study. We find that the total estimated effect of anti-contagion policies on the growth rate of hospitalizations is similar to our estimates for infection growth rates (Extended Data Fig. 6c).

3.4 Methods

Data Collection and Processing

We provide a brief summary of our data collection processes here (see the Supplementary Notes for more details, including access dates). Epidemiological, case definition/testing regime, and policy data for each of the six countries in our sample were collected from a variety of in-country data sources, including government public health websites, regional newspaper articles, and crowd-sourced information on Wikipedia. The availability of epidemiological and policy data varied across the six countries, and preference was given to collecting data at the most granular administrative unit level. The country-specific panel datasets are at the region level in France, the state level in the US, the province level in South Korea, Italy and Iran, and the city level in China. Due to data availability, the sample dates differ across countries: in China we use data from January 16 - March 5, 2020; in South Korea from February 17 - April 6, 2020; in Italy from February 26 - April 6, 2020; in Iran from February 27 - March 22, 2020; in France from February 29 - March 25, 2020; and in the US from March 3 - April 6, 2020. Below, we describe our data sources.

China We acquired epidemiological data from an open source GitHub project (Lin, 2020) that scrapes time series data from Ding Xiang Yuan. We extended this dataset back in time to January 10, 2020 by manually collecting official daily statistics from the central and provincial (Hubei, Guangdong, and Zhejiang) Chinese government websites. We compiled policies by collecting data on the start dates of travel bans and lockdowns at the city-level from the "2020 Hubei lockdowns" Wikipedia page (Wikipedia, 2020c) and various other news reports. We suspect that most Chinese cities have implemented at least one anti-contagion policy due to their reported trends in infections; as such, we dropped cities where we could not identify a policy deployment date to avoid miscategorizing the policy status of these cities. Thus our results are only representative for the sample of 116 cities for which we obtained policy data.

South Korea We manually collected and compiled the epidemiological dataset in South Korea, based on provincial government reports, policy briefings, and news articles. We compiled policy actions from news articles and press releases from the Korean Centers for Disease Control and Prevention (KCDC), the Ministry of Foreign Affairs, and local governments' websites.

Iran We used epidemiological data from the table "New COVID-19 cases in Iran by province" (Wikipedia, 2020a) in the "2020 coronavirus pandemic in Iran" Wikipedia article, which were compiled from data provided on the Iranian Ministry of Health website (in Persian). We relied on news media reporting and two timelines of pandemic events in Iran (Think Global Health, 2020; Wikipedia, 2020a) to collate policy data. From March 2-3, Iran did not report subnational cases. Around this period the country implemented three national policies: a recommendation against local travel (3/1), work from home for government employees (3/3), and school closure (3/5). As the effects of these policies cannot be distinguished from each other due to the data gap, we group them for the purpose of this analysis.

Italy We used epidemiological data from the GitHub repository (Presidenza del Consiglio dei Ministri, 2020) maintained by the Italian Department of Civil Protection (Dipartimento della Protezione Civile). For policies, we primarily relied on the English version of the COVID-19 dossier "Chronology of main steps and legal acts taken by the Italian Government for the containment of the COVID-19 epidemiological emergency" written by the Dipartimento della Protezione Civile (Civil Protection Department Website - Presidency of the Council of Ministers, 2020), and Wikipedia (Wikipedia, 2020d).

France We used the region-level epidemiological dataset provided by France's government website (Roussel, 2020) and supplemented it with numbers of confirmed cases by region on France's public health website, which was previously updated daily through March 25 (Sante Publique France, 2020). We obtained data on France's policy response to the COVID-19 pandemic from the French government website, press releases from each regional public health site (Agence Régionale de Santé, 2020) and Wikipedia (Wikipedia, 2020b).

United States We used state-level epidemiological data from usafacts.org (USA Facts, 2020) which they compile from multiple sources. For policy responses, we relied on a number of sources, including the U.S. Centers for Disease Control (CDC), the National Governors Association, as well as various executive orders from county- and city-level governments, and press releases from media outlets.

Policy Data Policies in administrative units were coded as binary variables, where the policy was coded as either 1 (after the date that the policy was implemented, and before it was removed) or 0 otherwise, for the affected administrative units. When a policy only

affected a fraction of an administrative unit (e.g., half of the counties within a state), policy variables were weighted by the percentage of people within the administrative unit who were treated by the policy. We used the most recent population estimates we could find for countries' administrative units (see the Population Data section in the Appendix). In order to standardize policy types across countries, we mapped each country-specific policy to one of the broader policy category variables in our analysis. In this exercise, we collected 168 policies for China, 59 for South Korea, 214 for Italy, 23 for Iran, 59 for France, and 1,194 for the United States (see Supplementary Table 1). There are some cases where we encode policies that are necessarily in effect whenever another policy is in place, due in particular to the far-reaching implications of home isolation policies. In China, wherever home isolation is documented, we assume a local travel ban is enacted on the same day if we have not found an explicit local travel ban policy for a given locality. In France, we assume home isolation is accompanied by event cancellations, social distancing, and no-gathering policies; in Italy, we assume home isolation entails no-gathering, local travel ban, work from home, and social distancing policies; in the US, we assume shelter-in-place orders indicate that non-essential business closures, work from home policies, and no-gathering policies are in effect. For policy types that are enacted multiple times at increasing degrees of intensity within a locality, we add weights to the variable by escalating the intensity from 0 pre-policy in steps up to 1 for the final version of the policy (see the Policy Data section in the Appendix).

Epidemiological Data We collected information on cumulative confirmed cases, cumulative recoveries, cumulative deaths, active cases, and any changes to domestic COVID-19 testing regimes, such as case definitions or testing methodology. For our regression analysis (Figure 3.2), we use active cases when they are available (for China and South Korea) and cumulative confirmed cases otherwise. We document quality control steps in the Appendix. Notably, for China and South Korea we acquired more granular data than the data hosted on the Johns Hopkins University (JHU) interactive dashboard;(the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, 2020) we confirm that the number of confirmed cases closely match between the two data sources (see Extended Data Fig. 1). To conduct the econometric analysis, we merge the epidemiological and policy data to form a single data set for each country.

Econometric analysis

Reduced-Form Approach The reduced-form econometric approach that we apply here is a "top down" approach that describes the behavior of aggregate outcomes y in data (here, infection rates). This approach can identify plausibly causal effects (Angrist and Pischke, 2008; Greene, 2003) induced by exogenous changes in independent policy variables z (e.g., school closure) without explicitly describing all underlying mechanisms that link z to y, without observing intermediary variables x (e.g., behavior) that might link z to y, or without other determinants of y unrelated to z (e.g., demographics), denoted w. Let $f(\cdot)$

describe a complex and unobserved process that generates infection rates y:

$$y = f(x_1(z_1, \ldots, z_K), \ldots, x_N(z_1, \ldots, z_K), w_1, \ldots, w_M) \tag{3.1}$$

Process-based epidemiological models aim to capture elements of $f(\cdot)$ explicitly, and then simulate how changes in z, x, or w affect y. This approach is particularly important and useful in forward-looking simulations where future conditions are likely to be different than historical conditions. However, a challenge faced by this approach is that we may not know the full structure of $f(\cdot)$, for example if a pathogen is new and many key biological and societal parameters remain uncertain. Crucially, we may not know the effect that large-scale policy (z) will have on behavior ($x(z)$) or how this behavior change will affect infection rates ($f(\cdot)$).

Alternatively, one can differentiate Equation 3.1 with respect to the k^{th} policy z_k:

$$\frac{\partial y}{\partial z_k} = \sum_{j=1}^{N} \frac{\partial y}{\partial x_j} \frac{\partial x_j}{\partial z_k} \tag{3.2}$$

which describes how changes in the policy affects infections through all N potential pathways mediated by $x_1, ..., x_N$. Usefully, for a fixed population observed over time, empirically estimating an average value of the local derivative on the left-hand-side in Equation 3.2 does not depend on explicit knowledge of w. If we can observe y and z directly and estimate changes over time $\frac{\partial y}{\partial z_k}$ with data, then intermediate variables x also need not be observed nor modeled. The reduced-form econometric approach(Angrist and Pischke, 2008; Greene, 2003) thus attempts to measure $\frac{\partial y}{\partial z_k}$ directly, exploiting exogenous variation in policies z.

Model Active infections grow exponentially during the initial phase of an epidemic, when the proportion of immune individuals in a population is near zero. Assuming a simple Susceptible-Infected-Recovered (SIR) disease model (Ma, 2020), the growth in infections during the early period is

$$\frac{dI_t}{dt} = (S_t\beta - \gamma)I_t \underset{S_t \to 1}{=} (\beta - \gamma)I_t, \tag{3.3}$$

where I_t is the number of infected individuals at time t, β is the transmission rate (new infections per day per infected individual), γ is the removal rate (proportion of infected individuals recovering or dying each day) and S is the fraction of the population susceptible to the disease. The second equality holds in the limit $S \to 1$, which describes the current conditions during the beginning of the COVID-19 pandemic. The solution to this ordinary differential equation is the exponential function

$$\frac{I_{t_2}}{I_{t_1}} = e^{g \cdot (t_2 - t_1)}, \tag{3.4}$$

where I_{t_1} is the initial condition. Taking the natural logarithm and rearranging, we have

$$\log(I_{t_2}) - \log(I_{t_1}) = g \cdot (t_2 - t_1). \tag{3.5}$$

Anti-contagion policies are designed to alter g, through changes to β, by reducing contact between susceptible and infected individuals. Holding the time-step between observations fixed at one day ($t_2 - t_1 = 1$), we thus model g as a time-varying outcome that is a linear function of a time-varying policy

$$g_t = \log(I_t) - \log(I_{t-1}) = \theta_0 + \theta \cdot policy_t + \epsilon_t, \tag{3.6}$$

where θ_0 is the average growth rate absent policy, $policy_t$ is a binary variable describing whether a policy is deployed at time t, and θ is the average effect of the policy on growth rate g over all periods subsequent to the policy's introduction, thereby encompassing any lagged effects of policies. ϵ_t is a mean-zero disturbance term that captures inter-period changes not described by $policy_t$. Using this approach, infections each day are treated as the initial conditions for integrating Equation 3.4 through to the following day.

We compute the first differences $\log(I_t) - \log(I_{t-1})$ using active infections where they are available, otherwise we use cumulative infections, noting that they are almost identical during this early period (except in China, where we use active infections). We then match these data to policy variables that we construct using the novel data sets we assemble and apply a reduced-form approach to estimate a version of Equation 3.6, although the actual expression has additional terms detailed below.

Estimation To estimate a multi-variable version of Equation 3.6, we estimate a separate regression for each country c. Observations are for subnational units indexed by i observed for each day t. Because not all localities began testing for COVID-19 on the same date, these samples are unbalanced panels. To ensure data quality, we restrict our analysis to localities after they have reported at least ten cumulative infections.

A necessary condition for unbiased estimates is that the timing of policy deployment is independent of natural infection growth rates (Angrist and Pischke, 2008), a mathematical condition that should be true in the context of a new epidemic. In established epidemiological models, including the standard SIR model above, early rates of infection within a susceptible population are characterized by constant exponential growth. This phenomenon is well understood theoretically (Chowell et al., 2016; Anderson and May, 1992; Kermack and McKendrick, 1927), has been repeatedly documented in past epidemics (WHO Ebola Response Team, 2014; Nishiura et al., 2010; Mills et al., 2004) as well as the current COVID-19 pandemic (Ma, 2020; Muniz-Rodriguez et al., 2020), and implies constant infection growth rates in the absence of policy intervention. Thus, we treat changes in infection growth rates as conditionally independent of policy deployments since the correlation between a constant variable and any other variable is zero in expectation.

We estimate a multiple regression version of Equation 3.6 using ordinary least squares. We include a vector of subnational unit-fixed effects θ_0 (i.e., varying intercepts captured

as coefficients to dummy variables) to account for all time-invariant factors that affect the local growth rate of infections, such as differences in demographics, socio-economic status, culture, and health systems (Greene, 2003). We include a vector of day-of-week-fixed effects δ to account for weekly patterns in the growth rate of infections that are common across locations within a country, however, in China, we omit day-of-week effects because we find no evidence they are present in the data – perhaps due to the fact that the outbreak of COVID-19 began during a national holiday and workers never returned to work. We also include a separate single-day dummy variable each time there is an abrupt change in the availability of COVID-19 testing or a change in the procedure to diagnose positive cases. Such changes generally manifest as a discontinuous jump in infections and a re-scaling of subsequent infection rates (e.g., See China in Figure 3.1), effects that are flexibly absorbed by a single-day dummy variable because the dependent variable is the first-difference of the logarithm of infections. We denote the vector of these testing dummies μ.

Lastly, we include a vector of P_c country-specific policy variables for each location and day. These policy variables take on values between zero and one (inclusive) where zero indicates no policy action and one indicates a policy is fully enacted. In cases where a policy variable captures the effects of collections of policies (e.g., museum closures and library closures), a policy variable is computed for each, then they are averaged, so the coefficient on this type of variable is interpreted as the effect if all policies in the collection are fully enacted. There are also instances where multiple policies are deployed on the same date in numerous locations, in which case we group policies that have similar objectives (e.g., suspension of transit and travel ban, or cancelling of events and no gathering) and keep other policies separate (i.e., business closure, school closure). The grouping of policies is useful for reducing the number of estimated parameters in our limited sample of data, allowing us to examine the impact of subsets of policies (e.g. Fig. 3.2c). However, policy grouping does not have a material impact on the estimated effect of all policies combined nor on the effect of actual policies, which we demonstrate by estimating a regression model where no policies are grouped and these values are recalculated (Supplementary Table 4, Extended Data Fig. 6).

In some cases (for Italy and the US), policy data is available at a more spatially granular level than infection data (e.g., city policies and state-level infections in the US). In these cases, we code binary policy variables at the more granular level and use population-weights to aggregate them to the level of the infection data. Thus, policy variables may take on continuous values between zero and one, with a value of one indicating that the policy is fully enacted for the entire population. Given the limited quantity of data currently available, we use a parsimonious model that assumes the effects of policies on infection growth rates are approximately linear and additively separable. However, future work that possesses more data may be able to identify important nonlinearities or interactions between policies.

For each country, our general multiple regression model is thus

$$g_{cit} = \log(I_{cit}) - \log(I_{ci,t-1}) = \theta_{0,ci} + \delta_{ct} + \mu_{cit} + \sum_{p=1}^{P_c} (\theta_{cp} \cdot policy_{pcit}) + \epsilon_{cit} \qquad (3.7)$$

where observations are indexed by country c, subnational unit i, and day t. The parameters of interest are the country-by-policy specific coefficients θ_{cp}. We display the estimated residuals ϵ_{cit} in Extended Data Fig. 10, which are mean zero but not strictly normal (normality is not a requirement of our modeling and inference strategy), and we estimate uncertainty over all parameters by calculating our standard errors robust to error clustering at the day level (Angrist and Pischke, 2008). This approach allows the covariance in ϵ_{cit} across different locations within a country, observed on the same day, to be nonzero. Such clustering is important in this context because idiosyncratic events within a country, such as a holiday or a backlog in testing laboratories, could generate nonuniform country-wide changes in infection growth for individual days not explicitly captured in our model. Thus, this approach non-parametrically accounts for both arbitrary forms of spatial auto-correlation or systematic misreporting in regions of a country on any given day (we note that it generates larger estimates for uncertainty than clustering by i). When we report the effect of all policies combined (e.g., Figure 3.2b) we are reporting the sum of coefficient estimates for all policies $\sum_{p=1}^{P_c} \theta_{cp}$, accounting for the covariance of errors in these estimates when computing the uncertainty of this sum.

Note that our estimates of θ and θ_0 in Equation 3.7 are robust to systematic under-reporting of infections, a major concern in the ongoing pandemic, due to the construction of our dependent variable. This remains true even if different localities have different rates of under-reporting, so long as the rate of under-reporting is relatively constant. To see this, note that if each locality i has a medical system that reports only a fraction ψ_i of infections such that we observe $\tilde{I}_{it} = \psi_i I_{it}$ rather an actual infections I_{it}, then the left-hand-side of Equation 3.7 will be

$$\begin{aligned}\log(\tilde{I}_{it}) - \log(\tilde{I}_{i,t-1}) &= \log(\psi_i I_{it}) - \log(\psi_i I_{i,t-1}) \\ &= \log(\psi_i) - \log(\psi_i) + \log(I_{it}) - \log(I_{i,t-1}) \\ &= \log(I_{it}) - \log(I_{i,t-1}) = g_t\end{aligned}$$

and is therefore unaffected by location-specific and time-invariant under-reporting. Thus systematic under-reporting does not affect our estimates for the effects of policy θ. As discussed above, potential biases associated with non-systematic under-reporting resulting from documented changes in testing regimes over space and time are absorbed by region-day specific dummies μ.

However, if the rate of under-reporting within a locality is changing day-to-day, this could bias infection growth rates. We estimate the magnitude of this bias (see Extended Data Fig. 2), and verify that it is quantitatively small. Specifically, if $\tilde{I}_{it} = \psi_{it} I_{it}$ where ψ_{it} changes day-to-day, then

$$\log(\tilde{I}_{it}) - \log(\tilde{I}_{i,t-1}) = \log(\psi_{it}) - \log(\psi_{i,t-1}) + g_t \tag{3.8}$$

where $\log(\psi_{it}) - \log(\psi_{i,t-1})$ is the day-over-day growth rate of the case-detection probability. Disease surveillance has evolved slowly in some locations as governments gradually expand

testing, which would cause ψ_{it} to change over time, but these changes in testing capacity do not appear to significantly alter our estimates of infection growth rates. In Extended Data Fig. 2, we show one set of epidemiological estimates (Russell et al., 2020) for $\log(\psi_{it}) - \log(\psi_{i,t-1})$. Despite random day-to-day variations, which do not cause systematic biases in our point estimates, the mean of $\log(\psi_{it}) - \log(\psi_{i,t-1})$ is consistently small across the different countries: 0.05 in China, 0.064 in Iran, 0.019 in South Korea, -0.058 in France, 0.031 in Italy, and 0.049 in the US. The average of these estimates is 0.026, potentially accounting for 7.3% of our global average estimate for the no-policy infection growth rate (0.36). These estimates of $\log(\psi_{it}) - \log(\psi_{i,t-1})$ also do not display strong temporal trends, alleviating concerns that time-varying under-reporting generates sizable biases in our estimated effects of anti-contagion policies.

Transient dynamics In China, we are able to examine the transient response of infection growth rates following policy deployment because only three policies were deployed early in a seven-week sample period during which we observe many cities simultaneously. This provides us with sufficient data to estimate the temporal structure of policy effects without imposing assumptions regarding this structure. To do this, we estimate a distributed-lag model that encodes policy parameters using weekly lags based on the date that each policy is first implemented in locality i. This means the effect of a policy implemented one week ago is allowed to differ arbitrarily from the effect of that same policy in the following week, etc. These effects are then estimated simultaneously and are displayed in Fig. 3.2 (also Supplementary Table 3). Such a distributed lag approach did not provide statistically meaningful insight in other countries using currently available data because there were fewer administrative units and shorter periods of observation (i.e. smaller samples), and more policies (i.e. more parameters to estimate) in all other countries. Future work may be able to successfully explore these dynamics outside of China.

As a robustness check, we examine whether excluding the transient response from the estimated effects of policy substantially alters our results. We do this by estimating a "fixed lag" model, where we assume that policies cannot influence infection growth rates for L days, recoding a policy variable at time t as zero if a policy was implemented fewer than L days before t. We re-estimate Equation 3.7 for each value of L and present results in Extended Data Fig. 5 and Supplementary Table 5.

Alternative disease models Our main empirical specification is motivated with an SIR model of disease contagion, which assumes zero latent period between exposure to COVID-19 and infectiousness. If we relax this assumption to allow for a latent period of infection, as in a Susceptible-Exposed-Infected-Recovered (SEIR) model, the growth of the outbreak is only asymptotically exponential (Ma, 2020). Nonetheless, we demonstrate that SEIR dynamics have only a minor potential impact on the coefficients recovered by using our empirical approach in this context. In Extended Data Figs. 8 and 9 we present results from a simulation exercise which uses Equations 3.9–3.11, along with a generalization to the SEIR

model (Ma, 2020) to generate synthetic outbreaks (see Supplementary Methods Section 2). We use these simulated data to test the ability of our statistical model (Equation 3.7) to recover both the unimpeded growth rate (Extended Data Fig. 8) as well as the impact of simulated policies on growth rates (Extended Data Fig. 9) when applied to data generated by SIR or SEIR dynamics over a wide range of epidemiological conditions.

Projections

Daily growth rates of infections To estimate the instantaneous daily growth rate of infections if policies were removed, we obtain fitted values from Equation 3.7 and compute a predicted value for the dependent variable when all P_c policy variables are set to zero. Thus, these estimated growth rates $\hat{g}_{cit}^{no\ policy}$ capture the effect of all locality-specific factors on the growth rate of infections (e.g., demographics), day-of-week-effects, and adjustments based on the way in which infection cases are reported. This counterfactual does not account for changes in information that are triggered by policy deployment, since those should be considered a pathway through which policies affect outcomes, as discussed in the main text. Additionally, the "no-policy" counterfactual does not model previously unobserved changes in behavior that might occur if fundamentally new behaviors emerge even in the absence of government intervention. When we report an average no-policy growth rate of infections (Figure 3.2a), it is the average value of these predictions for all observations in the original sample. Location-and-day specific counterfactual predictions ($\hat{g}_{cit}^{no\ policy}$), accounting for the covariance of errors in estimated parameters, are shown as red markers in Figure 3.3.

Cumulative infections To provide a sense of scale for the estimated cumulative benefits of effects shown in Figure 3.3, we link our reduced-form empirical estimates to the key structures in a simple SIR system and simulate this dynamical system over the course of our sample. The system is defined as the following:

$$\frac{dS_t}{dt} = -\beta_t S_t I_t \tag{3.9}$$

$$\frac{dI_t}{dt} = (\beta_t S_t - \gamma) I_t \tag{3.10}$$

$$\frac{dR_t}{dt} = \gamma I_t \tag{3.11}$$

where S_t is the susceptible population and R_t is the removed population. Here β_t is a time-evolving parameter, determined via our empirical estimates as described below. Accounting for changes in S becomes increasingly important as the size of cumulative infections $(I_t + R_t)$ becomes a substantial fraction of the local subnational population, which occurs in some no-policy scenarios. Our reduced-form analysis provides estimates for the growth rate of active infections $(\hat{g})$ for each locality and day, in a regime where $S_t \approx 1$. Thus we know

$$\frac{dI_t}{dt}/I_t\Big|_{S\approx 1} = \hat{g}_t = \beta_t - \gamma \tag{3.12}$$

but we do not know the values of either of the two right-hand-side terms, which are required to simulate Equations 3.9–3.11. To estimate γ, we note that the left-hand-side term of Equation 3.11 is

$$\frac{dR_t}{dt} \approx \frac{d}{dt}(\text{cumulative\_recoveries} + \text{cumulative\_deaths})$$

which we can observe in our data for China and South Korea. Computing first differences in these two variables (to differentiate with respect to time), summing them, and then dividing by active cases gives us estimates of γ (medians: China=0.11, Korea=0.05). These values differ slightly from the classical SIR interpretation of γ because in the public data we are able to obtain, individuals are coded as "recovered" when they no longer test positive for COVID-19, whereas in the classical SIR model this occurs when they are no longer infectious. We adopt the average of these two medians, setting $\gamma = .08$. We use medians rather than simple averages because low values for I induce a long right-tail in daily estimates of γ and medians are less vulnerable to this distortion. We then use our empirically-based reduced-form estimates of $\hat{g}$ (both with and without policy) combined with Equations 3.9–3.11 to project total cumulative cases in all countries, shown in Figure 3.4. We simulate infections and cases for each administrative unit in our sample beginning on the first day for which we observe 10 or more cases (for that unit) using a time-step of 4 hours. Because we observe confirmed cases rather than total infections, we seed each simulation by adjusting observed I_t on the first day using country-specific estimates of case detection rates. We adjust existing estimates of case under-reporting (Russell et al., 2020) to further account for asymptomatic infections assuming an infection-fatality ratio of 0.75% (Meyerowitz-Katz and Merone, 2020). We assume $R_t = 0$ on the first day. To maintain consistency with the reported data, we report our output in confirmed cases by multiplying our simulated $I_t + R_t$ values by the aforementioned proportion of infections confirmed. We estimate uncertainty by resampling from the estimated variance-covariance matrix of all regression parameters. In Extended Data Fig. 7, we show sensitivity of this simulation to the estimated value of γ as well as to the use of a Susceptible-Exposed-Infected-Recovered (SEIR) framework. In Supplementary Table 6, we show sensitivity of this simulation to the assumed infection-fatality ratio (see Supplementary Methods Section 1).

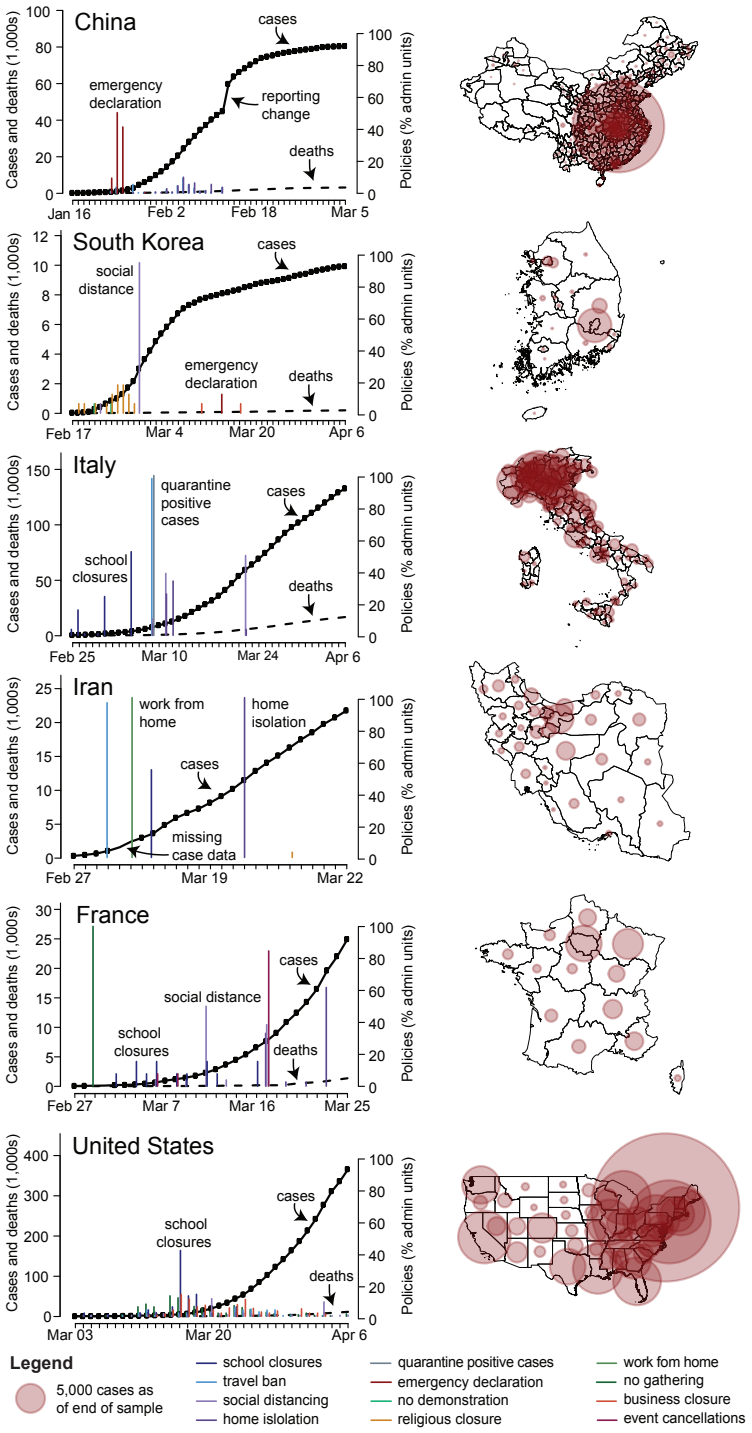

Figure 3.1: **Data on COVID-19 infections and large-scale anti-contagion policies.** Left: Daily cumulative confirmed cases of COVID-19 (solid black line, left axis) and deaths (dashed black line) over time. Vertical lines are deployments of anti-contagion policies, with height indicating the number of administrative units instituting a policy that day (right axis). For display purposes only, ≤ 5 policy types are shown per country and missing case data are imputed unless all sub-national units are missing. Right: Maps of cumulative confirmed cases by administrative unit on the last date of each sample.

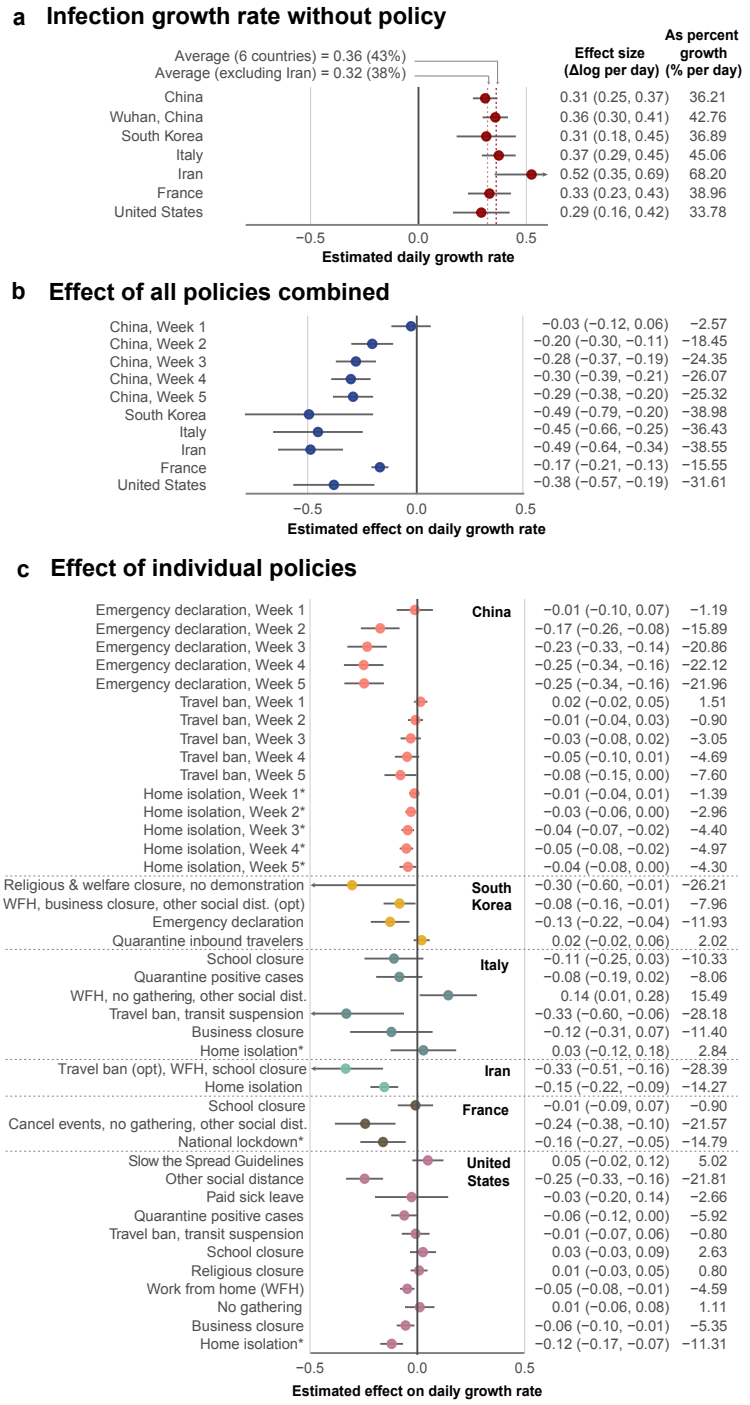

Figure 3.2: **Empirical estimates of unmitigated COVID-19 infection growth rates and the effect of anti-contagion policies.** Markers are country-specific estimates, whiskers are 95% CI. Columns report effect sizes as a change in the continuous-time growth rate (95% CI in parentheses) and the day-over-day percentage growth rate. (**a**) Estimates of daily COVID-19 infection growth rates in the absence of policy (dashed lines = averages with and without Iran, both excluding Wuhan-specific estimate). (**b**) Estimated combined effect of all policies on infection growth rates. (**c**) Estimated effects of individual policies or policy groups on the daily growth rate of infections, jointly estimated and ordered roughly chronologically within each country. *Reported effect of "home isolation" includes effects of other implied policies (see Methods).

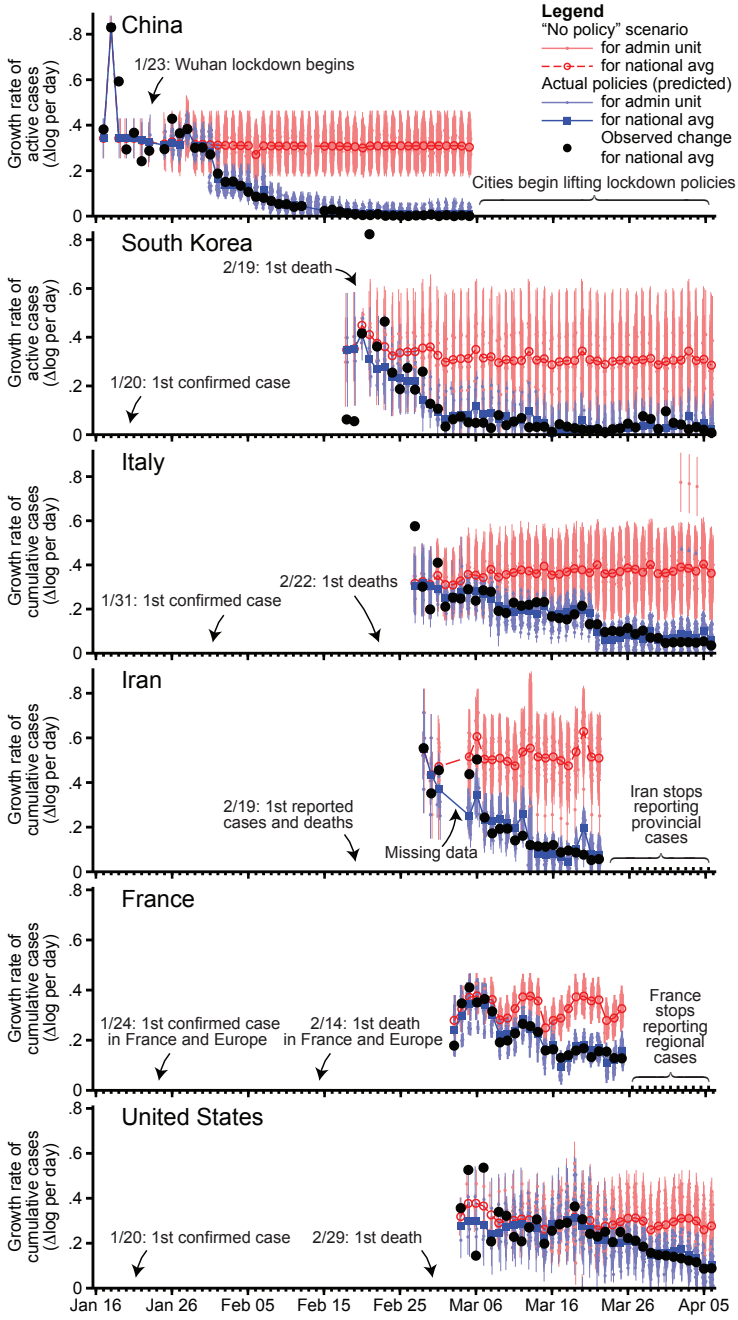

Figure 3.3: **Estimated infection growth rates based on actual anti-contagion policies and in a "no policy" counterfactual scenario.** Predicted daily growth rates of active (China, South Korea) or cumulative (all others) COVID-19 infections based on the observed timing of all policy deployments within each country (blue) and in a scenario where no policies were deployed (red). The difference between these two predictions is our estimated effect of actual anti-contagion policies on the growth rate of infections. Small markers are daily estimates for sub-national administrative units (vertical lines are 95% CI). Large markers are national averages. Black circles are observed daily changes in $\log(infections)$, averaged across administrative units. Sample sizes are the same as Figure 3.2.

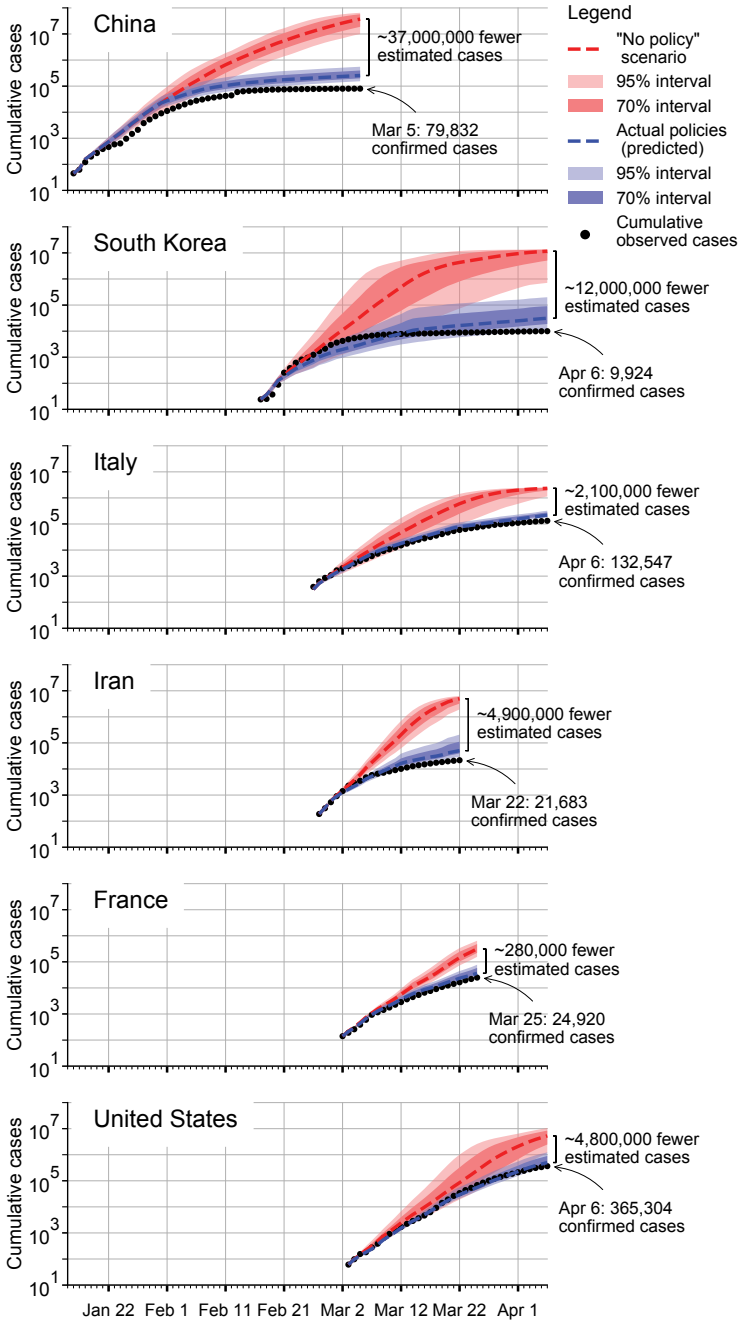

Figure 3.4: **Estimated cumulative confirmed COVID-19 infections with and without anti-contagion policies.** The predicted cumulative number of confirmed COVID-19 infections based on actual policy deployments (blue) and in the no-policy counterfactual scenario (red). Shaded areas show uncertainty based on 1,000 simulations where empirically estimated parameters are resampled from their joint distribution (dark = inner 70% of predictions; light = inner 95%). Black dotted line is observed cumulative infections. Infections are not projected for administrative units that never report infections in the sample, but which might have experienced infections in a no-policy scenario.

Chapter 4

Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics

4.1 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 and Shapiro, 2016), as have spe-

The materials in this chapter have been published as "Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics" in *Annual Review of Economics* (Bouguen et al., 2019). It was coauthored with Adrien Bouguen, Michael Kremer and Edward Miguel. The published version can be found online at https://www.annualreviews.org/doi/abs/10.1146/annurev-economics-080218-030333. The Supplementary Materials can be found at https://www.annualreviews.org/doi/suppl/10.1146/annurev-economics-080218-030333/suppl_file/EC11_Miguel_SupMat.pdf.

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 are also grateful to David Roodman of Open Philanthropy, who sparked GiveWell's support for the LID project with his 2017 blog post, "How thin the reed?" We also benefited from suggestions and comments provided by the LID faculty advisory committee at the Center for Effective Global Action, including Lia Fernald, Paul Gertler, Marco Gonzalez-Navarro, and Manisha Shah, as well as from Oriana Bandiera, Robin Burgess, Xavier Jaravel, and Rachael Meager.

cific 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 4.1 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 and McClafferty, 2001), the Kenya deworming study (Miguel and 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 4.2 presents the cumulative number of RCT studies in the cash transfer and child

[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.

health areas that 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 and 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.

[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.

Bleakley and 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 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 and 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 and Michelmore, 2018). Bastian and 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 and 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 4.2 summarizes and evaluates the growing body of evidence from RCTs on the long-term impacts of international development interventions, and finds that most (though

[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 and 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).

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.

Section 4.3 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 4.4 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 and 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 4.5 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.

4.2 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 Millán et al. (2019) 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. (2018) 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 4.1 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 4.1 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 4.1 are new unpublished working papers (at the time of writing this article).

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.

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

[5]We omit this study from Table 4.1 because it only reports medium-run outcomes at a time horizon of less than a decade.

produce meaningful increases in educational attainment and lead them to marry significantly more educated husbands, which may lead to long-run benefits.

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, 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

[6]We omit this study from Table 4.1 because it only reports medium-run outcomes at a time horizon of less than a decade.

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 and Shapiro, 2018) and the 19th century Georgia land lottery (Bleakley and 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.

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 4.4), 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 earn-

ings; 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 and 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 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 and Vogl (2018) use a difference-in-differences strategy and leverage both the spatio-temporal variation in pro-

[7]We omit PROGRESA/Oportunidades from Table 4.1 because the follow-up studies that are based on the original randomization only report medium-run outcomes.

[8]We omit these studies from Table 4.1 because they do not rely on RCT based estimators.

gram 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 and 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. (2018)'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]

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.

[9] We omit this study from Table 4.1 because it does not report labor market outcomes.

[10] We refer interested readers to the contemporaneous work by Millán et al. (2019), 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.

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 winners 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.

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.

Deworming

As discussed in Section 4.1, 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 and 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 and 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 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

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

from the Miguel and 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 and 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.

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

[12]We omit this study from Table 4.1 because it does not report labor market outcomes.

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.

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 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.

[13] We omit this study from Table 4.1 because it does not report labor market outcomes.

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.

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 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

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

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.

Baranov et al. (2020) 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.

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 4.1, 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"

[15]We omit this study from Table 4.1 because it does not report long-run labor market outcomes.

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.

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. (2018) 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.

4.3 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 4.1. 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]

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.

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),

[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.

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 4.2). 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 4.4 below, we discuss several approaches that could improve this yield rate for future experiments.

Eligible Studies

Table 4.2 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 4.2 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 4.4.

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 4.2, 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.

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.

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]

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,

[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.

nutrition, perinatal, sanitation, specific diseases, and stimulation). Studies in the nutrition literature, listed in Table 4.3, represent the largest group (32 studies), while the other categories, as shown in Appendix 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.

Eligible Nutrition Studies

Studies of two interventions clearly stand out in Table 4.3: 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., 1986b) 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 4.3). While a more recent and large-scale study has led to some questions regarding these magnitudes (Awasthi et al., 2013b), 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 4.3 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. (1986b), 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 4.3 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 4.3. 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).

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 4.3. However, many will become viable candidates in the coming years.

4.4 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.

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 and Todd, 2017),

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

the deworming program in Kenya (Miguel and 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 and 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 the amount of time group j has already been exposed to the intervention at time t. We specify the data-generating process to be

$$Y_{ijt} = \alpha_t + \sum_{k=0}^{K} \tau_k \mathbf{1}\{K_{jt} = kT\} + \epsilon_{ijt} \tag{4.1}$$

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 and 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.

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

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

$$Y_{ijt} = \alpha_t + \sum_{k=0}^{K} \hat{\tau}_k \mathbf{1}\{K_{jt} = kT\} + \epsilon_{ijt} \tag{4.2}$$

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

$$\text{Var}(\hat{\tau}_k) = \frac{2}{N}(k+1)\sigma^2 \tag{4.3}$$

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

[22]The derivation is shown in Appendix C.

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 and 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.

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 4.4, and Appendix A and B.

Data

Follow-up Surveys

Table 4.1 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 4.1, 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 4.1 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.

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 poor-

est 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. (2018). 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.

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 and 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. (2016b). 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.

4.5 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.

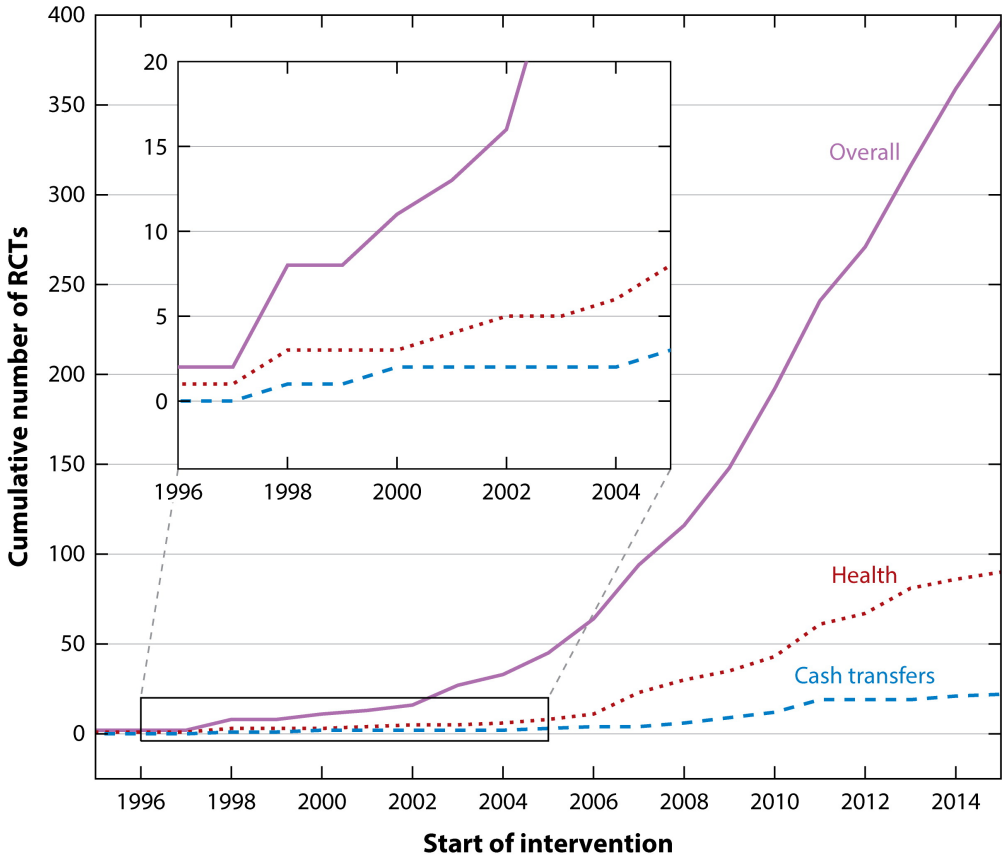

Figure 4.1: Cumulative number of completed randomized controlled trials (RCTs) in low- and middle-income countries from 1995 to 2015 in the American Economic Association's RCT Registry (`https://www.socialscienceregistry.org`).

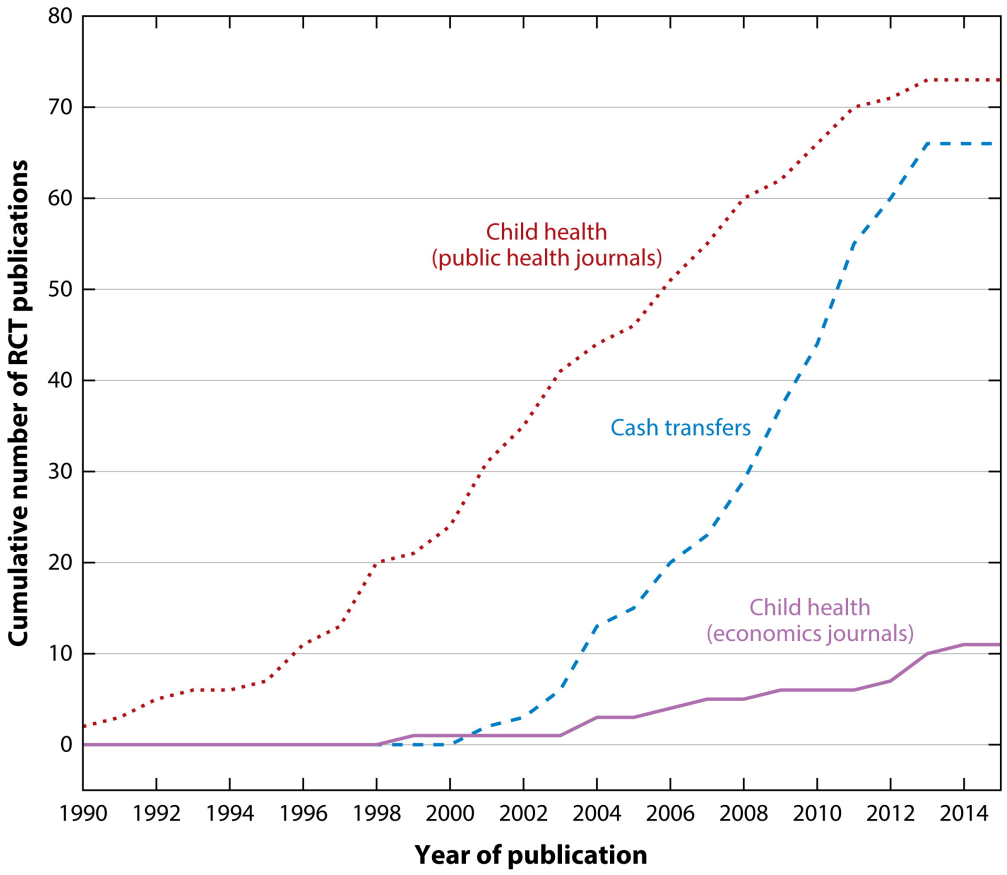

Figure 4.2: Cumulative number of randomized controlled trial (RCT) publications on cash transfers and child health in the AidGrade database (`http://www.aidgrade.org`).

Table 4.1: Existing Evidence on Long-Run Impacts of Development RCTs

	Study	Country	Intervention	Start of Intervention	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.

Table 4.1: Existing Evidence on Long-Run Impacts of Development RCTs (Continued)

	Study	Country	Intervention	Start of Intervention	Years to Follow-up	Data	Attrition Rate	Long-Run Impacts
8	PRAF II (Molina Millán et al., 2018)	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.
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.

Table 4.1: 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
14	Bangladesh Child Marriage (Buchmann et al., 2018)	Bangladesh	CCT	2008	9	Survey	13%	The conditional incentives delayed marriage and child-bearing, 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.

Table 4.2: Selected Cash Transfer Studies for Potential Long-term Follow-up

	Study Acronym	Country	Type	Start of Intervention	Phase-in Design[a]	Already Followed-Up (> 5 years)	ST Impacts			
							E	D	H	L
1	PROGRESA (Behrman et al., 2005)	Mexico	CCT	1998	yes	yes	+	+	+	0
2	PRAF II (Galiani and McEwan, 2013)	Honduras	UCT	2000	no	no		+		0
3	RPS (Maluccio and Flores, 2005)	Nicaragua	CCT	2000	yes	yes	+	+	+	−
4	BDH (Paxson and 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 and 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., 2015b)	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	+	+	+	+

ST Impacts: Short-Term Impacts; E: Economic; D: Education; H: Health; L: Adult Labor Market.
UCT: Unconditional Cash Transfer; CCT: Conditional Cash Transfer.
+/− indicate significant and positive/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.

Table 4.3: Selected Child Nutrition Studies for Potential Long-term Follow-up

	Study	Country	Description	Start of Intervention	Clustered RCT	Sample Size[a]	Age of Children	ST Impacts	
								H	C
1	Sommer et al. (1986a)	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. (2013a) (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	+	

Table 4.3: Selection of Child Nutrition Studies for Long-term Follow-up (Continued)

	Study	Country	Description	Start of Intervention	Clustered RCT	Sample Size[a]	Age of Children	ST Impacts H	ST Impacts C
26	Aboud and Akhter (2011)	Bangladesh	MMN & Responsive Feeding	2008	yes	45	8–20 mo	+	+
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	

ST Impacts: Short-Term Impacts; H: Health; C: Cognition.
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.

Bibliography

Frances E Aboud and Sadika Akhter. A cluster-randomized evaluation of a responsive stimulation and feeding intervention in bangladesh. *Pediatrics*, pages peds–2010, 2011.

Frances E Aboud, Sohana Shafique, and Sadika Akhter. A responsive feeding intervention increases children's self-feeding and maternal responsiveness but not weight gain. *The Journal of nutrition*, 139(9):1738–1743, 2009.

Daron Acemoglu and Simon Johnson. Disease and development: the effect of life expectancy on economic growth. *Journal of political Economy*, 115(6):925–985, 2007.

Achyuta Adhvaryu, Anant Nyshadham, Teresa Molina, and Jorge Tamayo. Helping children catch up: Early life shocks and the progresa experiment. Technical report, National Bureau of Economic Research, 2018.

Agence Régionale de Santé. Agir pour la santé de tous. `https://www.ars.sante.fr/`, 2020.

Emily L. Aiken, Guadalupe Bedoya, Aidan Coville, and Joshua E. Blumenstock. Targeting development aid with machine learning and mobile phone data: Evidence from an anti-poverty intervention in afghanistan. Unpublished, 2020.

Anna Aizer, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney. The long-run impact of cash transfers to poor families. *American Economic Review*, 106(4):935–71, 2016.

Richard Akresh, Damien de Walque, and Harounan Kazianga. *Evidence from a Randomized Evaluation of the Household Welfare Impacts of Conditional and Unconditional Cash Transfers Given to Mothers or Fathers*. The World Bank, 2016. doi: 10.1596/1813-9450-7730. URL `https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-7730`.

Jennifer Alix-Garcia, Craig McIntosh, Katharine RE Sims, and Jarrod R Welch. The ecological footprint of poverty alleviation: evidence from mexico's oportunidades program. *Review of Economics and Statistics*, 95(2):417–435, 2013.

Douglas Almond. 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(4):672–712, 2006.

Douglas Almond, Janet Currie, and Valentina Duque. Childhood circumstances and adult outcomes: Act ii. Technical report, National Bureau of Economic Research, 2017.

Michael L Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495, 2008.

Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

Alison Andrew, Orazio Attanasio, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 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(4):e1002556, 2018.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

M Caridad Araujo, Mariano Bosch, and Norbert Schady. Can cash transfers help households escape an inter-generational poverty trap? In *The Economics of Poverty Traps*. University of Chicago Press, 2017.

David Atkin, Benjamin Faber, Thibault Fally, and Marco Gonzalez-Navarro. A new engel on price index and welfare estimation. Technical report, National Bureau of Economic Research, 2020.

Orazio P Attanasio, Camila Fernández, Emla OA Fitzsimons, Sally M Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 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, 2014.

Shally Awasthi, Richard Peto, Simon Read, Sarah Clark, Vinod Pande, Donald Bundy, DEVTA (Deworming, Enhanced Vitamin A) team, et al. 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(9876):1469–1477, 2013a.

Shally Awasthi, Richard Peto, Simon Read, Sarah Clark, Vinod Pande, Donald Bundy, DEVTA (Deworming, Enhanced Vitamin A) team, et al. 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(9876):1469–1477, 2013b.

Boris Babenko, Jonathan Hersh, David Newhouse, Anusha Ramakrishnan, and Tom Swartz. Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in mexico. In *Proceedings of NIPS 2017 Workshop on Machine Learning for the Developing World*, 2017. URL `https://arxiv.org/abs/1711.06323`.

Sarah Baird, Joan Hamory, and Edward Miguel. Tracking, attrition and data quality in the kenyan life panel survey round 1 (klps-1). Technical report, Center for International and Development Economics Research Working Paper, 2008. URL `https://escholarship.org/uc/item/3cw7p1hx`.

Sarah Baird, Craig McIntosh, and Berk Özler. Cash or condition? evidence from a cash transfer experiment. *The Quarterly journal of economics*, 126(4):1709–1753, 2011.

Sarah Baird, Joan Hamory Hicks, Michael Kremer, and Edward Miguel. Worms at work: Long-run impacts of a child health investment. *The Quarterly Journal of Economics*, 131 (4):1637–1680, 2016a.

Sarah Baird, Craig McIntosh, and B Ozler. When the money runs out: Evaluating the longer-term impacts of a two year cash transfer program. Technical report, Working Paper, 2016b.

Sarah Baird, Joan Hamory Hicks, Michael Kremer, and Edward Miguel. Worms and well-being: 15 year economic impacts from kenya. Unpublished, 2018.

Oriana Bandiera, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. Labor markets and poverty in village economies. *The Quarterly Journal of Economics*, 132(2):811–870, 2017. doi: 10.1093/qje/qjx003. URL `http://dx.doi.org/10.1093/qje/qjx003`.

Oriana Bandiera, Clare Balboni, Robin Burgess, Maitreesh Ghatak, and Anton Heil. Why do people stay poor? Unpublished, 2018.

Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1):22–53, 2015a.

Abhijit Banerjee, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348 (6236):1260799, 2015b.

Abhijit Banerjee, Dean Karlan, and Jonathan Zinman. Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, 7(1):1–21, 2015c.

Abhijit Banerjee, Esther Duflo, Raghabendra Chattopadhyay, and Jeremy Shapiro. The long term impacts of a "graduation" program: Evidence from west bengal. Unpublished, 2016.

Abhijit V Banerjee and Esther Duflo. *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs, 2011.

Victoria Baranov, Sonia Bhalotra, Pietro Biroli, and Joanna Maselko. Maternal depression, women's empowerment, and parental investment: evidence from a randomized controlled trial. *American economic review*, 110(3):824–59, 2020.

Tania Barham, Karen Macours, and John A Maluccio. Are conditional cash transfers fulfilling their promise? schooling, learning, and earnings after 10 years. Technical report, CEPR Discussion Papers, 2017.

Tania Barham, Karen Macours, and JOHN A Maluccio. Experimental evidence of exposure to a conditional cash transfer during early teenage years: Young women's fertility and labor market outcomes. Technical report, CEPR discussion paper, 2018.

Felipe Barrera-Osorio, Marianne Bertrand, Leigh L Linden, and Francisco Perez-Calle. Improving the design of conditional transfer programs: Evidence from a randomized education experiment in colombia. *American Economic Journal: Applied Economics*, 3(2): 167–95, 2011.

Felipe Barrera-Osorio, Leigh L Linden, and Juan Saavedra. Medium-and long-term educational consequences of alternative conditional cash transfer designs: Experimental evidence from colombia. Technical report, National Bureau of Economic Research, 2017.

Mauricio Lima Barreto, GG Farenzena, RL Fiaccone, LMP Santos, Ana Marlucia de Oliveira Assis, MPN Araújo, and PAB Santos. Effect of vitamin a supplementation on diarrhoea and acute lower-respiratory-tract infections in young children in brazil. *The Lancet*, 344 (8917):228–231, 1994.

Jacob Bastian and Katherine Michelmore. The long-term impact of the earned income tax credit on children's education and employment outcomes. *Journal of Labor Economics*, 36(4):1127–1163, 2018.

BBC News. Wuhan pneumonia: 30 days from outbreak to out of control. `https://www.bbc.com/zhongwen/simp/chinese-news-51290945`, 2020.

Jere R Behrman, Piyali Sengupta, and Petra Todd. Progressing through progresa: An impact assessment of a school subsidy experiment in rural mexico. *Economic development and cultural change*, 54(1):237–275, 2005.

Jere R Behrman, Maria C Calderon, Samuel H Preston, John Hoddinott, Reynaldo Martorell, and Aryeh D Stein. Nutritional supplementation in girls influences the growth of their children: prospective study in guatemala. *The American Journal of Clinical Nutrition*, 90 (5):1372–1379, 2009.

Jere R Behrman, Susan W Parker, and Petra E Todd. Do conditional cash transfers for schooling generate lasting benefits? a five-year followup of progresa/oportunidades. *Journal of Human Resources*, 46(1):93–122, 2011.

Fiorella Benedetti, Pablo Ibarrarán, and Patrick J McEwan. Do education and health conditions matter in a large cash transfer? evidence from a honduran experiment. *Economic Development and Cultural Change*, 64(4):759–793, 2016.

Najy Benhassine, Florencia Devoto, Esther Duflo, Pascaline Dupas, and Victor Pouliquen. Turning a shove into a nudge? a "labeled cash transfer" for education. *American Economic Journal: Economic Policy*, 7(3):86–125, 2015a.

Najy Benhassine, Florencia Devoto, Esther Duflo, Pascaline Dupas, and Victor Pouliquen. Turning a shove into a nudge? a" labeled cash transfer" for education. *American Economic Journal: Economic Policy*, 7(3):86–125, 2015b.

Eric Bettinger, Michael Kremer, Maurice Kugler, Carlos Medina, Christian Posso, and Juan Saavedra. School vouchers, labor markets and vocational education. Unpublished, 2018.

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

Christopher Blattman, Nathan Fiala, and Sebastian Martinez. Generating skilled self-employment in developing countries: Experimental evidence from uganda. *The Quarterly Journal of Economics*, 129(2):697–752, 2013.

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

Hoyt Bleakley. Disease and development: evidence from hookworm eradication in the american south. *The Quarterly Journal of Economics*, 122(1):73–117, 2007.

Hoyt Bleakley and Joseph P Ferrie. Up from poverty? the 1832 cherokee land lottery and the long-run distribution of wealth. Technical report, National Bureau of Economic Research, 2013.

Joshua Blumenstock. Don't forget people in the use of big data for development. *Nature*, 561(7722):170–172, 2018.

Joshua Blumenstock. Machine learning can help get covid-19 aid to those who need it most. *Nature*, 2020.

Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

Joshua Blumenstock, Michael Callen, Tarek Ghani, Niall Keleher, and Jacob Shapiro. Measuring poverty and vulnerability in real-time. Unpublished, 2018.

Joshua Evan Blumenstock. Fighting poverty with data. *Science*, 353(6301):753–754, 2016.

Martin C. J. Bootsma and Neil M. Ferguson. The effect of public health measures on the 1918 influenza pandemic in U.S. cities. *Proceedings of the National Academy of Sciences*, 104(18):7588–7593, 5 2007. ISSN 0027-8424. doi: 10.1073/pnas.0611071104.

Kirill Borusyak and Xavier Jaravel. Revisiting event study designs. Unpublished, 2016.

Adrien Bouguen, Yue Huang, Michael Kremer, and Edward Miguel. Using randomized controlled trials to estimate long-run impacts in development economics. *Annual Review of Economics*, 2019.

Lasse Brune, Dean Karlan, Sikandra Kurdi, and Christopher R Udry. Social protection amidst social upheaval: Examining the impact of a multi-faceted program for ultra-poor households in yemen. Technical report, National Bureau of Economic Research, 2020.

Giorgio Brunello and Simona Comi. Education and earnings growth: evidence from 11 european countries. *Economics of Education Review*, 23(1):75–83, 2004.

Nina Buchmann, Erica Field, Rachel Glennerster, Shahana Nazneen, Svetlana Pimkina, and Iman Sen. Power vs money: Alternative approaches to reducing child marriage in bangladesh, a randomized control trial. Unpublished, 2018.

Marshall Burke and David B Lobell. Satellite-based assessment of yield variation and its determinants in smallholder african systems. *Proceedings of the National Academy of Sciences*, 114(9):2189–2194, 2017.

Marshall Burke, Solomon M Hsiang, and Edward Miguel. Global non-linear effect of temperature on economic production. *Nature*, 527(7577):235–239, 2015.

Fiona Burlig and Matt Woerman. Are 212 section 10: Non-standard standard errors ii. `https://static1.squarespace.com/static/558eff8ce4b023b6b855320a/t/573bd63745bf21da74c080a8/1463539276997/ARE_212_Section_10.pdf`, 2016. accessed 10 May 2020.

Bettye M Caldwell, Robert H Bradley, et al. *Home observation for measurement of the environment*. University of Arkansas at Little Rock Little Rock, 1984.

Nathalie Charpak, Rejean Tessier, Juan G Ruiz, Jose Tiberio Hernandez, Felipe Uriza, Julieta Villegas, Line Nadeau, Catherine Mercier, Francoise Maheu, Jorge Marin, et al. Twenty-year follow-up of kangaroo mother care versus traditional care. *Pediatrics*, page e20162063, 2016.

Xi Chen and William D Nordhaus. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594, 2011.

Raj Chetty, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics*, 126(4):1593–1660, 2011a.

Raj Chetty, John N Friedman, and Jonah E Rockoff. New evidence on the long-term impacts of tax credits. *IRS Statistics of Income White Paper*, 2011b.

Matteo Chinazzi, Jessica T. Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, Cécile Viboud, Xinyue Xiong, Hongjie Yu, M. Elizabeth Halloran, Ira M. Longini, and Alessandro Vespignani. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 2020. ISSN 0036-8075. doi: 10.1126/science.aba9757.

Gerardo Chowell, Lisa Sattenspiel, Shweta Bansal, and Cécile Viboud. Mathematical models to characterize early epidemic growth: A review. *Physics of Life Reviews*, 18:66–97, 2016.

Civil Protection Department Website - Presidency of the Council of Ministers. Coronavirus emergency. `http://www.protezionecivile.it/web/guest/home`, 2020.

COCO. Coco - common objects in contexts. http://cocodataset.org, 2020. accessed 6 May 2020.

T Conley. Spatial econometrics. new palgrave dictionary of economics, eds durlauf sn, blume le, 2008.

Timothy G Conley. Gmm estimation with cross sectional dependence. *Journal of econometrics*, 92(1):1–45, 1999.

Kevin Croke, Joan Hamory Hicks, Eric Hsu, Michael Kremer, and Edward Miguel. Does mass deworming affect child nutrition? meta-analysis, cost-effectiveness, and statistical power. Working Paper 22382, National Bureau of Economic Research, July 2016. URL `http://www.nber.org/papers/w22382`.

Jesse M Cunha. Testing paternalism: Cash versus in-kind transfers. *American Economic Journal: Applied Economics*, 6(2):195–230, 2014.

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

Angus Deaton. *The analysis of household surveys: a microeconometric approach to development policy*. The World Bank, 1997.

David Deming. Early childhood intervention and life-cycle skill development: Evidence from head start. *American Economic Journal: Applied Economics*, 1(3):111–34, 2009.

Michael J Dibley, Tonny Sadjimin, Chris L Kjolhede, and Lawrence H Moulton. Vitamin a supplementation fails to reduce incidence of acute respiratory illness and diarrhea in preschool-age indonesian children. *The Journal of nutrition*, 126(2):434–442, 1996.

Esther Duflo, Pascaline Dupas, and Michael Kremer. The impact of free secondary education: Experimental evidence from ghana. Unpublished, 2018.

Chloe N East. The effect of food stamps on children's health: Evidence from immigrants' changing eligibility. *Journal of Human Resources*, pages 0916–8197R2, 2018.

EGAP. Metaketa initiative, 2018. URL `https://egap.org/metaketa`.

Dennis Egger, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael W Walker. General equilibrium effects of cash transfers: experimental evidence from kenya. Technical report, National Bureau of Economic Research, 2019.

Ans Eilander, Tarun Gera, Harshpal S Sachdev, Catherine Transler, Henk CM van der Knaap, Frans J Kok, and Saskia JM Osendarp. Multiple micronutrient supplementation for improving cognitive performance in children: systematic review of randomized controlled trials–. *The American journal of clinical nutrition*, 91(1):115–130, 2009.

Chris Elbers, Jean O Lanjouw, and Peter Lanjouw. Micro–level estimation of poverty and inequality. *Econometrica*, 71(1):355–364, 2003.

Christopher D Elvidge, Kimberly E Baugh, Mikhail Zhizhin, and Feng-Chi Hsu. Why viirs data are superior to dmsp for mapping nighttime lights. *Proceedings of the Asia-Pacific Advanced Network*, 35(62), 2013.

Christopher D Elvidge, Kimberly Baugh, Mikhail Zhizhin, Feng Chi Hsu, and Tilottama Ghosh. Viirs night-time lights. *International Journal of Remote Sensing*, 38(21):5860–5879, 2017.

Ryan Engstrom, Jonathan Hersh, and David Newhouse. Poverty from space: using high-resolution satellite imagery for estimating economic well-being. Technical report, The World Bank, 2017.

David Evans, Stephanie Hausladen, Katrina Kosec, and Natasha Reese. *Community-based conditional cash transfers in Tanzania: results from a randomized trial*. The World Bank, 2014.

Mieke Faber, Jane D Kvalsvig, Carl J Lombard, and AJ Spinnler Benadé. Effect of a fortified maize-meal porridge on anemia, micronutrient status, and motor development of infants–. *The American journal of clinical nutrition*, 82(5):1032–1039, 2005.

Facebook. How ai-powered maps help improve vaccination campaigns and rural electrification. https://tech.fb.com/ai-powered-maps-help-vaccination-campaigns/, 2019. accessed 3 March 2021.

Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Technical report, Imperial College London, 3 2020.

David Fisman, Edwin Khoo, and Ashleigh Tuite. Early epidemic dynamics of the west african 2014 ebola outbreak: estimates derived with a simple two-parameter model. *PLoS currents*, 6, 2014.

Seth Flaxman, Swapnil Mishra, Axel Gandy, H Unwin, H Coupland, T Mellan, H Zhu, T Berah, J Eaton, P Perez Guzman, et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Technical report, Imperial College London, 2020.

Sebastian Galiani and Patrick J McEwan. The heterogeneous impact of conditional cash transfers. *Journal of Public Economics*, 103:85–96, 2013.

Nicola Gennaioli, Rafael La Porta, Florencio Lopez-de Silanes, and Andrei Shleifer. Human capital and regional development. *The Quarterly journal of economics*, 128(1):105–164, 2012.

Paul Gertler, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M Chang, and Sally Grantham-McGregor. Labor market returns to an early childhood stimulation intervention in jamaica. *Science*, 344(6187):998–1001, 2014.

Google Earth Engine. Viirs stray light corrected nighttime day/night band composites version 1. https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_MONTHLY_V1_VCMSLCFG, 2020. accessed 6 May 2020.

Google Static Maps. Google static maps. https://developers.google.com/maps/documentation/maps-static/intro, 2020. accessed 6 May 2020.

Eric P Green, Christopher Blattman, Julian Jamison, and Jeannie Annan. 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, 2015.

William H Greene. *Econometric Analysis*. Prentice Hall, 2003. Upper Saddle River, NJ.

SUMMIT Study Group. Effect of maternal multiple micronutrient supplementation on fetal loss and infant death in indonesia: a double-blind cluster-randomised trial. *The Lancet*, 371(9608):215–227, 2008.

Richard J. Hatchett, Carter E. Mecher, and Marc Lipsitch. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proceedings of the National Academy of Sciences*, 104(18):7582–7587, May 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0610941104.

Johannes Haushofer and Jeremy Shapiro. The short-term impact of unconditional cash transfers to the poor: experimental evidence from kenya. *The Quarterly Journal of Economics*, 131(4):1973–2042, 2016.

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

James J Heckman. Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902, 2006.

J Vernon Henderson, Adam Storeygard, and David N Weil. Measuring economic growth from outer space. *American economic review*, 102(2):994–1028, 2012.

M Guillermo Herrera, Penelope Nestel, L Weld, A El Amin, KAMAL AHMED Mohamed, and WW Fawzi. Vitamin a supplementation and child survival. *The Lancet*, 340(8814): 267–271, 1992.

Sonja Y Hess, Janet M Peerson, Elodie Becquey, Souheila Abbeddou, Césaire T Ouédraogo, Jérôme W Somé, Elizabeth Yakes Jimenez, Jean-Bosco Ouédraogo, Stephen A Vosti, Noël Rouamba, et al. 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(8):e0181770, 2017.

John Hoddinott, John A Maluccio, Jere R Behrman, Rafael Flores, and Reynaldo Martorell. Effect of a nutrition intervention during early childhood on economic productivity in guatemalan adults. *The lancet*, 371(9610):411–416, 2008.

Hilary Hoynes, Diane Whitmore Schanzenbach, and Douglas Almond. Long-run impacts of childhood access to the safety net. *American Economic Review*, 106(4):903–34, 2016.

Hilary W Hoynes and Diane Whitmore Schanzenbach. Safety net investments in children. Technical report, National Bureau of Economic Research, 2018.

Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Luna Yue Huang, Andrew Hultgren, Emma Krasovich, et al. The effect of large-scale anti-contagion policies on the covid-19 pandemic. *Nature*, 584(7820):262–267, 2020.

Solomon M Hsiang. Temperatures and cyclones strongly associated with economic production in the caribbean and central america. *Proceedings of the National Academy of sciences*, 107(35):15367–15372, 2010.

Chang-Tai Hsieh, Edward Miguel, Daniel Ortega, and Francisco Rodriguez. The price of political opposition: Evidence from venezuela's maisanta. *American Economic Journal: Applied Economics*, 3(2):196–214, 2011.

INEGI. 2010 population and housing census of mexico. https://www.inegi.org.mx/programas/ccpv/2010/default.html, 2010. accessed 5 May 2020.

Institute for Health Metrics and Evaluation. Global health data exchange, 2018. URL `http://ghdx.healthdata.org/gbd-results-tool`.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016a.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016b.

Champak C Jinabhai, Myra Taylor, Anna Coutsoudis, Hoosen M Coovadia, Andrew M Tomkins, and Keith R Sullivan. A randomized controlled trial of the effect of anti-helminthic treatment and micronutrient fortification on health status and school performance of rural primary school children. *Annals of tropical paediatrics*, 21(4):319–333, 2001.

Cigdem Kagitcibasi, Diane Sunar, Sevda Bekman, Nazli Baydar, and Zeynep Cemalcilar. Continuing effects of early enrichment in adult life: The turkish early enrichment project 22 years later. *Journal of Applied Developmental Psychology*, 30(6):764–779, 2009.

Sasikiran Kandula, Teresa Yamana, Sen Pei, Wan Yang, Haruka Morita, and Jeffrey Shaman. Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. *Journal of The Royal Society Interface*, 15(144):20180174, 2018.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical*
and physical character, 115(772):700–721, 1927.

Hans-Peter Kohler and Rebecca L Thornton. Conditional cash transfers and hiv/aids prevention: unconditionally promising? *The World Bank Economic Review*, 26(2):165–190, 2011.

Timothy A Kohler, Michael E Smith, Amy Bogaard, Gary M Feinman, Christian E Peterson, Alleen Betzenhauser, Matthew Pailes, Elizabeth C Stone, Anna Marie Prentiss, Timothy J Dennehy, et al. Greater post-neolithic wealth disparities in eurasia than in north america and mesoamerica. *Nature*, 551(7682):619–622, 2017.

Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

Michael Kremer. Randomized evaluations of educational programs in developing countries: Some lessons. *American Economic Review*, 93(2):102–106, 2003.

Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 2020.

Adriana D Kugler and Ingrid Rojas. Do ccts improve employment and earnings in the very long-term? evidence from mexico. Technical report, National Bureau of Economic Research, 2018.

Peter Lanjouw, Nicholas Stern, et al. *How Lives Change: Palanpur, India, and Development Economics*. Oxford University Press, 2018.

Kenneth Lee, Edward Miguel, and Catherine Wolfram. Experimental evidence on the economics of rural electrification. *Journal of Political Economy*, 128(4):1523–1565, 2020.

Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020a.

Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*, 2020b.

Jiabao Lin. COVID-19/2019-nCoV Time Series Infection Data Warehouse. `https://github.com/BlankerL/DXY-COVID-19-Data`, 2020.

Torbjörn Lind, Bo Lönnerdal, Hans Stenlund, Indria L Gamayanti, Djauhar Ismail, Rosadi Seswandhana, and Lars-Åke Persson. 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(3):729–736, 2004.

David Lindenbaum. 2nd spacenet competition winners code release. `https://medium.com/the-downlinq/2nd-spacenet-competition-winners-code-release-c7473eea7c11`, 2017. accessed 29 Jan 2021.

Kurt Z Long, Yura Montoya, Ellen Hertzmark, Jose I Santos, and Jorge L Rosado. 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(3):693–700, 2006.

José Lourenço, Robert Paton, Mahan Ghafari, Moritz Kraemer, Craig Thompson, Peter Simmonds, Paul Klenerman, and Sunetra Gupta. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. *medRxiv*, 2020.

Jens Ludwig and Douglas L Miller. Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly journal of economics*, 122(1):159–208, 2007.

Junling Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 2020.

Karen Macours, Norbert Schady, and Renos Vakis. Cash transfers, behavioral changes, and cognitive development in early childhood: Evidence from a randomized experiment. *American Economic Journal: Applied Economics*, 4(2):247–73, April 2012. doi: 10.1257/app.4.2.247. URL http://www.aeaweb.org/articles?id=10.1257/app.4.2.247.

Benjamin F Maier and Dirk Brockmann. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, 2020.

Kenneth M Maleta, John Phuka, Lotta Alho, Yin Bun Cheung, Kathryn G Dewey, Ulla Ashorn, Nozgechi Phiri, Thokozani E Phiri, Stephen A Vosti, Mamane Zeilani, et al. 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(8):1909–1915, 2015.

John Maluccio and Rafael Flores. *Impact evaluation of a conditional cash transfer program: The Nicaraguan Red de Protección Social*. Intl Food Policy Res Inst, 2005.

John A Maluccio, John Hoddinott, Jere R Behrman, Reynaldo Martorell, Agnes R Quisumbing, and Aryeh D Stein. The impact of improving nutrition during early childhood on education among guatemalan adults. *The Economic Journal*, 119(537):734–763, 2009.

Mari Skar Manger, Joanne E McKenzie, Pattanee Winichagoon, Andrew Gray, Visith Chavasit, Tippawan Pongcharoen, Sueppong Gowachirapant, Bruce Ryan, Emorn Wasantwisut, and Rosalind S Gibson. 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(6):1715–1722, 2008.

Richard M Martin, Michael S Kramer, Rita Patel, Sheryl L Rifas-Shiman, Jennifer Thompson, Seungmi Yang, Konstantin Vilchuck, Natalia Bogdanovich, Mikhail Hameza, Kate

Tilling, et al. 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(7):e170698–e170698, 2017.

Reynaldo Martorell, Jean-Pierre Habicht, and Juan A Rivera. History and design of the incap longitudinal study (1969–77) and its follow-up (1988–89). *The Journal of nutrition*, 125(suppl_4):1027S–1041S, 1995.

Benjamin Marx, Thomas M Stoker, and Tavneet Suri. There is no free house: Ethnic patronage in a kenyan slum. *American Economic Journal: Applied Economics*, 11(4): 36–70, 2019.

Susana L Matias, Alejandro Vargas-Vásquez, Ricardo Bado Pérez, Lorena Alcázar Valdivia, Oscar Aquino Vivanco, Amelia Rodriguez Martín, and Jose Pedro Novalbos Ruiz. Effects of lipid-based nutrient supplements v. micronutrient powders on nutritional and developmental outcomes among peruvian infants. *Public health nutrition*, 20(16):2998–3007, 2017.

Maxar. Eliminating malaria with the power of the crowd. https://blog.maxar.com/earth-intelligence/2017/eliminating-malaria-with-the-power-of-the-crowd, 2017. accessed 25 Feb 2021.

Sarmila Mazumder, Sunita Taneja, Kiran Bhatia, Sachiyo Yoshida, Jasmine Kaur, Brinda Dube, GS Toteja, Rajiv Bahl, Olivier Fontaine, Jose Martines, et al. 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(9975): 1333–1342, 2015.

David McKenzie. Beyond baseline and follow-up: The case for more t in experiments. *Journal of development Economics*, 99(2):210–221, 2012.

Gideon Meyerowitz-Katz and Lea Merone. A systematic review and meta-analysis of published research data on covid-19 infection-fatality rates. *medRxiv*, 2020. URL `https://www.medrxiv.org/content/10.1101/2020.05.03.20089854v1`.

Guy Michaels, Dzhamilya Nigmatulina, Ferdinand Rauch, Tanner Regan, Neeraj Baruah, and Amanda Dahlstrand-Rudin. Planning ahead for better neighborhoods: Long run evidence from tanzania. Technical report, IZA Institute of Labor Economics, 2017.

Stelios Michalopoulos and Elias Papaioannou. National institutions and subnational development in africa. *The Quarterly journal of economics*, 129(1):151–213, 2014.

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

Christina E Mills, James M Robins, and Marc Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906, 2004.

Stephen Mills, Stephanie Weiss, and Calvin Liang. Viirs day/night band (dnb) stray light characterization and correction. In *Earth Observing Systems XVIII*, volume 8866, page 88661P. International Society for Optics and Photonics, 2013.

Teresa Molina Millán, Tania Barham, Karen Macours, John A Maluccio, and Marco Stampini. Long-Term Impacts of Conditional Cash Transfers: Review of the Evidence. *The World Bank Research Observer*, 34(1):119–159, 05 2019. ISSN 0257-3032. doi: 10.1093/wbro/lky005. URL https://doi.org/10.1093/wbro/lky005.

Ministère des Solidarités et de la Santé. PRÉPARATION AU RISQUE ÉPIDÉMIQUE COVID-19. https://solidarites-sante.gouv.fr/IMG/pdf/guide_methodologique_covid-19-2.pdf, 2020.

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

Sendhil Mullainathan. Satellite images can pinpoint poverty where surveys can't. https://www.nytimes.com/2016/04/03/upshot/satellite-images-can-pinpoint-poverty-where-surveys-cant.html, 2016. accessed 25 Feb 2021.

K Muniz-Rodriguez, G Chowell, CH Cheung, D Jia, PY Lai, Y Lee, M Liu, SK Ofori, KM Roosa, L Simonsen, et al. Doubling time of the COVID-19 epidemic by province, China. *Emerging Infectious Diseases*, 26(8), 2020.

Luisa Natali, Sudhanshu Handa, Amber Peterman, David Seidenfeld, Gelson Tembo, et al. Making money work: unconditional cash transfers allow women to save and re-invest in rural zambia. Technical report, UNICEF Office of Research - Innocenti, Florence, 2016.

NEMO Study Group. 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(4):1082–1093, 2007.

Hiroshi Nishiura, Gerardo Chowell, Muntaser Safan, and Carlos Castillo-Chavez. Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza a (H1N1) 2009. *Theoretical Biology and Medical Modelling*, 7(1):1, 2010.

OECD. Household final expenditure on housing. In *National Accounts at a Glance 2014*. OECD Publishing, Paris, 2014.

Open AI Tanzania. 2018 open ai tanzania building footprint segmentation challenge. https://competitions.codalab.org/competitions/20100, 2020. accessed 6 May 2020.

Owen Ozier. Exploiting externalities to estimate the long-term effects of early childhood deworming. *American Economic Journal: Applied Economics*, 10(3):235–62, 2018.

Stéphanie Pamies-Sumner. Development impact evaluations: State of play and new challenges. Technical report, Agence Française de Développement, 2015.

Susan W Parker and Petra E Todd. Conditional cash transfers: The case of progresa/oportunidades. *Journal of Economic Literature*, 55(3):866–915, 2017.

Susan W Parker and Tom Vogl. Do conditional cash transfers improve economic outcomes in the next generation? evidence from mexico. Technical report, National Bureau of Economic Research, 2018.

Christina Paxson and Norbert Schady. Does Money Matter? The Effects of Cash Transfers on Child Development in Rural Ecuador. *Economic Development and Cultural Change*, 59(1):187–229, October 2010. URL `https://ideas.repec.org/a/ucp/ecdecc/v59y2010i1p187-229.html`.

Mark M Pitt, Mark R Rosenzweig, and Mohammad Nazmul Hassan. Human capital investment and the gender division of labor in a brawn-based economy. *American Economic Review*, 102(7):3531–60, 2012.

Elizabeth L Prado, Susy K Sebayang, Mandri Apriatni, Siti R Adawiyah, Nina Hidayati, Ayuniarti Islamiyah, Sudirman Siddiq, Benyamin Harefa, Jarrad Lum, Katherine J Alcock, et al. 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(2):e217–e228, 2017.

Presidenza del Consiglio dei Ministri. COVID-19. `https://github.com/pcm-dpc/COVID-19`, 2020.

David J Price and Jae Song. The long-term effects of cash assistance. Technical report, Working Paper, Stanford University, 2016.

Mohammad M Rahman, Sten H Vermund, Mohammad A Wahed, George J Fuchs, Abdullah H Baqui, and Jose O Alvarez. Simultaneous zinc and vitamin a supplementation in bangladeshi children: randomised double blind controlled trial. *Bmj*, 323(7308):314–318, 2001.

Laxmi Rahmathullah, Barbara A Underwood, Ravilla D Thulasiraj, Roy C Milton, Kala Ramaswamy, Raheem Rahmathullah, and Ganeesh Babu. Reduced mortality among children in southern india receiving a small weekly dose of vitamin a. *New England journal of medicine*, 323(14):929–935, 1990.

Christina D Romer and David H Romer. The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks. *American Economic Review*, 100(3): 763–801, 2010.

D.A Ross, N Dollimore, P.G Smith, B.R Kirkwood, P Arthur, S.S Morris, H.A Addy, FredN Binka, P Arthur, J.O Gyapong, and A.M Tomkins. Vitamin a supplementation in northern ghana: effects on clinic attendances, hospital admissions, and child mortality. *The Lancet*, 342(8862):7 – 12, 1993. ISSN 0140-6736. doi: https://doi.org/10.1016/0140-6736(93)91879-Q. URL http://www.sciencedirect.com/science/article/pii/0140673693918790. Originally published as Volume 2, Issue 8862.

Olivier Roussel. Open platform for french public data - Fr-SARS-CoV-2. https://www.data.gouv.fr/en/datasets/fr-sars-cov-2, 2020.

Timothy W Russell, Joel Hellewell, Sam Abbott, Nick Golding, Hamish Gibbs, Christopher I Jarvis, Kevin van Zandvoort, Stefan Flasche, Rosalind Eggo, Edmunds John W, Kucharski Adam J, and CMMID nCov working group. Using a delay-adjusted case fatality ratio to estimate under-reporting. Technical report, Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, 2020. Accessed: 2020-04-09.

Sante Publique France. Covid-19. https://www.santepubliquefrance.fr/, 2020.

Sunil Sazawal, Usha Dhingra, Pratibha Dhingra, Girish Hiremath, Archana Sarkar, Arup Dutta, Venugopal P Menon, and Robert E Black. Micronutrient fortified milk improves iron status, anemia and growth among children 1–4 years: a double masked, randomized, controlled trial. *PloS one*, 5(8):e12167, 2010.

Fernando Sempértegui, Bertha Estrella, Verónica Camaniero, Valeria Betancourt, Ricardo Izurieta, Wilma Ortiz, Elizabeth Fiallo, Sheyla Troya, Alicia Rodrıguez, and Jeffrey K Griffiths. The beneficial effects of weekly low-dose vitamin a supplementation on acute lower respiratory infections and diarrhea in ecuadorian children. *Pediatrics*, 104(1):e1–e1, 1999.

Anuraj H Shankar, Blaise Genton, Richard D Semba, Moses Baisor, Joseph Paino, Steven Tamja, Thomas Adiguma, Lee Wu, Lawrence Rare, James M Tielsch, et al. Effect of vitamin a supplementation on morbidity due to plasmodium falciparum in young children in papua new guinea: a randomised trial. *The Lancet*, 354(9174):203–209, 1999.

Bhattiprolu Sivakumar, Kamasundaram Vijayaraghavan, Shahnaz Vazir, Nagalla Balakrishna, Veena Shatrugna, Kramadhati Venkata Rameshwar Sarma, Krishnapillai Madhavan Nair, Namala Raghuramulu, and Kamala Krishnaswamy. Effect of micronutrient supplement on health and nutritional status of schoolchildren: study design. *Nutrition*, 22(1): S1–S7, 2006.

Emmanuel Skoufias and Bonnie McClafferty. Is progresa working: Summary of the results of an evaluation by ifpri. Technical report, International Food Policy Research Institute Washington, DC, 2001.

Florentino S Solon, Jesus N Sarol Jr, Allan BI Bernardo, Juan Antonio A Solon, Haile Mehansho, Liza E Sanchez-Fermin, Lorena S Wambangco, and Kenton D Juhlin. 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(4_suppl2):S129–S140, 2003.

Alfred Sommer. Vitamin a deficiency and clinical disease: an historical overview. *The Journal of nutrition*, 138(10):1835–1839, 2008.

Alfred Sommer, Edi Djunaedi, AA Loeden, Ignatius Tarwotjo, KeithP West JR, Robert Tilden, Lisa Mele, Aceh Study Group, et al. Impact of vitamin a supplementation on childhood mortality: a randomised controlled community trial. *The Lancet*, 327(8491): 1169–1173, 1986a.

Alfred Sommer, Edi Djunaedi, AA Loeden, Ignatius Tarwotjo, KeithP West JR, Robert Tilden, Lisa Mele, Aceh Study Group, et al. Impact of vitamin a supplementation on childhood mortality: a randomised controlled community trial. *The Lancet*, 327(8491): 1169–1173, 1986b.

Sally K Stansfield, Muller Pierre-Louis, A Augustin, and G Lerebours. Vitamin a supplementation and increased prevalence of childhood diarrhoea and acute respiratory infections. *The Lancet*, 342(8871):578–582, 1993.

John Strauss, Firman Witoelar, and Bondan Sikoki. *The fifth wave of the Indonesia family life survey: overview and field report*, volume 1. RAND Santa Monica, CA, 2016.

Parminder S Suchdev, Laird J Ruth, Bradley A Woodruff, Charles Mbakaya, Usha Mandava, Rafael Flores-Ayala, Maria Elena D Jefferds, and Robert Quick. 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(5):1223–1230, 2012.

Biao Tang, Xia Wang, Qian Li, Nicola Luigi Bragazzi, Sanyi Tang, Yanni Xiao, and Jianhong Wu. Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *Journal of Clinical Medicine*, 9(2):462, 2020.

Jeffery C Tanner, Tara Candland, and Whitney S Odden. Later impacts of early childhood interventions: a systematic review. *Washington: Independent Evaluation Group, World Bank Group*, 2015.

Alessandro Tarozzi and Angus Deaton. Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics*, 91(4):773–792, 2009.

the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. `https://github.com/CSSEGISandData/COVID-19`, 2020.

Think Global Health. Timeline of the coronavirus. `https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus`, 2020.

Duncan Thomas, Elizabeth Frankenberg, and James P Smith. Lost but not forgotten: Attrition and follow-up in the indonesia family life survey. *Journal of Human resources*, pages 556–592, 2001.

Duncan Thomas, Firman Witoelar, Elizabeth Frankenberg, Bondan Sikoki, John Strauss, Cecep Sumantri, and Wayan Suriastini. Cutting the costs of attrition: Results from the indonesia family life survey. *Journal of Development Economics*, 98(1):108–123, 2012.

Huaiyu Tian, Yonghong Liu, Yidan Li, Chieh-Hsi Wu, Bin Chen, Moritz UG Kraemer, Bingying Li, Jun Cai, Bo Xu, Qiqi Yang, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science*, 2020.

Tim K Tsang, Peng Wu, Yun Lin Yun Lin, Eric Lau, Gabriel M Leung, and Benjamin J Cowling. Impact of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China. *medRxiv*, 2020.

UN. Big data for sustainable development. https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/, 2021. accessed 25 Feb 2021.

Unicef et al. Levels and trends in child malnutrition. Technical report, eSocialSciences, 2018.

USA Facts. Coronavirus locations: COVID-19 map by county and state. `https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/`, 2020.

Jacobien Veenemans, Paul Milligan, Andrew M Prentice, Laura RA Schouten, Nienke Inja, Aafke C van der Heijden, Linsey CC de Boer, Esther JS Jansen, Anna E Koopmans, Wendy TM Enthoven, et al. Effect of supplementation with zinc and other micronutrients on malaria in tanzanian children: a randomised trial. *PLoS medicine*, 8(11):e1001125, 2011.

K Vijayaraghavan, G Radhaiah, B Surya Prakasam, KV R Sarma, and Vinodini Reddy. Effect of massive dose vitamin a on morbidity and mortality in indian children. *The Lancet*, 336(8727):1342–1345, 1990.

Gary R Watmough, Charlotte LJ Marcinko, Clare Sullivan, Kevin Tschirhart, Patrick K Mutuo, Cheryl A Palm, and Jens-Christian Svenning. Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences*, 116(4):1213–1218, 2019.

Keith P West Jr, Joanne Katz, Steven Charles LeClerq, EK Pradhan, James M Tielsch, Alfred Sommer, RP Pokhrel, SK Khatry, SR Shrestha, and MR Pandey. Efficacy of vitamin a in reducing preschool child mortality in nepal. *The Lancet*, 338(8759):67–71, 1991.

WHO. WHO novel coronavirus (COVID-19) situation. https://who.sprinklr.com/, 2020. Accessed: 2020-04-13.

WHO Ebola Response Team. Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *New England Journal of Medicine*, 371(16):1481–1495, 2014.

Wikipedia. COVID-19 pandemic in Iran — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_Iran&oldid=956402285, 2020a.

Wikipedia. COVID-19 pandemic in France — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_France&oldid=956505489, 2020b.

Wikipedia. COVID-19 pandemic lockdown in Hubei — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_lockdown_in_Hubei&oldid=955933271, 2020c.

Wikipedia. COVID-19 pandemic lockdown in Italy — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_lockdown_in_Italy&oldid=956053371, 2020d.

Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, 2020a.

Joseph T Wu, Kathy Leung, Mary Bushman, Nishant Kishore, Rene Niehus, Pablo M de Salazar, Benjamin J Cowling, Marc Lipsitch, and Gabriel M Leung. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine*, pages 1–5, 2020b.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):1–11, 2020.

Alwyn Young. The african growth miracle. *Journal of Political Economy*, 120(4):696–739, 2012.

Appendix A

Appendix

A.1 Supplementary Figures

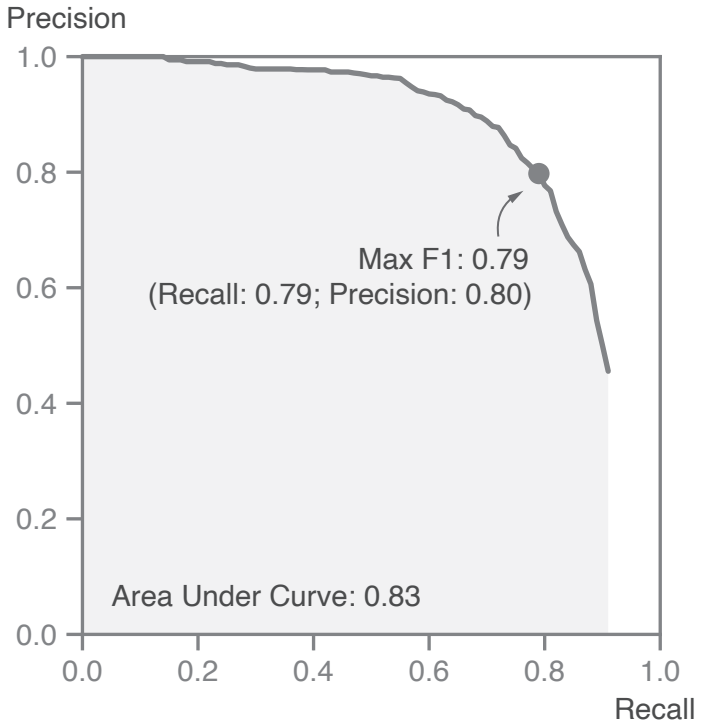

Figure A.1: **The precision-recall curve of the Mask R-CNN model shows satisfactory predictive performance.** The Mask R-CNN model is trained and evaluated with 3-fold cross validation. The evaluation is based on 120 annotated images, which were randomly sampled from all the input satellite images in Siaya, Kenya. The Mask R-CNN model outputs a confidence score for every predicted building instance, and the precision-recall curve is generated by varying the confidence score threshold, below which predicted instances are dropped. A higher threshold makes the model more conservative and corresponds to the left portion of the curve (with high precision and low recall), and vice versa. The dot represents the optimal confidence score threshold, obtained by maximizing F1, the harmonic mean of precision and recall. The main model used in this study employs the optimal threshold, and has a recall of 0.79 and a precision of 0.80.

Figure A.2: **Ten randomly sampled pairs of input images and deep learning predictions.** Ten images are randomly sampled from all the input satellite images in the GiveDirectly study area. Each predicted building is outlined in white and filled with the "representative" roof color.

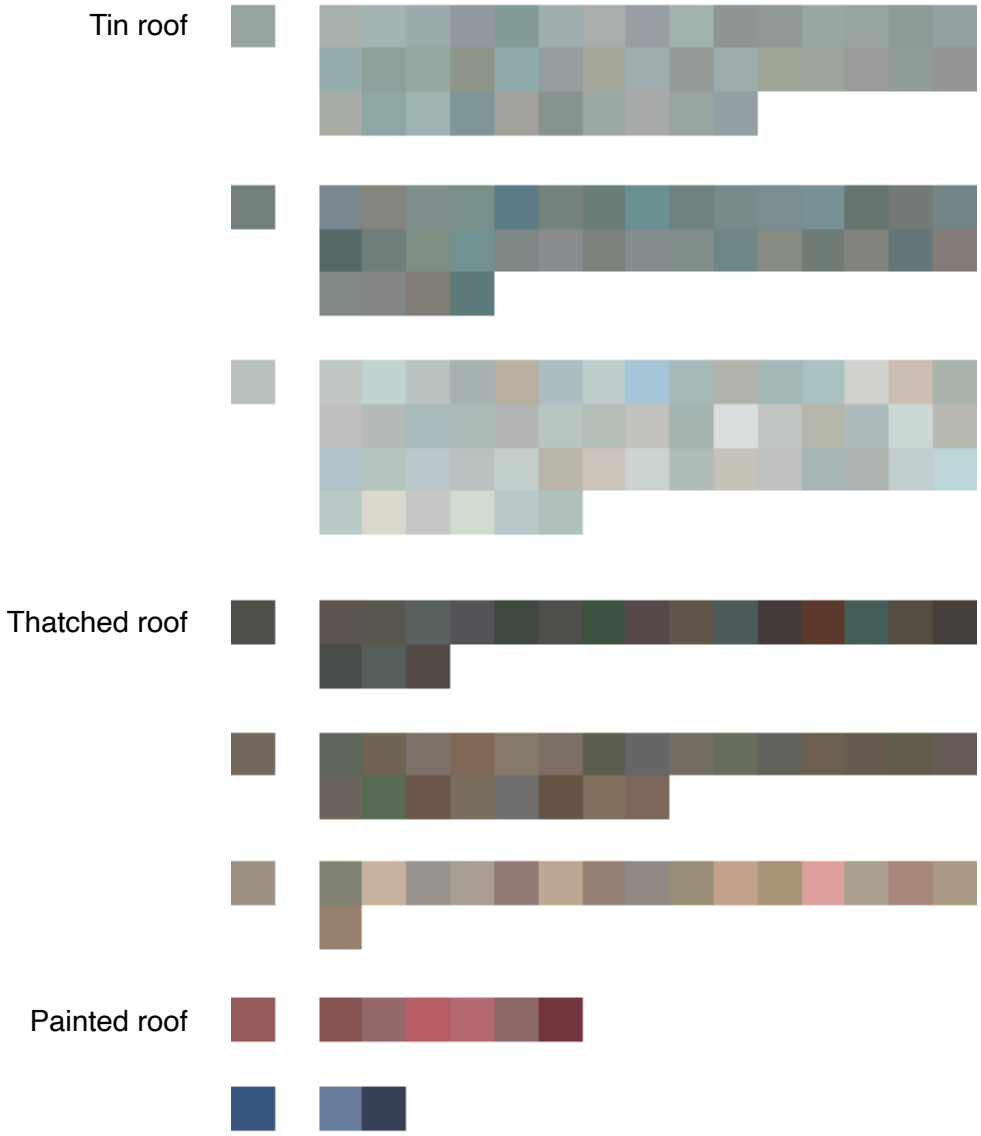

Figure A.3: **The distribution and grouping of roof colors.** All the buildings in the GiveDirectly study area are split into eight groups by a K-means clustering algorithm, based on their roof colors. The color block on the left represents the "average" roof color of the cluster, and the color blocks on the right represent a random subset of all the roof colors in the given cluster. The number of color blocks on the right is proportional to the size of the cluster. The eight groups are further grouped into tin roof, thatched roof, and painted roof.

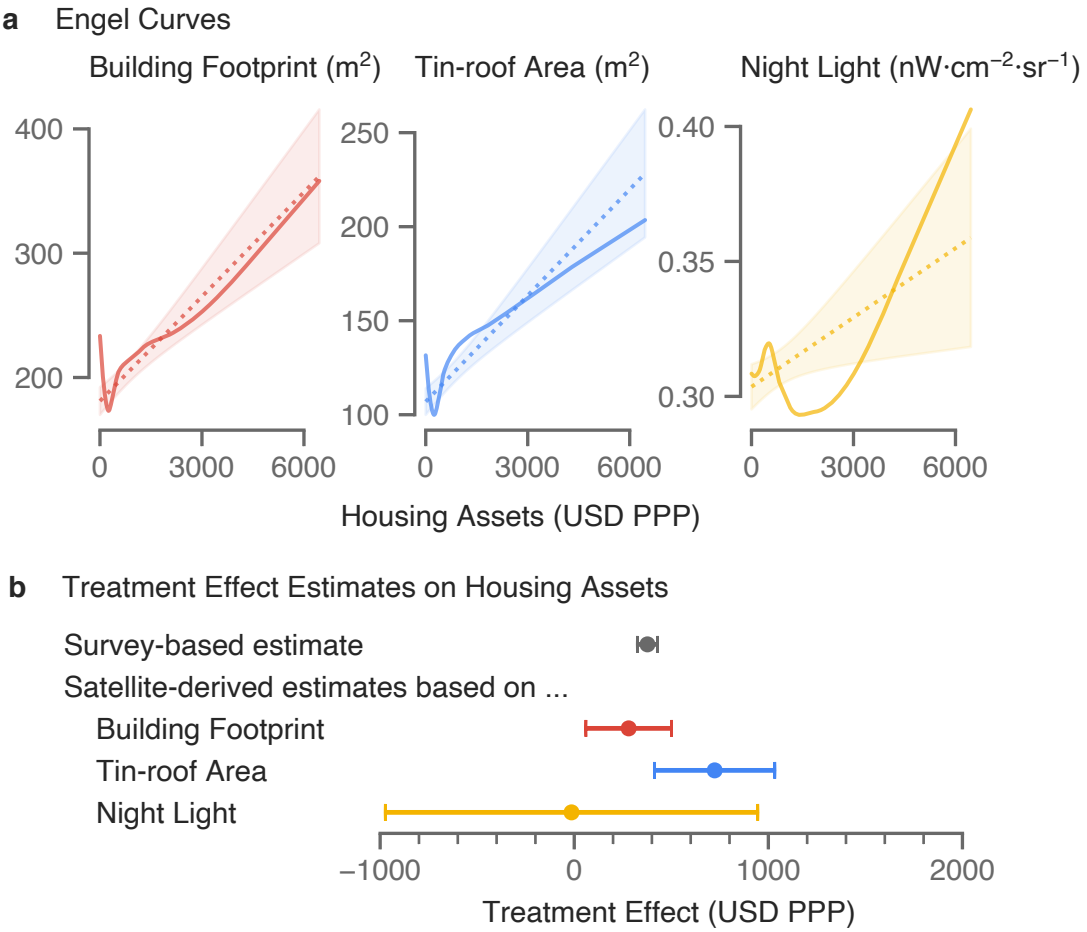

Figure A.4: **The treatment effect on housing assets can be similarly recovered by scaling the effect on building footprint.** **a** The Engel curves of building footprint, tin-roof area, and night light, estimated with LOESS (solid line) or a linear regression (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the survey-based versus satellite-derived treatment effects. The dots show the point estimates. The error bars show the 95% confidence intervals, with the arrow(s) marking upper/lower bounds that are out of range (if any).

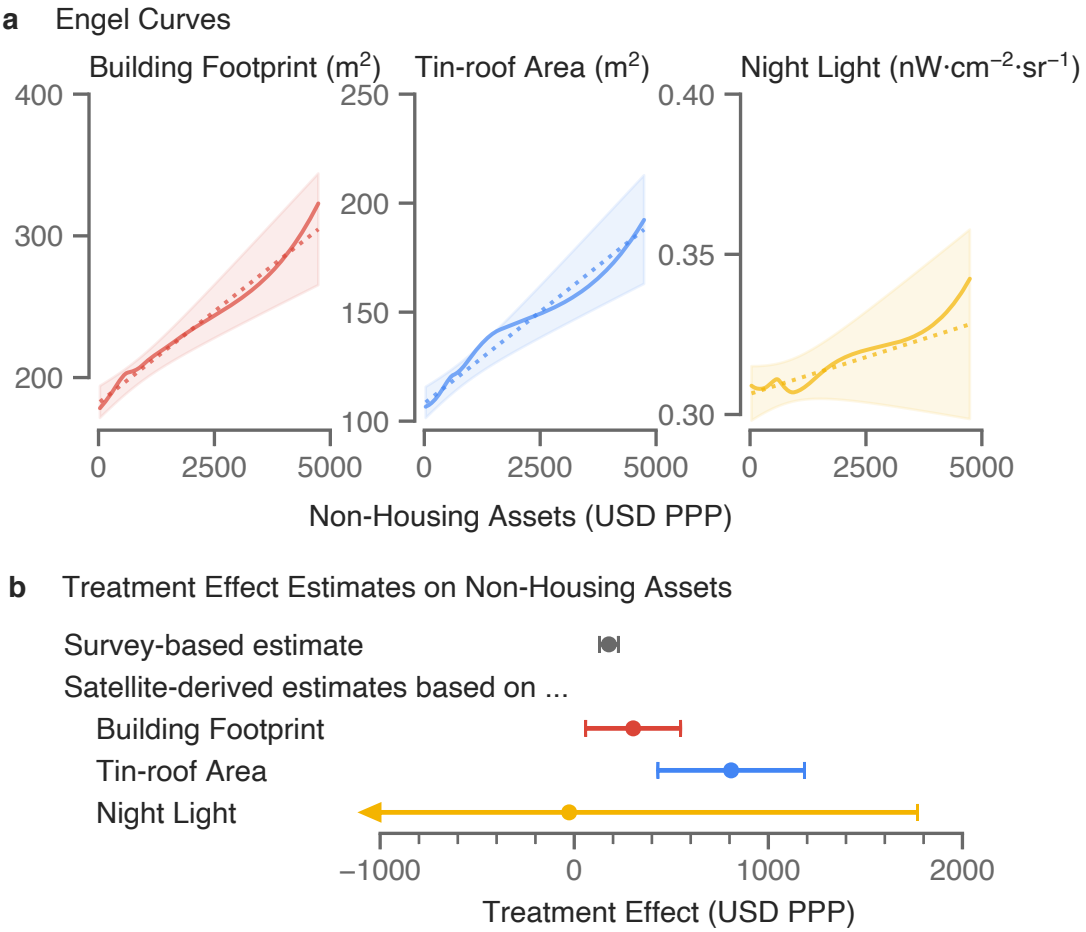

Figure A.5: **The treatment effect on non-housing assets can be similarly recovered by scaling the effect on building footprint. a** The Engel curves of building footprint, tin-roof area, and night light, estimated with LOESS (solid line) or a linear regression (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the survey-based versus satellite-derived treatment effects. The dots show the point estimates. The error bars show the 95% confidence intervals, with the arrow(s) marking upper/lower bounds that are out of range (if any).

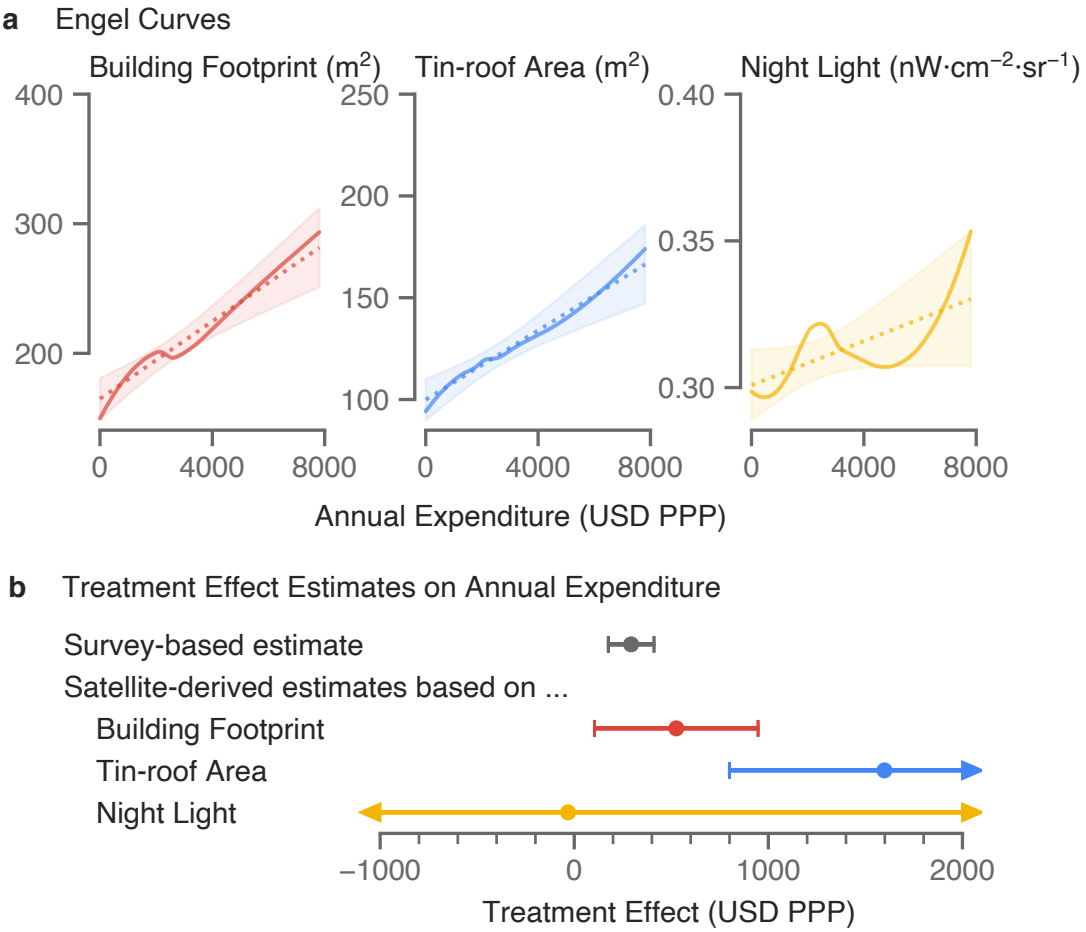

Figure A.6: **The treatment effect on annual consumption expenditure can be similarly recovered by scaling the effect on building footprint.** **a** The Engel curves of building footprint, tin-roof area, and night light, estimated with LOESS (solid line) or a linear regression (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the survey-based versus satellite-derived treatment effects. The dots show the point estimates. The error bars show the 95% confidence intervals, with the arrow(s) marking upper/lower bounds that are out of range (if any).

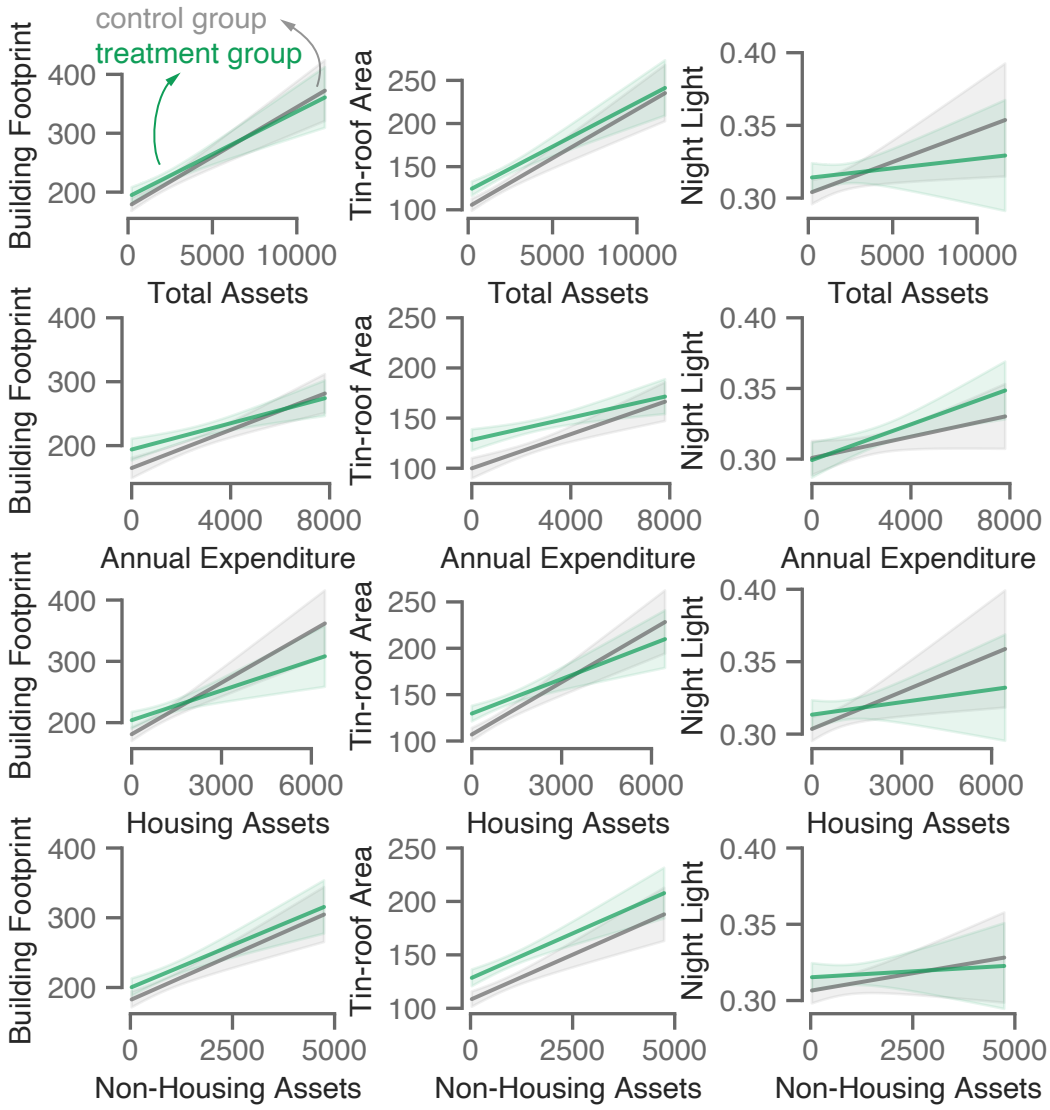

Figure A.7: **The Engel curves for tin-roof area shifted in response to the cash transfer.** The Engel curves for the treatment households (in green) and the control households (in gray). The shaded regions represent the 95% confidence intervals.

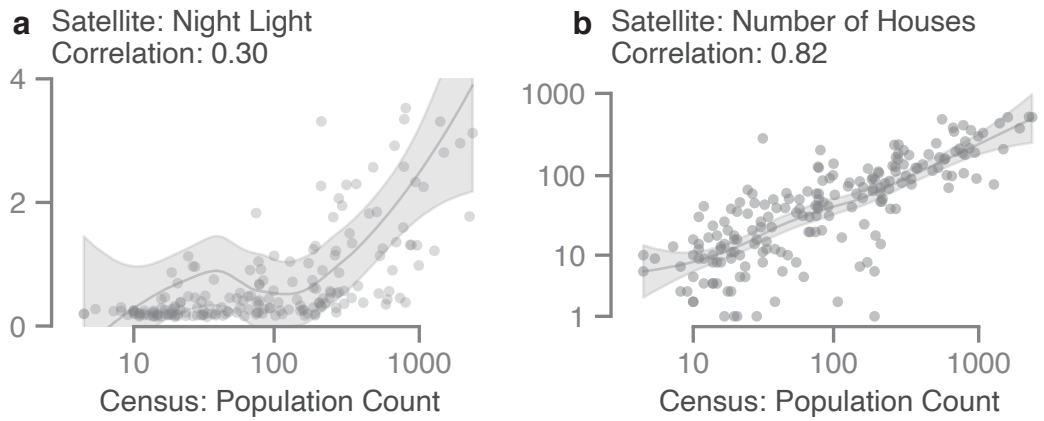

Figure A.8: **Population count in Mexican villages is more strongly correlated with the number of houses in satellite imagery, compared to night light.** The population count is shown in log scale. Each point corresponds to a randomly sampled rural locality in Mexico. Gray lines are estimated LOESS curves, and the shaded regions are the 95% confidence intervals. The (Pearson) correlation coefficients are reported in the panel subtitles.

A.2 Training the Deep Learning Model

Creating In-sample Building Footprint Annotations

I create in-sample building footprint annotations to train the model, and to objectively and quantitatively evaluate model performance. Among the 71,012 satellite images that cover all of the Siaya county in Kenya, I randomly sample 120 images for annotation. I use the Supervisely image annotation web platform to create annotations. On any given image, I outline the boundaries of all the instances of buildings on the image. Buildings that border each other are annotated as separate instances, if there are reasons to believe that they are separate structures (e.g., if they appear to use different roof materials). Half-finished buildings are annotated, although they are fairly rare in the analysis sample.

Some measurement errors can arise from the annotation process, which may in turn impact the predictions of the deep learning model. First, the Google Static Maps logo blocks 1.05% of the total area of any given image, and structures covered by the logos are not annotated. Second, only the visible parts of the buildings are annotated, but a very small part of some buildings may be partially occluded by trees. Third, the annotation accuracy (and thus potentially prediction accuracy) may be different across buildings with different roof materials. In particular, thatched-roof houses tend to be harder to identify for human annotators than metal-roof houses, because they are typically smaller, not as reflective, and may resemble trees in the overhead imagery.

Training the Mask R-CNN Model

I use the Mask R-CNN model (He et al., 2017) for instance segmentation of buildings on satellite images. The backbone architecture used is ResNet50 with the Feature Pyramid Networks. The model is trained with a learning rate of 5×10^{-4} and a batch size of 10. Optimization is conducted with the Adam optimizer. I implement the deep learning pipeline with Python and PyTorch. In particular, I use the official Torchvision implementation of Mask R-CNN. I train the Mask R-CNN model in a transfer learning framework, with a multi-step process as follows.

1. COCO (Common Objects in Context) The model is first pre-trained with the COCO (Common Objects in Context) data set, a large-scale natural image data set containing 80 object categories and around 1.5 million object instances (COCO, 2020). Despite the fact that input images and object categories in COCO are different from target satellite images, pre-training the model with a large-scale dataset often provides meaningful performance gains, even when the model is later transferred across domains.

2. Open AI Tanzania The model is then fine-tuned on the Open AI Tanzania building footprint segmentation data set, a collection of high-resolution aerial imagery collected by

consumer drones in Zanzibar, Tanzania (Open AI Tanzania, 2020). These images are representative of the rural or peri-urban scenes in a developing country context, in terms of the distribution of the density, sizes and heights of the buildings. All the buildings in the drone images are identified, outlined and classified into three categories (completed building, unfinished building, and foundation) by human annotators. This somewhat unusual categorization is due to the fact that there are a large number of unfinished structures in Zanzibar. Most input satellite images in this study contain very few unfinished structures, so I collapse the first two categories into one and drop the third category. The native resolution of the drone images is 7cm, and I down-sample the images to about 30cm to match with the resolution of the target satellite images.

In training time, 90% of the data are used for training, and the remaining 10% for validation. In order to guard against overfitting, and choose the best model, in each epoch, I evaluate the performance of the model with the validation set, using average precision with an Intersection over Union (IoU) cutoff of 0.5 as the main evaluation metric. The model is trained for 50 epochs, and the best model (at epoch 43) is saved and loaded in subsequent steps.

3. Supplementary Annotations in Mexico, Tanzania and Kenya The model is then fine-tuned on a set of 587 annotated high-resolution satellite images from Mexico, Tanzania, and Kenya. The Mexico dataset consists of 199 satellite images corresponding to 8 randomly sampled rural localities studied in Supplementary Figure A.8. Some of these are historical images with lower data quality and more cloud coverage. These images are pooled and randomly split into a training set (90%) and a validation set (10%). The model is trained for 25 epochs, and achieves the best performance at epoch 17.

4. In-sample Annotations Finally, the model is fine-tuned on a set of 120 in-sample annotated images in Siaya, Kenya (see Section A.2 for details). This ensures that training images and inference images belong to the same data distribution. The model is trained on 90% of the images for 25 epochs, and evaluated with the 10% held out set. I keep the best-performing model (at epoch 15). This is the main model used for conducting inference on input satellite images in the GiveDirectly study area.

Throughout the training process, I conduct extensive data augmentation to increase the transferability of the model from one dataset to another. I randomly flip the training images horizontally and vertically, randomly jitter the brightness, contrast, saturation, and hue of the images. For the Open AI Tanzania dataset, I also randomly blur and crop the images.

A.3 Validation in Mexico

Results

I provide additional validation results in rural Mexico, using the 2010 Population and Housing Census (INEGI, 2010). Population count in a rural village (as reported in the 2010 census), is highly correlated with the number of houses in that village (as identified by the deep learning model), with a Pearson correlation coefficient of 0.82 (Supplementary Figure A.8b). Population count, however, is only modestly correlated with night light (Supplementary Figure A.8a). Night light is less sensitive in smaller, less populated villages, a finding that is consistent with prior work (Jean et al., 2016a).

Methods

This comparison is based on the locality-level data set, Principales Resultados por Localidad, or ITER. (A locality is equivalent to a village in rural areas.) To form the analysis sample, I drop all urban localities (defined as having more than 2,500 residents), small localities where the relevant asset measures are masked in the census to protect privacy, and localities where these measures are missing. To avoid covering neighboring urban or rural localities in the satellite images, I exclude rural localities that are closer than 0.01 degree (1.1 km) from other rural localities, or 0.1 degree (11.1 km) from urban localities. Finally, to reduce computation, I randomly sample 200 rural localities, and drop 3 of them, for which Google Static Maps does not have satellite image coverage for.

In the census, each rural locality is geo-coded as a point. Most of the rural localities are small, isolated and surrounded by vegetation or open space, making it feasible to match census records to corresponding satellite images. For each locality, I obtain satellite images that cover an area of roughly 1×1 km, with the locality coordinate at the center. The images are retrieved from the Google Static Maps API on October 10, 2019, and are likely taken several years after the census. I generate deep learning predictions on these images with the method described in Methods and Supplementary Materials A.2, but only train the model for the first three steps in Supplementary Materials A.2. For the comparison, I count the number of houses in a locality in the deep learning predictions, and extract the population count variable from the census. Additionally, I download night light data, the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images from 2019.