Randomized Controlled Trials in West Africa

Practice and Theory

Universität Hamburg

Fakultät für Wirtschafts- und Sozialwissenschaften

(kumulative) Dissertation

Zur Erlangung der Würde des Doktors der
Wirtschafts- und Sozialwissenschaften

– Dr. phil. –

(gemäß der PromO vom 24. August 2010)

vorgelegt von

Felipe Alexander Dunsch
aus Hamburg

Hamburg, 11.11.2019

UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Kumulative Dissertation

Vorsitz der Prüfungskommission: Prof. Dr. Kai-Uwe Schnapp

Erstgutachter: Prof. Dr. Cord Jakobeit

Zweitgutachterin: Prof. Dr. Vera Troeger

Disputationsdatum: 14.4.2022

Table of Contents

List of Tables

List of Tables – continued

List of Figures

Title of Figure *Page*

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of Acronyms & Abbreviations

3IE	International Initiative for Impact Evaluation
ACT	Artemisinin-based Combination Therapy
AF	Available and Functioning
AIDS	Acquired Immune Deficiency Syndrome
ANC	Antenatal care
ANCOVA	Analysis of Covariance
ANF	Available but not Functioning
ATESP	Apprenticeship Training and Entrepreneurial Support Program
ATE	Average Treatment Effect
ATET	Average Treatment Effect on the Treated
BCG	Bacille Calmette-Guérin
BMJ	British Medical Journal
BMZ	Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
CCT	Conditional Cash Transfer
CDD	Community Driven Development
CDF	Sudan Community Development Fund
CEGA	Center for Effective Global Action
CGD	Center for Global Development
CHEW	Community Health Extension Worker
CHO	Community Health Officer
CHPS	Community-Based Health Planning & Services
CHT	Community Health Team
CHV	Community Health Volunteer
CPBF	Community Performance Based Financing
DCE	Discrete Choice Experiment
DFID	Department for International Development
DIME	Development Impact Evaluation Unit – World Bank
EFinA	Enhancing Financial Innovation and Access
FDR	False Discovery Rate
FGD	Focus Group Discussion
FMOH	Federal Ministry of Health
FWER	Family-wise Error Rate
GDP	Gross Domestic Product
GHS	Ghana Health Services
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit
GPS	Global Positioning System
GUP	Graduation from Ultra Poverty
HIV	Human Immunodeficiency Viruses
HRH	Human Resources for Health
HRITF	World Bank Health Results Innovation Trust Fund
HSDF	The Health Strategy and Delivery Foundation
IBM	International Business Machines Corporation
IDA	International Development Association
IE	Impact Evaluation
IFPRI	International Food Policy Research Institute
IID	independent and identically distributed
IPA	Innovations for Poverty Action
INR	Item Non-Response
IPT	Intermittent Preventive Therapy
IRB	Institutional Review Board

ITT	Intention to Treat
JCHEW	Junior Community Health Extension Worker
J-PAL	Jameel Abdul Latif Poverty Action Lab
LATE	Local Average Treatment Effect
LG	Local Government
LSE	London School of Economics
LSS	Life Saving Skills
MA	Master of Arts
MCA	Millennium Challenge Account
MCH	Maternal and Child Health
MFB	Micro-Finance Bank
MG	Magnesium
MHT	Multiple-Hypotheses Testing
MIT	Massachusetts Institute of Technology
MLSS	Modified Life Saving Skills
MSc	Master of Science
MSME	Micro, Small & Medium Enterprises
MV	Millennium Villages
NA	Not available
NGO	Non-Governmental Organization
NHIS	National Health Insurance Scheme
NICE	National Institute for Health and Care Excellence
OIC	Officer in Charge
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
OPD	Outpatient Department
ORS	Oral Rehydration Salts
PC	Pure Control
PEQ	Patient Experience Questionnaire
PHC	Primary Health Care Center
PNC	Postnatal care
PPE	Picker Patient Experience Questionnaire
PROGRESA	Programa de Educacion, Salud y Alimentacion
QIP	Quality Improvement Plan
RCT	Randomized Controlled Trial
SD	Standard Deviation
SDI	Service Delivery Indicators
SE	Standard Error
SIEF	Strategic Impact Evaluation Fund – World Bank
SIGN	Scottish Intercollegiate Guidelines Network
SP	Stated Preference
SSA	Sub Saharan Africa
SURE-P	Subsidy Reinvestment and Empowerment Program
UHC	Universal Health Care
UK	United Kingdom
UNICEF	United Nations Children's Fund
US	United States of America
USD	United States Dollars
WDC	Ward Development Committee
WHO	World Health Organisation
WTAC	Willingness to Accept Compensation
WTP	Willingness to Pay
WWGS	What Works Global Summit

1. Introduction & Motivation

Since the turn of the century, there has been a gradual shift towards a professionalization of the evaluation practice of development projects. This shift can be partly seen as a response to relatively disappointing development results in large regions of the world (namely Sub-Saharan Africa, South-East Asia, and parts of Central- and Latin-America) despite decades of aid flows since the end World War II. Sub-Saharan Africa for example – despite some successes – lags other developing regions in regard to economic growth, employment, health, education, and sanitation. The GDP per capita in Sub-Saharan Africa (SSA) was a meagre 1,574 USD in 2017, which equals less than a 3-fold increase since the 1980s. Latin America saw 5-fold increases of GDP per capita in the same time period to now over 10,000 USD per capita. South Asia, while starting from even lower levels, has outpaced SSA by now (World Bank, 2019).

Thus, for the large majority of Sub-Saharan Africa, the massive increases in development aid flows in past decades have not translated to sustainable reductions in poverty levels. Even more astounding is that multi- and bi-lateral donors regularly give themselves good grades at the project level (World Bank, 2009; Wolff, 2005; Stockmann et al., 2015). Paul Mosley (1987) was one of the first to recognize that development aid successes at the project level often do not translate into growth effects for a region or nation as a whole. With this observation, he coined the "micro-macro-paradox" term. Peter Boone's studies in the mid-1990s were then the first that questioned the utility of development aid for economic growth, employing modern econometric methods (Boone, 1994; Boone, 1996; Faust and Leiderer, 2008).[1] He found that development aid mainly contributes to boosting private consumption but does not impact investments so that there are only marginal long-term growth effects. The discipline subsequently fell into sort of a "knowledge vacuum" – while some tried to explain the micro-macro-paradox, others denied its existence (Easterly, 2006).

Only a few years after Boone's publications, Craig Burnside and David Dollar (2000) published a paper in the *American Economic Review*, which turned out to be very influential and was recognized as a breakthrough in the discipline. Their regression models showed that development aid has no measurable impact on economic growth unless it is interacted with "good governance". Large multilateral donors such as the IMF and the World Bank changed their disbursement practices to take into account these new results (Dalgaard, 2004). The Millennium Challenge Account (MCA), founded in 2004 by the US, oriented itself according to this finding to increase the effectiveness of its spending (Banerjee, 2005).

[1] This includes the use of instrumental variables such as population size and controlling for reverse causality of development aid and economic growth.

However, a few more years later, William Easterly et al. wrote a counter-article (2004), also published in the *American Economic Review*, which used the same econometric model as Burnside and Dollar, however using updated data. The initially significant relationship from the original article disappeared using this updated data, apparently lacking statistical robustness (in this case due to the updated data). The ensuing back-and-forth of renowned economists (and the debates that ensued) exemplifies the dilemma of modern effectiveness research at the macro level. David Roodman (2007, p. 20) determined: "The quantitative approach to answering grand questions about aid effectiveness has repeatedly offered hope and repeatedly disappointed. (...) On balance, it seems that the macro research has attracted attention out of proportion to its value." Similarly, MIT economist Abhijit Banerjee (2008) states:

> It is not clear to us that the best way to get growth is to do growth policy of any form. Perhaps making growth happen is ultimately beyond our control. Maybe all that happens is something goes right for once (privatized agriculture raises incomes in rural China) and then that sparks growth somewhere else in the economy, and so on. Perhaps we will never learn where it will start or what will make it continue. The best we can do in that world is to hold the fort till that initial spark arrives: make sure that there is not too much human misery, maintain the social equilibrium, try to make sure that there is enough human capital around to take advantage of the spark when it arrives. (Banerjee, 2008, p. 15)

This should be a clear rejection of the holistic-teleological approaches of the past, which have implied that the “West" has the power to bring about growth and development through external interventions.

The knowledge „vacuum“ around core assumptions of development aid at the macro level opened the scientific playing field for new theories and approaches. Among others, it also allowed the “Sachs-Easterly-Moyo dispute” to play out (Dunsch, 2012). While Jeffrey Sachs propagated the thesis of “poverty traps” and advocated for development aid to vastly increase to “push” people out of it (Sachs et al., 2004; Sachs, 2005; Sachs, 2008), William Easterly argued for ending large scale promises and declarations and advocated instead a more incremental approach to aid (Easterly, 2006). Sachs is convinced that only integrated approaches of numerous programs at the same time with a massive increase in aid spending can achieve the goal of ending poverty (Sachs, 2005). Easterly, on the other hand, preferred a step-by-step trial-and-error approach to advance knowledge (Easterly, 2006). Another, more radical, voice in the debate heated debate was Dambisa Moyo, who gained a lot of attention with her book “Dead Aid” in 2009.[2] Moyo recognized

[2] In Germany the Sachs-Easterly-Moyo dispute (around the utility of foreign aid and reasons why economic growth does not necessarily go hand in hand with poverty reduction in sub-Saharan Africa) also stimulated a series of publications, e.g. by Andresen (2012), Neudeck (2010), Seitz (2014) and Johnson (2011).

development aid as the main *problem* for African development. Development aid would create dependencies, she argued, as well as entail economic inefficiencies. In sum, these would be detrimental for economic growth. Next to their implicit and seemingly not always fully thought-out acceptance of modernization theory (Sachs) and neo-liberal ideas (Moyo)[3], the three protagonists at their core agree that development aid (and its evaluation practice) requires urgent reform.

The limited research successes on the macro-level paired with the confusions around what the best ways would be to implement development aid led to the realization that we know a lot less about development than initially believed. It became clear that existing monitoring and evaluation systems have grave weaknesses and that the disbursements for projects only weakly correlate with the actual performance of donors and recipients (Radelet and Levine, 2008; CGD, 2006), but aid flows increased. Already early, Peter Bauer identified one of the causes for the seemingly "escalation" of interventions in the development space: „Whatever happens, progress or failure becomes an argument for more aid" (Bauer, 1976). These issues were also gradually recognized by the global donor community, for example through the 4 successive High-Level Forums on Aid Effectiveness in Rome (2003), Paris (2005), Accra (2008), and Busan (2011).[4] While evaluations have been a byproduct of development cooperation and development aid for a long time, these global debates around aid effectiveness have contributed to the recent re-focus and the said professionalization of evaluations. Previously, evaluations were mostly focused on simple input-output relationships, as the *correct* use of public funds was at the center. Now, the paradigm has shifted towards utilizing evaluations to robustly assess cause-and-effect relationships (causal impacts) of aid, development projects, and policies (Caspari and Barbu, 2008).[5] This trend has is most visible through the expanded use of "impact evaluations" (IEs), of which (experimental) randomized controlled trials (RCTs) are the most prominent method. With this, the focus of evaluating the effectiveness of development aid became "smaller" and moved to the level of projects: If it is impossible to ascertain whether aid is effective or creates growth for a country, it might at least be possible to show that aid has effects on the program or project level.[6] The trend of using

[3] See also: Jakobeit, 2009.

[4] The overarching goal of the Paris Declaration (OECD, 2005) was to increase aid effectiveness following 5 main principles: *Ownership, Alignment, Harmonisation, Managing for Results* and *Accountability*. This dissertation discusses effectiveness in its narrower sense, mainly touching on *Accountability* and *Managing for Results* (Ashoff, 2010). Higher-level issues, for example lack of coordination and harmonization among donors are obviously as (if not more) important that focusing on effectiveness on the micro-level. These are outside of the scope of this dissertation.

[5] In a widely recognized study of evaluations in development, the Center for Global Development (CGD) states: „Yet after decades in which development agencies have disbursed billions of dollars for social programs, and developing country governments and nongovernmental organizations (NGOs) have spent hundreds of billions more, it is deeply disappointing to recognize that we know relatively little about the net impacts of most of these social programs" (CGD, 2006, p. 1).

[6] Karlan and Appel (2011) even state that the debate (on larger questions of development) is futile and leads to "stagnation and inertia" (cited in de Souza Leão and Eyal, 2019, p. 404)

IEs can paradoxically be interpreted as a regression from measuring impacts of aid as a holistic concept. Impact evaluations are a step "back" to a level of more manageable outputs and outcomes using specific indicators, which are, in contrast to more complex (and some would argue more important) concepts such as the quality of education or the health care systems of a country, easier to measure. This is in line with the more "step-wise" (Easterly 2006; 2009) approach to conduct development aid and implement projects in general, which has found its way into the mainstream as opposed to the previously more prominent "big push" models (Rostow, 1959; Sachs, 2005),[7] and an acknowledgement that it is nearly impossible to measure the impact of all of aid on complex systems such as growth (Nuscheler, 2008). IEs have now found their way into the mainstream of development, as exemplified by the World Bank's IDA-16 replenishments, where the use of impact evaluations was explicitly mentioned for the first time[8] (World Bank, 2011), and most prominently with Esther Duflo, Abhijit Banerjee, and Michael Kremer winning the Nobel Prize in Economics in 2019 for their pioneering use of RCTs in developing countries.

While evaluations with a narrower focus are not a novelty, but the fast expansion and professionalization of methods during the last 10 to 15 years is astounding.[9] Of all policy evaluation fields, somewhat surprisingly development economics to date appears to use the most advanced quantitative methods to measure causal impacts (even though, however, only as a small percentage of all evaluation efforts). Experiments have a long tradition in medicine, economics, and psychology, but only more recently gained traction in development economics and political science (Faas and Huber, 2010; Druckmann et al., 2011).[10]

The existing debates around the value of IEs and RCTs, in addition to the apparent dearth of existing experimental research especially for and in West Africa, sparked the motivation for this dissertation. Figure 1.1 shows the impact evaluation study density for low- and middle-income countries. The map illustrates that experimental evidence is scarce in the West African region, especially when compared to South and East Asia, East and South Africa, and Latin America. The

[7] The Millennium Villages (MVs) of the United Nations (led by Jeffrey Sachs) are based on the "big push" idea. The MVs receive a lot of attention from mainstream media, but they are seen critically by the majority of the development economics discipline (Clemens and Demombynes, 2011).

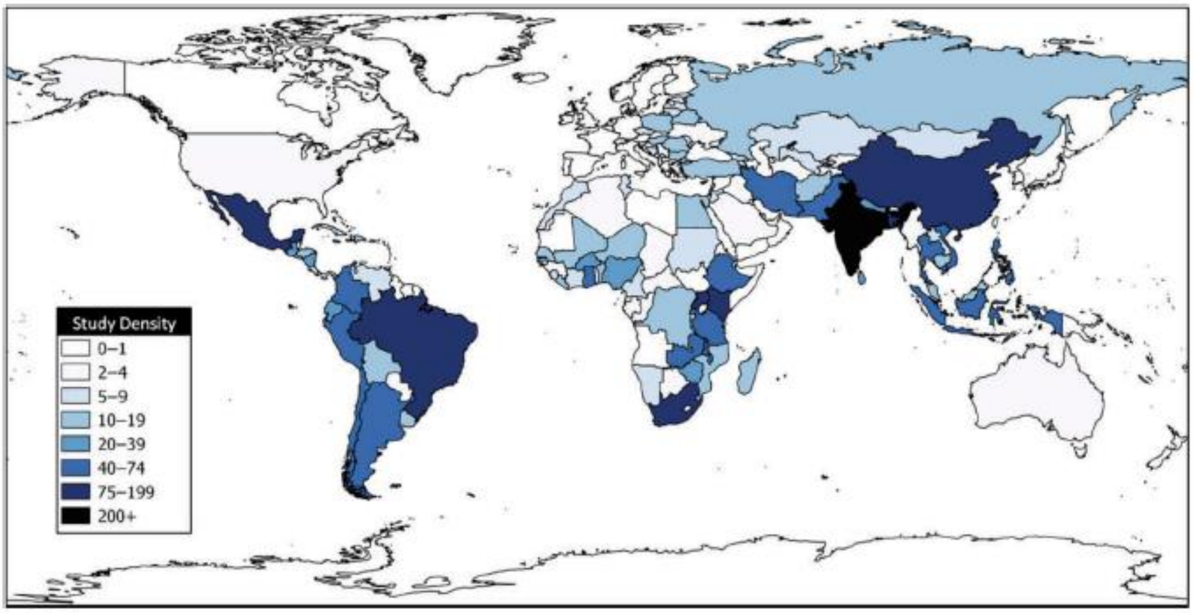
[8] In a 2011 position paper of the German Federal Ministry for Economic Cooperation and Development, German Minister for Development Cooperation Dirk Niebel also explicitly emphasized that "Effectiveness is important to us" (BMZ, 2011).

[9] De Souza Leão and Eyal (2019) speak of "waves" of RCTs in this context (the trend we are seeing now is the 2nd wave of RCTs in development).

[10] Not only the "classic" fields of evaluation research, such as government interventions in social, labor market, or education policies, but also in other fields the question of impact has gained ground and has become a central challenge (Hegemann et al., 2013), for example in peace and conflict studies (Paffenholz, 2006; Pearson et al., 2012), international environment policy (Miles, 2002; Oberthür and Stokke, 2011) or in general the research on international regimes (Hasenclever et al., 1997). However, the boundaries of causal impact measurements in these fields are reached quickly. Quasi-experimental research remained the exception and it requires a lot of statistical maneuvering to credibly exclude impacts of confounding variables.

leading theme of this dissertation is thus the question to what extend these IEs/RCTs in the field of development can successfully identify cause-and-effect relationships and ultimately to assess the wider utility of RCTs to provide actionable policy advice.

Figure 1.1 – Density of low- and middle-income country impact evaluations (1981–2012) – (Cameron et al., 2016)



To do so this dissertation is combining 5 individual articles. The first 4 articles are evaluations (case studies) in the form of self-contained RCTs, that were conducted in Ghana and Nigeria (chapters 2 to 5), in collaboration with World Bank economists. They cover aspects of the sectors of private sector development and health, each with different epistemological and practical (policy-relevant) goals. While each study is answering – by definition – a rather small development question, the results can nevertheless be important to inform policies and to ignite future research, which is important for West Africa that – in addition to deep poverty – suffers from a dearth of solid quantitative development research of any kind. All collected data for the 4 RCTs is primary data that was gathered under my leadership in the countries during numerous trips using self-designed surveys. Chapter 6 is a theoretical and summarizing essay discussing the important problem of extrapolating results from experiments to other contexts (external validity), which could be perceived as the most significant aspect when one attempts to move from experimental results towards actionable policy advice. Chapter 7 combines, concludes, and provides an outlook.

The rest of this introductory chapter provides a cursory overview of what IEs are (1.1). Then, the pros and cons of the method are briefly addressed (1.2 and 1.3). A more extensive discussion on the advantages and disadvantages of IEs can also be found in a previous publication (Dunsch, 2012) and in chapter 6.

1.1 What are Impact Evaluations?

As stated, a popular academic trend in the past years has been the increased usage of “Impact Evaluations” (IEs), which became a new standard in evaluation practice (DeGEval 2008, 2010; Quack and Sprenger 2010) but also advanced to “state-of-the-art” in development economics. (Faust and Neubert 2010; Stockmann and Meyer, 2009). De Souza Leão and Eyal (2019) describe RCTs as a “hinge” between academic economics and development aid. “Impact“ is understood by the OECD as „positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended” (OECD, 2002, p. 24). IEs are then “studies that measure the impact directly attributable to a specific program or policy, as distinct from other potential explanatory factors” (CGD, 2006, p. 10). For IEs, of which *Randomized Controlled Trials* (RCTs) are a subgroup – creating a control group – the counterfactual – is key (Gertler et al., 2016). Some proponents of RCTs are convinced that this method is essentially the only way to make statements about causal chains and ultimately about effectiveness of programs.

Especially in the USA there has been a number of recent publications praising the value of RCTs.[11] RCT frontrunners Abhijit Banerjee and Esther Duflo speak of an “explosion” of the method (Banerjee and Duflo, 2009) and Angrist and Pischke (2010) suggest that microeconomics is undergoing a “credibility revolution” with an increased academic and political weight. In his column in the New York Times, Pulitzer Prize winner suggested that RCTs are “the hottest thing in the fight against poverty (Kristof, 2011) and Duflo and Kremer (2005, p. 230) stated: "Just as randomized trials revolutionized medicine in the twentieth century, they have the possibility to revolutionize social policy during the twenty-first". These proponents see RCTs as “global public goods” which can serve as guidance to development practitioners and policymakers. Deaton describes the debate surrounding the origins and objectives of RCTs as follows:

> Skepticism about econometrics, doubts about the usefulness of structural models in economics, and the endless wrangling over identification and instrumental variables has led to a search for alternative ways of learning about development. There has also been frustration with the World Bank’s apparent failure to learn from its own projects and its inability to provide a convincing argument that its past activities have enhanced economic growth and poverty reduction. (…) For many economists, and particularly for the group at the Poverty Action Lab at MIT, the solution has been to move toward randomized controlled trials of projects, programs, and policies. RCTs are seen as generating gold

[11] The *International Initiative for Impact Evaluation* (3ie) suggested the introduction of an indicator which evaluates donors in regard to their use of rigorous evaluation methods (Nohr and Schmidt, 2012).

> standard evidence that is superior to econometric evidence and that is immune to the methodological criticisms that are typically directed at econometric analyses. Another aim of the program is to persuade the World Bank to replace its current evaluation methods with RCTs (…) (Deaton, 2010, p. 437-8).

Figures 1.1 and 1.2 show the trend of the use of impact evaluation for development. Especially the founding of the Jameel Abdul Latif Poverty Action Lab (J-PAL; affiliated with the Massachusetts Institute of Technology) in 2003 and the Development Impact Evaluation (DIME) Unit in the World Bank (2005), as long with the World Bank's Spanish Impact Evaluation Fund (SIEF) accelerated this trend. While being a "staple" of international development economics, RCTs are now also being commissioned more frequently to study behavioral trends of the population to optimize domestic public policy. The UK has started the "Behavioral Insights Team", and the US White House has also emphasized the value of randomized controlled trials (Burwell et al., 2013). There are also other donor and research organizations that exclusively fund or employ RCTs or quasi-experimental studies, such as the International Initiative for Impact Evaluation (3IE), the Center for Effective Global Action (CEGA) of the University of California at Berkeley and others.[12]

However, there is also a more skeptical or critical view on RCTs, underscoring ethical, technical, practical, financial, and theoretical constraints and limitations of the method. Angus Deaton (2010), a leading critic who stresses the overemphasized use and value of RCTs states, for example:

> Past development practice is seen as a succession of fads, with one supposed magic bullet replacing another – from planning to infrastructure to human capital to structural adjustment to health and social capital to the environment and back to infrastructure – a process that seems not to be guided by progressive learning. (Deaton, 2010, p. 437)

[12] For RCTs Kucklick (2012) estimates the count of involved economists at „about 60", that have published around 320 studies (as of 2012).

Figure 1.2: Trend of using Impact Evaluations in the field of development (Savedoff, 2013)

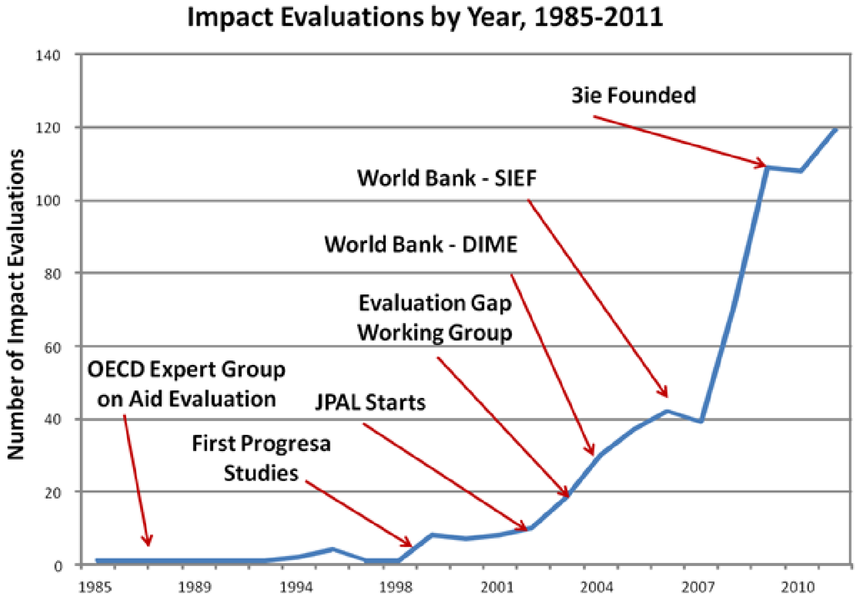
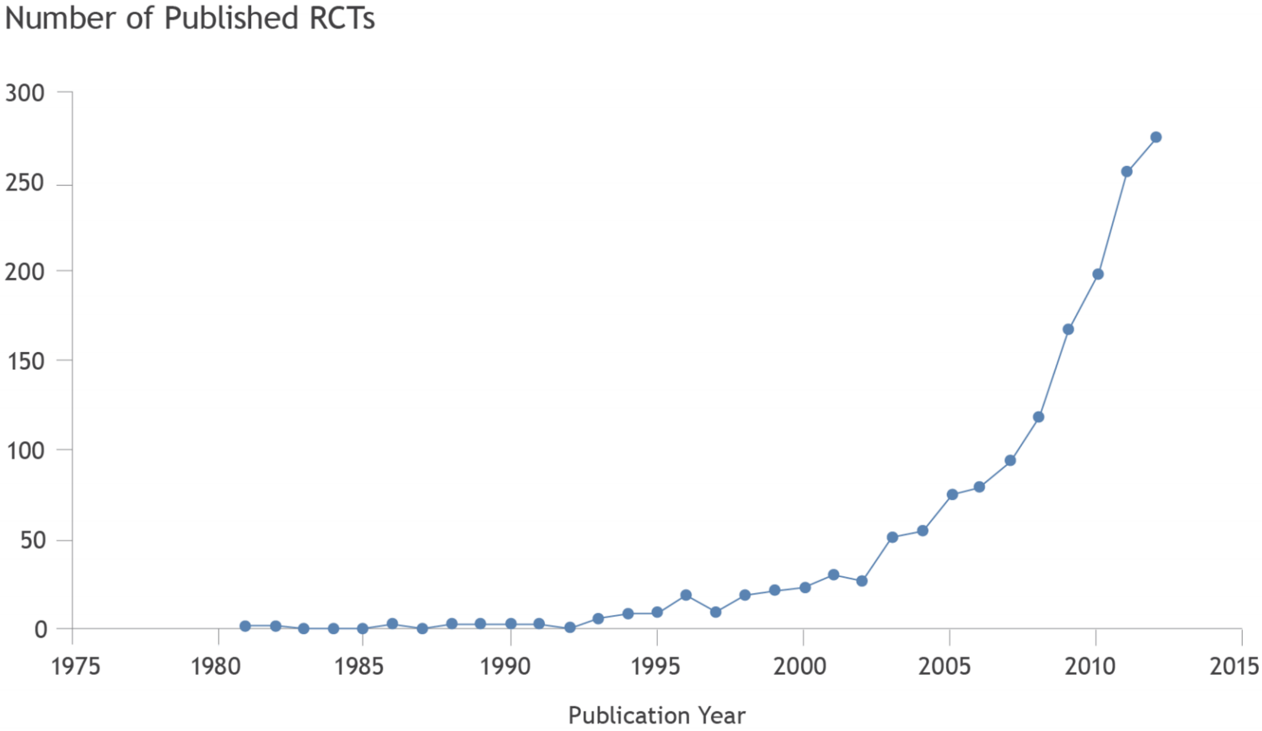


Figure 1.3: Number of Published RCTs (Banerjee et al., 2016)



On the micro-level impact evaluations are distinguished from traditional evaluations by clearly articulating an identification strategy of the program effects. The leading research questions become, “what would have happened in absence of the program?”, instead of the formerly “were positive trends observed during the time of the intervention?” (Gertler et al., 2016). Only the former question truly tackles the question of the benefit of the respective intervention, the latter asks for the sum of all developments in the "target area" and therefore misses the original goal: "The empirical gold standard in the social sciences is to estimate a causal effect of some action" (Al-Ubaydli and List, 2013, p. 3).[13]

The "traditional" way of measuring impact, which is still the norm for the majority of development project evaluations, is in most cases an unsystematic before-and-after comparison of project goals, or even more simplistic: a subjective ex-post experience report. Tracking progress of participating individuals over time, however, does not suffice as a plethora of other factors might have an impact (positive or negative) on the desired results. In addition, this approach lacks a valid comparison group, tracking similar individuals that were not exposed to the project. This before-and-after approach therefore makes two mistakes which prohibit claims to causality.

1) The counterfactual, or the situation the subjects would be in an alternative state without the project, and

2) Correlation is mistaken for causation.

The mere temporal coincidence of two events does not imply that the intervention caused the effect, which is why establishing a counterfactual is required (for example by means of randomization).

Table 1.1 shows the simple example of a fictional development project. The base level of the outcome variable x is 100. After the end of the project, the level has risen to 110 (+10%). Using traditional evaluation methods (before-and-after), the implementing organization would most like assess this project as a success. They would omit the fact that x has also risen in the area without the project (by 20%). Comparing the project and the no-project group, we see that the project actually slowed down progress in x, rather than enhancing it. Most evaluations in the development sector do not capture this effect.

[13] The belief that experiments can reveal the effects of interventions is grounded in a positivist world view. From a post-positivist perspective, methodological rigor itself does not suffice to make true statements about the world (Lee, 2000). What appears to be comprehensible in a specific context, can turn out to be a social construct in another context. As a third theoretical approach is transformative-emancipatory, in which the recipients of an intervention become the focus and the neutral position of the observer is given up for a position of advocacy for a marginalized group (Mertens 2000). In evaluation research, the three approaches – positivism, post-positivism, and transformative-emancipatory – are not entirely antagonistic. Until recently an implicit consensus ensured that multidisciplinary approaches were possible (Lee, 2000). It was not until the recent "hype" around RCTs started that this consensus lost some of its weight.

Table 1.1: Comparison of a Hypothetical Project

	Project	No Project *(counterfactual)*
Level x at baseline	100	100
Level x + t at endline	110	120
Subjective assessment (only looking at the project group	+10%	(+20%)
True effect of project identified by a RCT	-10%	

RCTs can be understood as a tool that allows to isolate the impacts of a project from other sources of impacts, i.e. excluding alternative explanations of the change in the outcome variable. To create a counterfactual, the target group subjects are randomly separated into one or more subgroups (control and treatment). Is the sample large enough, it can be assured that the groups, on average have the same observable and non-observable characteristics. The only difference between the groups would be that the treatment group receives the respective intervention and the control group does not. After a set time frame, data is collected from all subjects (most often in the form of surveys), and the averages of the outcome variables are compared. As the groups are identical on average, the difference can be declared the impact of the project. Most RCTs employ a baseline survey and at least one follow-up survey. The baseline is useful to check whether the randomization was successful.[14] Baseline values of the outcome variable can also be used as control variables for the final analysis to increase statistical precision (McKenzie, 2012). The timing of the follow-up surveys is important as it can be used to measure short-term or longer-term effects.

In case it is not possible to use randomization, there are a series of quasi-experimental methods meant to simulate experiments. This includes "matching" techniques (e.g. Al-Ubaydli and List, 2013), regression discontinuity designs (akin to natural experiments; Campbell, 1969), and using instrumental variables (Gertler et al., 2016).[15]

There is an active ongoing academic debate about the value of RCTs. While there is a broad consensus on the general usefulness of research method, disagreements, however, are centered around ethics, the correct application of the method, the role of biases, as well as the conclusions that can be drawn from study results to be applied in other contexts (external validity).

[14] As many studies deal with small samples, it is advisable to check whether the treatment and control groups have the same values for the outcome variables that are important. In practice, this is not always done (Barrett und Carter, 2010).

[15] In technical terms, the RCT method is the application of a strictly exogenous instrument.

The next section (1.2) briefly describes the main advantage of RCTs when compared to other methods and then I will briefly explore some fundamental problems of RCTs (1.3), which can limit their usefulness.[16]

1.2 Advantages of Randomized Controlled Trials

The quality and usefulness of RCTs are often discussed along two main dimensions, internal and external validity. Internal validity describes whether an evaluation was conducted technically sound and whether therefore project impacts were identified correctly. However, in order to improve effectiveness of development aid as a whole, external validity is the more important component. If external validity is achieved, it is possible to directly extend results from one context and apply them to other contexts (across space and time) to inform policymaking.

This section is shorter than the next one on challenges, as the biggest advantage of RCTs is powerful and quickly stated: In the ideal case, they allow to isolate causal effects of development projects and to separate them unequivocally from other external influence factors (establishing *internal validity*): „The core purpose of RCTs is to use random assignment in order to ensure that the unconfoundedness assumption essential to identifying an average treatment effect holds" (Barrett and Carter, 2010, p. 522).

Proponents are therefore of the opinion that evidence, generated through RCTs, is superior to other forms of knowledge generation (Imbens, 2010) and is therefore best suited to serve as guidance for policymakers. (Whether this claim for *external validity* is correct is the topic of chapter 6.)

Proponents like Olken (2009) for example think that policymakers cannot afford to ignore the "hard" evidence that RCTs create, and that they should only start or continue projects that were proven to work. As already mentioned, Duflo and Kremer even assign RCT results the status of global public goods:[17]

> The benefits of knowing which programs work and which do not extend far beyond any program or agency, and credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and nongovernmental organizations (NGOs) beyond national borders. (Duflo and Kremer, 2005, p. 205)

[16] A more extensive discussion can be found in a previous publication (Dunsch, 2012).

[17] In a "New Yorker" article of 2010 Duflo is cited that the use of RCTs „takes the guesswork, the wizardry (...) out of whether something works or not." Without RCTs we wouldn't be better than „medieval doctors and their leeches" (Parker, 2010).

RCTs are also credited with adding more "credibility" to the evaluation field. By design, RCTs are more transparent as they define the control group already ahead of the baseline survey or the implementation of the project. This ex-ante definition of the control group should reduce the likelihood of data mining or specification searching to some degree (Rasmussen, 2011). Randomization avoids the hand-picking of control groups and diminishes selection bias, which would allow researchers to confirm their preferred theories (Ogden, 2017, p. xxiv), which are unfortunately not uncommon in today's research and evaluation practice.[18]

The most famous *randomistas* Duflo and Banerjee ("Poor Economics - A Radical Rethinking of the Way to Fight Global Poverty", 2011) and Karlan and Appel ("More than Good Intentions – Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy", 2011) wrote entire books on the usefulness of RCTs and their real-life applications. In 2019, Duflo, Banerjee, and Michael Kremer won the Nobel Prize in Economics for having "introduced a new approach to obtaining reliable answers about the best ways to fight global poverty" (The Prize in Economic Sciences, 2019).

For example, in regard to external validity stemming from a well-conducted RCT, an often-cited practical example is the PROGRESA project in Mexico.[19] It is a so-called c*onditional cash transfer* (CCT) program. To augment the presence of girls in Mexican schools, parents were financially incentivized to send them to school. Payouts were tied to the appearance of the girls at school and participation in preventive health care (Gertler and Boyce, 2001). The RCT results show that the disease burden in the treatment group was reduced by 23% in the treatment group and anemia was reduced by 18%. Girls in the treatment group spent 3.4% more time at school. Due to the project's success, the project was scaled up and continued even after a change in government. It is also being applied in other countries (Easterly, 2006). Some argue that RCTs also have contributed to finding out more about the effectiveness of routine mechanisms in the health sector. This includes studies that show that breastfeeding, immunization, oral rehydration and supplementation are all methods to reduce diarrhea diseases in children.[20]

[18] This does not mean that RCTs are totally exempt of publication bias problems. Even when conducting an RCT, some researchers "hedge" against null-results by looking at many outcome variables at the same time and only reporting those that show positive results. Some would argue therefore that, ideally, RCTs are being registered before they are being conducted. In a registry, the researchers name their theories, evaluation questions, analysis methods, and outcome variables. However, this practice is still the exception from the norm. Abhijit Banerjee has proposed a central registry a while ago (Banerjee, 2005). It would eliminate ex-post data mining. Duflo (2006), Duflo and Kremer (2005), and Rasmussen (2011) also wrote in favor of creating such an institution. Next to the potential reduction in data mining practices, a central institution would make it easier to identify evidence gaps and to systematize research. In 2012, the American Economic Association has created such a registry but it is not yet common practice to register trials in advance. It remains to be seen how popular it can become: https://www.socialscienceregistry.org/

[19] After a change in government it was renamed to „OPORTUNIDADES".

[20] However, Easterly (2009) mentions that to come to these conclusions, no RCTs would have been necessary. Imbens (2010) adds in this context, that while it is common knowledge that smokers are more prone to develop lung cancer,

Another field in which RCTs seemingly have showed interesting results are irrationally high price elasticities in the demand for health services in developing countries (Dunsch, 2012). Raising prices on interventions such as vaccines, deworming medication etc., often reduces demand for these products drastically, although these are proven to save lives. These and other results indicate that people who live in poverty (as those that are wealthier) do not strictly behave according to the Homo Oeconomicus model, even when decisions are seemingly easy and can decide about life and death (Al-Ubaydli and List, 2013). Duflo (2006) believes that RCTs contributed to finding out that poor people have relatively low self-control and prefer short-term benefits over longer term-benefits more than people with more resources (Duflo, 2006). Another simple example would be that farmers often do not save enough money to purchase fertilizer for the next season. As a result, good harvests are often not contributing to building longer term wealth (Easterly, 2009). Next to bilateral and multilateral donors, some Governments have started to use the method to evaluate their own development programs.[21]

1.3 Problems with Randomized Controlled Trials

This section discusses problems of IEs. It is first touching on issues surrounding problems of internal validity before briefly touching on external validity, which is also the central theme of chapter 6.

1.3.1 Ethical concerns

A key criticism repeatedly made about RCTs is the allegation of breach of ethical principles when people, sometimes without their knowledge, become the subjects of scientific experiments or are denied support as a member of the control group for scientific reasons (Barrett and Carter, 2010). Most institutions that carry out RCTs examine the research projects through so-called institutional review boards (IRBs). Nevertheless, questionable research designs are not uncommon. Opponents object to RCTs as a mere means to place as many scientific articles as possible with little use for actual policy influence.[22] It is best practice to gather "informed consent" of people that take part in a study. However, only if participants in a study do *not* know that they are part of a study it can be assumed that participation does not affect their behavior. This so-called

this causal relationship was never uncovered using a RCT. Unreasonably high standards to establish causality could prevent researchers to even starting to engage in complex research projects.

[21] Platz (2012) concludes after surveying 10 German NGOs that IEs are being utilized by those that work with international partner organizations and have accumulated more capacity. These NGOs also state, however, that donors exhibit preferences for the correct and efficient use of funds, rather than using additional money to rigorously prove the effectiveness of a certain intervention.

[22] Barrett and Carter (2010, p. 519) accuse "randomistas" of "herd behavior".

"Hawthorne effect" can only be avoided if the subjects are not informed, which of course can pose an ethical problem, especially if there is uncertainty about the potential social benefits of the intervention (or even potentially detrimental outcomes). However, in most cases, participants are aware of their participation, which in turn is likely to lead to behavioral changes in the subjects, thus potentially distorting the outcome (Barrett and Carter, 2010).

RCTs are usually not only trying to gauge whether something works, but also how big the effect is (cost-benefit calculation). Especially in the field of health care, these methods are controversial: "[V]aluable (possibly lifesaving) treatments (...) being withheld for scientific research purposes (...) is morally objectionable, of course" (Cohen and Easterly, 2009, p. 19).

RCT proponent Esther Duflo admits that she would not carry out any experiments where it would be clear that the control group would be at a great disadvantage (2006). The belief that RCTs violate ethical principles also depends heavily on the assessment of the usefulness of these experiments. Proponents such as Duflo or Banerjee would say that RCTs are not objectionable, because this is the only way to generate knowledge about effectiveness, which ultimately could benefit all the poor. Of course, this is only true if this knowledge drives development cooperation, and at this point opinions differ. Even RCT critics do not think that ethical issues are a major drawback: "Ethical issues seems a poor explanation of why there are evaluations in medicine where it is a matter of life and death and are not evaluations of, for instance, educational innovations. Finally, one would think the ethical issue of 'policy malpractice' (…) through perpetuation of ineffective action is at least as serious an issue as structuring participation in programs of unknown efficacy in order to learn if they are effective" (Pritchett, 2002).

RCTs are easier to justify if funding for an intervention is limited anyway, or intervention is only gradually introduced (phase-in). Since only part of the population can receive the program because of financial constraints, it makes sense to base the classification on a random selection rather than on arbitrariness or other rationalities. On the contrary, if the budget is limited, randomized selection is the fairest distribution method.

Closely related to ethical issues is another point of criticism, the lack of "local ownership". To produce robust internal and - even more difficult - external validity for an experiment requires extensive data collection, data preparation, and data analysis expertise that most developing countries do not have. A sizable group of development economists therefore has developed sort of a monopoly position, which on the one hand is used to produce paper after paper, and on the other hand tries to universalize the RCT method as a "gold standard" and to conduct "capacity building", at the same time falling back into the well-known mistakes of development cooperation: "They [RCT proponents Duflo and Banerjee, A.N.] unapologetically propose a solution they acknowledge to be paternalistic: outside interventions by those who know best" (Besley, 2012, p.

162). It is noteworthy - and yet another paradox – that scientists, on the one hand, are engaged in deep scientific debates about the statistical details of their publications, but on the other hand argue that it is easy for governments and researchers in developing countries to apply these procedures themselves.[23]

1.3.2 Technical and feasibility limitations

In addition to ethical concerns, there are a series of more technical threats to internal validity, some of which are listed below (not exhaustively), to provide an overview.

Attrition, Spillovers, and Distortions

The RCT supporters Duflo and Kremer (2005) describe three important problems related to data which can imperil the internal validity of studies: First, individuals assigned to treatment and control groups may choose not to participate in the intervention or the surveys (attrition). This is problematic if these non-participants are systematically different from "compliers" (those that remain in the study) and attrition might also be different in the treatment and the control group. Second, so-called "spillovers" can distort the result. For example, parts of the control group can receive the treatment through geographic proximity (Cohen and Easterly, 2009), for example through word of mouth in case the treatment is an information campaign. The clean separation of the two groups over the entire project cycle is very difficult. However, a mixing of the two groups makes research design obsolete at worst, as also voiced by Banerjee and Duflo (2017, p. 101): "spillover effects could lead one to misstate a program's overall effect." A third distorting influence would be overlapping work by other institutions, such as NGOs, within the study area. If the researchers are unaware of this, and external factors are not equally distributed among the treatment and control groups, this can pose a problem to the validity of the study.

Issues of Data Accuracy

RCTs require the availability and quality of a lot of data. Most data for RCTs are collected through extensive panel surveys, often in hard-to-access and remote areas of developing countries. To collect data, most researchers hire local companies that specialize in data collection. These companies hire teams varying in size from 20 to 50 enumerators that often work under limited supervision. Extensive training is required, as well as supervision in order to avoid cheating by

[23] There are different ways to run impact evaluations. The Development Impact Evaluation (DIME) unit of the World Bank, for example, tries to work with Governments agencies in developing countries directly and closely to design the research and to get their buy-in. This can mean that research projects end up not being "cutting edge" as they are constrained by trying to answer pressing policy questions for the Government. In the ideal case, capacity of local administrative staff, researchers or policymakers can be built. However, this added capacity often remains superficial, as important necessary statistical skills need years of education and practice and cannot be transferred through one or a few projects. However, at least the transfer of a passive understanding of the method can be valuable for local policymakers to guide future research to let science influence their decisions.

enumerators. There are many sources of concerns during this process. For the studies in this dissertation (chapters 2-5), we used tablet computers to administer the surveys, which allowed us to monitor the enumerators' work closely. This included monitoring the time they spent per survey, and sometimes even per survey module. Enumerators also had to record their location (via GPS) in the survey, so we could see their movements on a map. Lastly, the software also allows to conduct random audio-audits, which allows the researcher to "listen in" on certain portions of the survey. Ensuring data quality in an involved and time-intensive process that now all researchers can engage in. Unfortunately, the quality of the data is usually not discussed in most publications.

In addition to problems that might be caused by subpar work by the enumerators, the accuracy of interviews and the veracity of self-reported data can be questioned. Details on categories such as income, spending, wealth, personal possessions, duration of activities, etc. are essential components of RCTs and other methods. The majority of the poor in the world are subsistence farmers. Often, they cannot give accurate (monetary) information about these important variables.[24] In these cases, scientists like to use "proxy variables" or auxiliary variables. For example, livestock ownership is often used to measure the wealth of African subsistence farmers. However, these auxiliary variables are subject to high intertemporal fluctuation, and accurate ownership attributions are also difficult for large families. Survey answers can therefore be very sensitive to the context. Also, even seemingly trivial decisions by the research team, for example between using paper questionnaires or administering surveys on cell phones or laptops, can have measurable effects on results.[25] (See chapter 4 for a study on biases introduced to the data just by phrasing questions differently.) Similar problems of comparability of studies can arise from using different definitions of key concepts. The size of a household, an essential variable for determining wealth and poverty, may vary according to the definition (Beaman and Dillon, 2012). Common definitions of a household include "people cooking together" and "people living together in a compound".

It is also interesting to note that respondents in countries such as Ghana, Kenya or Zambia have probably already participated in multiple surveys on different topics due to the massive increase in surveys conducted: villagers know that research teams are coming and what questions they will ask - and that these surveys can be related to the distribution of aid or goods (even if this is often not the case in reality). They might adapt their answers, accordingly, providing responses they think the researchers want to hear ("social desirability bias"; Grimm, 2010).

[24] See Ravallion (2012) for a discussion of the value of self-reported data.

[25] An entire recent volume of the Journal of Development Economics (Vol. 98, 1, 2012) is dedicated to different forms of collecting data.

Heterogeneous effects

In general, RCTs are used to compare arithmetic means of the dependent variables of control and treatment groups, e.g. the mean of the test score of all children in a treatment group vs. the mean of a test score of all children in the control group (Banerjee and Duflo 2009; Deaton 2009). The average in medians, for example, cannot be recovered, but would be as useful (or even more useful) as the average in means. How, why, and to what extent the measure affects the units and subgroups (women, children, etc.) within a larger group remains hidden ("black box"). Larger sample sizes make it possible to compare subgroups, but it doesn't fully alleviate the problem. "As with all statistics – the evaluation of field experiments has implications for the mean of the population and may have little value in predicting individual behavior" (Banerjee and Duflo, 2017, p. 101).

Deaton (2009) warns against "disastrous" conclusions that could arise if highly heterogeneous effects within the groups can be expected. Thus, an increase in the mean could conceal a massive detrimental effect for a few participants (median < means): „Is LATE [Local Average Treatment Effect, A.N.] useful for the case in which a program has a positive average impact but causes a small share of people to suffer very negative consequences?" (Cohen and Easterly, 2009, p. 9). Similarly, a positive average program effect could be due to some very strong individual effects while at the same time the majority of the population is worse off (Deaton, 2010). The intervention is declared a success, although it causes great harm, which is a problem when experiments are being used as the basis for immediate policy advice.[26] It can even be assumed that the distributional effects on subgroups – in some cases – are more important for possible scale-up decisions than the arithmetic mean of the entire treatment and control group (Barrett and Carter, 2010).

Different population characteristics

The sample which is studied and surveyed will almost always be different than the population it is drawn from (Athey and Imbens, 2017; Deaton, 2010). Banerjee and Duflo (2017, p. 101) concede this point: "An obvious example of this is if an RCT finds a program has large impacts using a sample of poverty-stricken minority children, one cannot assume the program will have similar impacts on the universe of students in the United States." Thus, the RCT might not be representative of the region or country, especially if participants volunteered for a program (individuals selecting into treatment), or when the site was selected by the implementing organization is particularly prone to produce a larger effect size than any randomly selected site. This problem can

[26] One solution would be to select relevant subgroups ex-ante. If done ex-post, the researchers could be accused of "specification searching" (Banerjee and Duflo, 2009).

be pronounced in studies that use so-called "encouragement designs" as the population that takes up the treatment in such a design might be different that would take up the treatment absent of the encouragement. Similarly, impacts might be different in the real world, then they are when researchers monitor every step of the implementation process of a project (Barrett and Carter, 2014):

> As is true of any research method that pools data from distinct subpopulations, there is a nontrivial probability that no external population exists to whom the results of the experiment apply on average. Collecting the data experimentally does not solve this problem. If the inferential challenge largely revolves around essential heterogeneity rather than around endogeneity, experiments that address only the latter issue can at best claim to solve a problem of second-order relevance. (Barrett and Carter, 2014, p. 75).

General equilibrium effects

An education program might increase job prospects for the study sample, however, if scaled up to the entire population, might decrease the overall return to education and therefore overstate the program impacts (Mookherjee, 2005; Banerjee et al., 2017; Deaton and Cartwright, 2018). Programs also become costlier at scale in case they do work and might encounter political backlash that cannot be detected with smaller trials (Banerjee et al., 2017a). Indeed, Banerjee et al., (2017b) show political backlash for an anticorruption program to be implemented at scale in India and the scale-up was ultimately cancelled despite promising results at the pilot stage. Rothstein and von Wachter (2017) describe the risk of general equilibrium effects for a labor market intervention, where the subjects in the treatment group are incentivized to search for jobs (which in turn can reduce the chances of finding a job for the control group.[27]

Implementation challenges

In addition, organizations that partner with researchers for RCTs are often above average in management capacity. Once scaled-up, the government might run into motivational, budget or capacity constraints (Banerjee et al., 2017a). In smaller scale pilots, interventions can often be better monitored and controlled than when programs are implemented at scale and "gaming" might be more prevalent. Interventions also become costlier at scale, as more qualified nurses, teachers, or employees need to be hired (see e.g. Davis et al., 2017). These implementation challenges that lead to different forms of sampling bias (see e.g. Barrett and Carter, 2014). Thus, it is problematic to assume that a project that has shown a positive impact in a closely monitored RCT, has the same effects in the real world when it is run by the government without supervision of a research

[27] This example is also mentioned in Banerjee and Duflo (2017).

team (Cartwright, 2010). Most researchers that run RCT wear a "double-hat". On the one side they often partner with governments and want to provide actionable advice. On the other hand, they are driven by their publication record at their academic institutions. It is widely known that surprising or counterintuitive results have a higher chance of being published than those that are not ("publication bias"). It is also true that null-results are very hard to publish, so that there is an incentive to create "some" positive results with the data that was gathered, often running many specifications or through data mining.[28] Despite their claim to possess the "gold standard", RCTs also suffer, for example, from "p-hacking" as much as other studies and other fields of science (see for example Head et al., 2015).[29]

Time delays/non-linear impacts

RCTs often rely on one or more surveys to be conducted after the conclusion of the program. The timing of these surveys is crucial to measure the effects. If effects are non-linear, then the measurement might lead to wrong conclusions. Let's say, for example, an endline data collection is conducted 6 months after the conclusion of a project. If larger effects do not set in until then – they might come at a later time – the effect is found to be smaller than it would have been, when compared to a potential measurement at the peak of the effect, which could have materialized at a later point in time. An example might be an agricultural project, for example training on crop management. If the endline data collection comes before an important harvesting period, the researchers might conclude that the impacts were lower than if she had measured the effects after the harvest.[30] Likewise, if the follow-up survey is carried out too early, the interventions may be considered to have a benefit that is not sustainable (Woolcock, 2013). Data collection timing often depends more on feasibility, time and budget constraints on the side of the researcher. Anecdotal evidence, however, shows that e.g. job training programs can initially show a negative effect in the very short run and transition to positive effects later. The timing of surveys can therefore play a crucial role and ultimately maybe also have an effect on following policy decisions based on the

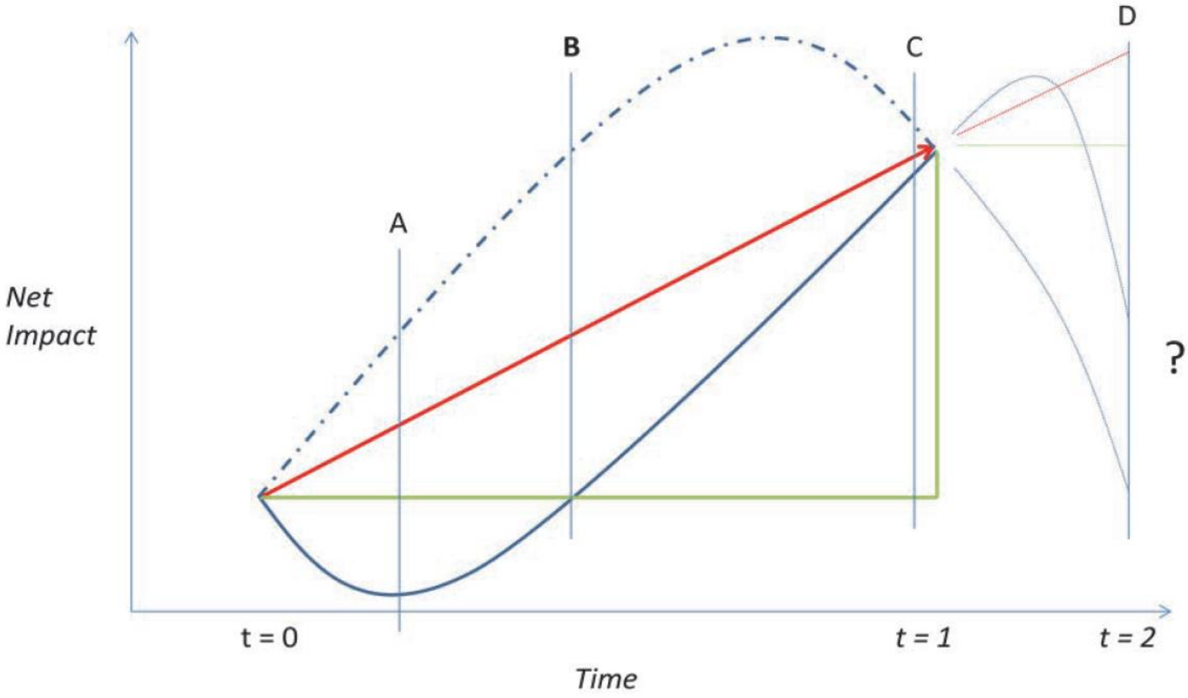
[28] "(…) Deaton (2010) expresses many concerns about the analyses and implementations of RCTs exploring heterogeneous treatment effects can be viewed as data mining and researchers should explore the implications of testing a large number of hypotheses in their studies; researchers rarely use appropriate standard errors when reporting results; exploring different combinations of baseline variables to include in regressions is another potential form of data mining; including baseline variables can lead to substantial biases in small samples; attrition from the study must be addressed; and it is not uncommon for RCTs to have implementation and operational issues that threaten the validity of the experiment" (Banerjee and Duflo, 2017, p. 101).

[29] Kaushik Basu (2005): "[The researchers] publish only what seems unexpected. Since the expected does not get published, we never get the larger global picture (…) and so think we have stumbled upon knowledge when, in fact, we have not." See also Earp and Trafimow (2015, p. 4) for some notes on this phenomenon in the field of social psychology.

[30] To capture effects that might be non-linear, my co-authors and I conducted an RCT in Nigeria with 7 follow-up rounds (Dunsch et al., 2017). However, cases with more than 2 survey rounds after the project are rare, mainly due to cost constraints.

studies.[31] Figure 1.4 illustrates this point graphically (Woolcock, 2013). Depending on when data is collected, project impacts might be assessed very differently.

Figure 1.4: Understanding Impact Trajectories (Woolcock, 2013)



Barrett and Carter (2010) summarize that the methodological elegance of RCTs is often counteracted by the variety of practical real-world problems:

> Problems arise, however, when pristine asymptotic properties confront the muddy realities of field applications, and strict control over fully exogenous assignment almost inevitably breaks down. (...) The end result is that the attractive asymptotic properties of RCTs often disappear in practice (...). (Barrett and Carter (2010, p. 522)

The list of potential problems underscores the importance for the implementers of such research to lay out good protocols from the beginning and ensure that these are followed throughout the lifetime of the project. It is important to note, however, that most of the listed points also apply to the quasi-experimental (or other) approaches to measuring efficacy described above.[32]

Despite these aforementioned issues, Banerjee and Duflo (2017) are still convinced:

[31] Kucklick (2012) argues that: „nobody has conducted a follow-up survey longer than 3 years after the intervention.“

[32] Al-Ubaydli and List (2013, p. 3) state: "In principle, generalizability requires no less of a leap of faith in conventional (non-experimental) empirical research than in experimental research. The issue is obfuscated in non-experimental research by the more pressing problem of identification: how to correctly estimate treatment effects in the absence of randomization."

> Still, with these important limitations in mind, the conventional wisdom is: if you can do a randomized field experiment, you should. (…) [I]f one designs the RCT in a way that helps validate a model of selection for observational data, then the only limitation appears to be the budget of the researcher. (Banerjee and Duflo, 2017, p. 101)

1.3.3 RCTs & Impact on development policymaking

The RCT critic Lant Pritchett points to an apparent contradiction: "[T]he randomization agenda as a methodological approach inherits in an enormous internal contradiction - that all empirical claims should be only when backed by evidence from randomization, excepting empirical claims about the impact of randomization on policy" (Pritchett, 2009, p. 162). Proponents of RCTs generally assume that it is sufficient to present "correct" or "true" assessments and outcomes regarding working policies. Politicians would take these recommendations for granted and implement policies accordingly. However, this fails to recognize the basic reality of political decision-making. It is neglected that policymakers (and it is debatable to what extent) are not just interested in the success of policies but also care for personal prestige or the survival of organizational structures. There might be only few practical benefits to conduct rigorous impact evaluations, but many risks. For example, it is very likely that many impact evaluation results would reduce donors' current high (self-reported) success rates.[33]

Lant Pritchett (2002) summarizes:

> If a program can already generate sufficient support to be adequately funded, then knowledge is a danger. No advocate would want to engage in research that potentially undermines support for his/her program. Endless, but less than compelling, controversy is preferred to knowing for sure the answer is "no".[34] (Pritchett, 2002, p. 268)

Seen in a positive light, it might also be that unrealistically high expectations lead to improved outcomes: „[I]t is possible that a combination of excessive subjective certainty among altruistic advocates and strategic maintained ignorance of true program effects is actually welfare improving" (Pritchett, 2002, p. 268).

Politicians operate under time pressure, decisions are taken in the absence of complete information, and politicians are accountable to their constituents in donor and recipient countries. So even if there is evidence to suggest that a particular program promises the greatest benefit on

[33] Banerjee and He (2008) provide a good overview oft he different evaluation practices of bi- and multilateral donors.

[34] To date, very few policies in Western countries are based on the results of RCTs, neither in the US nor in Europe. The phenomenal success of some Asian countries also is neither based on experiments nor vast flows of development aid (Sangmeister and Schönstedt, 2010).

average, it does not mean that it changes the political process. Political decision-making is by definition unscientific, although scientific results should of course influence the decision-making processes. Well-executed RCTs can provide clues for policy decisions. Nevertheless, the policymakers must still decide autonomously.

> A new drug might do better than a placebo in an RCT, yet a physician might be entirely correct in not prescribing it for a patient whose characteristics, according to the physician's theory of the disease, might lead her to suppose that the drug would be harmful. Similarly, if we find that dams in India do not reduce poverty on average (...) there is no implication about any specific dam, (...) yet it is always a specific dam that a policy maker has to decide about. (Deaton, 2010, p. 441)

These obvious realities about the political process is little understood by the *Randomistas*. Actual demand for "evidence" might be lower than expected.[35]

Although having a generally favorable opinion, William Easterly believes that many projects have impact mechanisms that are so obvious that they do not require an RCT to prove them (Easterly, 2009). RCTs would show successes first and foremost with simple intervention that require a steady routine, as for example vaccination campaigns (Woolcock, 2013). Similarly, de Souza Leão and Eyal (2019), referring to a seminal study by Miguel and Kremer (2004) on deworming question: "Do we really need an RCT to know that if children are less sick, they are more likely to go to school and less likely to get other kids sick?" Interventions that involve participants' own discretion and require a high level of involvement from implementers are very difficult to evaluate: "REs [randomized experiments, A.N.] can study incentives for teachers to show up to class, but not how well the teachers are doing the discretionary transaction-intensive job helping their students learn" (Easterly, 2009, p. 419). The same applies to interventions in health or education. It is easy to determine if bed nets have been distributed for malaria or textbooks. However, whether these are used correctly and thus provide welfare effects is much more difficult to determine (Dunsch, 2012).

Many important questions cannot be answered by means of IEs: "The big, philosophical questions such as whether development aid is fundamentally helpful or not, or what the root causes of global poverty are, cannot be answered" (de Souza Leão and Eyal, 2019, p. 404). Countries, for example, cannot be divided in half. Therefore, some phenomena that can be essential for development simply cannot be measured by IE methods (Imbens, 2010). This could be e.g. the interest rate or certain national laws. Thus, due to the high technical demands on the method (rigor), RCTs

[35] Also, until IE results see the light of day, many years can pass, which make them less useful for immediate policy-advice. The studies that are part of this dissertation were started in 2013 and took 4-6 years to be published.

inevitably must investigate micro-interventions (in most cases), which may be irrelevant to the overall development of a country (de Souza Leão and Eyal, 2019).[36] Barrett and Carter (2010) rightly recognize in the current expansion of RCTs the problem of not examining the most important issues, but especially those that can be investigated using RCTs. Therefore, RCTs are by definition mainly studying the symptoms of development problems rather than their root causes which might be at the macro level or otherwise not accessible to RCT studies. (Ogden, 2017, p. xxvii, calls this the "Trivial Significance Critique" of RCTs.) Critics mention that this turn towards smaller evaluations is also a mechanism for "randomistas to avoid the political debate about development by limiting themselves to testing behavioral hypotheses that often could be quite innocuous" (de Souza Leão and Eyal, 2019, p. 401). As a result, economists frequently conducting RCTs have been criticized to be searching only "under the street light" (ibid., p. 413).[37]

In this context, reference should also be made to the high costs of RCTs. Impact evaluations and or RCTs easily account for up to 10% of the project budget, whereas "conventional" evaluations usually don't impact the budget by more than 3%. If the total budget of the project is around 300,000 euros, the realization of a meaningful impact evaluation would therefore hardly be possible (Stockmann, 2010). Only larger budget envelopes currently allow for more extensive impact evaluations and economies of scale. However, if the reliability of large-scale data collection is not always clear and the cost remains high, other forms of evaluation, with greater involvement of the target groups, may allow for a better benefit-to-effort ratio.[38]

1.3.4 Problems of achieving external validity

Following the reasoning of the RCT proponents, experiments are the best way to build solid evidence, and to make forward-looking policy recommendations. A policy recommendation stemming from the results of an RCT can only be valid, if results hold or can be extrapolated into the future or across other, new contexts. However, even if internal validity can be established, there are several reasons why the claim to external validity or generalizability remains problematic. Dani Rodrik (2008) summarizes that "randomistas", however, are most interested in establishing internal validity, since without internal validity (the kind that only RCTs can produce), we need not start to worry about external validity. Internal validity is established when an experiment can

[36] Banerjee and Duflo (2017, p. 101) concede: "For instance, how much of the variance in achievement is explained by genetic endowment? Given we are not likely to alter genetics by means of a field experiment, if one is wed to RCTs then this question is unanswerable."

[37] In the words of Angus Deaton (2010, p. 429): "This goes beyond the old story of looking for an object where the light is strong enough to see; rather, we have at least some control over the light but choose to let it fall where it may and then proclaim that whatever it illuminates is what we were looking for all along."

[38] Peltzer (2012) for example recommends to use focus group discussions (FGD) more extensively. Nohr and Schmidt (2012) also argue for a methodological pluralism.

unequivocally determine the difference in outputs between the control group and the treatment group due to its flawless execution (data collection, data analysis). Deaton finds (in Ogden, 2017, p. 39): “Maybe I’m missing something, but my reading of the J-PAL webpage makes me think that when they list estimates, they seem to suggest that you can use them pretty much anywhere.”

External validity, however, does not automatically follow from internal validity. Peters et al. (2018) look at all RCTs published in major journals between 2009 and 2014 to see how these deal with threats to external validity. They find that the majority of the papers do not appropriately discuss the problems. Rothwell (2005, p. 82) discusses external validity in a widely cited paper for the medical field. He states: “Although what little systematic evidence we now have confirms that RCTs do often lack external validity, this issue is neglected by current researchers, medical journals, funding agencies, ethics committees, the pharmaceutical industry, and governmental regulators alike.” The important ongoing debate on the problems around external validity for RCTs in development, and whether it is possible to achieve it at all, is the focus of chapter 6.

1.4 References

Al-Ubaydli, O., & List, J. A. (2013). *On the generalizability of experimental results in economics: With a response to Camerer* (No. w19666). National Bureau of Economic Research.

Ashoff, G. (2010, July). Wirksamkeit als Legitimationsproblem und komplexe Herausforderung der Entwicklungspolitik. In *Wirksamere Entwicklungspolitik* (pp. 27-68). Nomos Verlagsgesellschaft mbH & Co. KG.

Andresen, H. (2012). *Staatlichkeit in Afrika: muss Entwicklungshilfe scheitern?*. Brandes & Apsel Verlag.

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, *24*(2), 3-30.

Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments* (Vol. 1, pp. 73-140). North-Holland.

Banerjee, A. V. (2005). 'New Development Economics' and the challenge to theory. *Economic and Political Weekly*, *40*(40), 4340-4344.

Banerjee, A. V. (2008, June). Big answers for big questions: the presumption of growth policy. In *Brookings Conference What Works in Development*.

Banerjee, A. V., & He, R. (2008). Making aid work. *Reinventing foreign aid*, 47-92.

Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. Annu. Rev. Econ., 1(1), 151-178.

Banerjee, A., Duflo, E., & Kremer, M. (2016, September). The influence of randomized controlled trials on development economics research and on development policy. In *The State of Economics, the State of the World Conference at the World Bank*.

Banerjee, A. & Duflo, E. (Eds.). (2017). *Handbook of Field Experiments* (Vol. 2). Elsevier.

Barrett, C. B., & Carter, M. R. (2010). The power and pitfalls of experiments in development economics: Some non-random reflections. *Applied economic perspectives and policy*, *32*(4), 515-548.

Barrett, C. B., & Carter, M. R. (2014). A retreat from radical skepticism: rebalancing theory, observational data, and randomization in development economics. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, 58-77.

Basu, K. (2014). Randomisation, causality and the role of reasoned intuition. *Oxford Development Studies*, *42*(4), 455-472.

Bauer, P. T. (1976). *Dissent on development*. Harvard University Press.

Beaman, L., & Dillon, A. (2012). Do household definitions matter in survey design? Results from a randomized survey experiment in Mali. *Journal of Development Economics*, *98*(1), 124-135.

Besley, T. (2012). Poor Choices: Poverty from the Ground Level. *Foreign Affairs*, *91*, 160.

BMZ - Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (2011). Chancen schaffen - Zukunft entwickeln. Bonn: BMZ.

Boone, P. (1994). The impact of foreign aid on savings and growth. London School of Economics. *CEP Working Paper*, *677*.

Boone, P. (1996). Politics and the effectiveness of foreign aid. *European economic review*, *40*(2), 289-329.

Burnside, C., & Dollar, D. (2000). Aid, policies, and growth. *American Economic Review*, *90*(4), 847-868.

Burwell, S., Muñoz, C., Holdren, J., & Krueger, A. (2013). Next steps in the evidence and innovation agenda (Memorandum to the Heads of Departments and Agencies). Washington, DC: Office of Management and Budget, Executive Office of the President. Retrieved July 15, 2013.

Cameron, D. B., Mishra, A., & Brown, A. N. (2016). The growth of impact evaluation for international development: how much have we learned?. *Journal of Development Effectiveness*, *8*(1), 1-21.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*(4), 409.

Caspari, A., & Barbu, R. (2010). Wirkungsevaluierungen: zum Stand der internationalen Diskussion und dessen Relevanz für die Evaluierung der deutschen Entwicklungszusammenarbeit. *BMZ Evaluation Working Papers*. Bonn: BMZ.

CGD – Center for Global Development (2006). When Will We Ever Learn? Improving Lives Through Impact Evaluation. (Report of the Evaluation Gap Working Group). Washington, DC.

Clemens, M. A., & Radelet, S. (2003). The Millennium Challenge Account: How much is too much, how long is long enough?. *Center for Global Development Working Paper*, (23).

Clemens, M. A., & Demombynes, G. (2011). When does rigorous impact evaluation make a difference? The case of the Millennium Villages. *Journal of Development Effectiveness*, *3*(3), 305-339.

Cohen, J., & Easterly, W. (Eds.). (2009). *What works in development?: Thinking big and thinking small*. Brookings Institution Press.

Dalgaard, C. J., Hansen, H., & Tarp, F. (2004). On the empirics of foreign aid and growth. *The Economic Journal*, *114*(496), F191-F216.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*(2), 424-55.

Degeval–Gesellschaft Für Evaluation, E. V. (2008). Standards für Evaluation.

Degeval–Gesellschaft Für Evaluation, E. V. (2010). Methoden der Evaluation. Positionspapier.

de Souza Leão, L., & Eyal, G. (2019). The rise of randomized controlled trials (RCTs) in international development in historical perspective. *Theory and Society*, 48(3), 383-418.

Druckman, J. N., Green, D. P., Kuklinski, J. H., & Lupia, A. (Eds.). (2011). *Cambridge handbook of experimental political science*. Cambridge University Press.

Duflo, E. (2006). Field experiments in development economics. *Econometric Society Monographs*, *42*, 322.

Duflo, E., & Kremer, M. (2005). Use of randomization in the evaluation of development effectiveness. *Evaluating development effectiveness*, *7*, 205-231.

Duflo, E., & Banerjee, A. (2011). *Poor economics*. Public Affairs.

Dunsch, F. (2012). *Conflicting Strategies to Enhance Foreign Aid Efficacy in Africa. The Millennium Villages, Randomized Trials and Free Trade.* Baden-Baden: Nomos.

Easterly, W. (2009). Can the West Save Africa? *Journal of Economic Literature,* 47, 373-447.

Easterly, W. (2006). *The white man's burden: why the West's efforts to aid the rest have done so much ill and so little good*. Penguin.

Easterly, W. (2008). *Reinventing foreign aid* (Vol. 1). The MIT Press.

Easterly, W., Levine, R., & Roodman, D. (2004). Aid, policies, and growth: comment. *American economic review*, *94*(3), 774-780.

Ebrahim-Zadeh, C. (2003). Dutch Disease: Too much wealth managed unwisely. *Finance & Development*, *40*(1), 50-50.

Faas, T., & Huber, S. (2010). Experimente in der Politikwissenschaft: Vom Mauerblümchen zum Mainstream. *Politische Vierteljahresschrift*, *51*(4), 721-749.

Faust, J., & Leiderer, S. (2008). Zur Effektivität und politischen Ökonomie der Entwicklungszusammenarbeit. *Politische Vierteljahresschrift*, 129-152.

Faust, J. (Ed.). (2010). *Wirksamere Entwicklungspolitik: Befunde, Reformen, Instrumente*. Nomos.

Gertler, P., & Boyce, S. (2001). An experiment in incentive-based welfare: The impact of PROGRESA on health in Mexico. *University of California, Berkeley*, 30-37.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice*. The World Bank.

Grimm, P. (2010). Social desirability bias. *Wiley international encyclopedia of marketing*.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic literature*, *48*(2), 399-423.

Hasenclever, A., Mayer, P., & Rittberger, V. (1997). *Theories of international regimes* (Vol. 55). Cambridge university press.

Hegemann, H., Heller, R., & Kahl, M. (Eds.). (2012). *Studying 'Effectiveness' in International Relations: A guide for students and scholars*. Verlag Barbara Budrich.

Jakobeit, C. (2009). Warum hilft die Hilfe nicht?. *GWP–Gesellschaft. Wirtschaft. Politik*, *58*(4).

Johnson, D. (2011). *Afrika vor dem großen Sprung*. Wagenbach.

Karlan, D., & Appel, J. (2011). More than good intentions. *Dutton, New York*.

Kristof, N. (2011). Getting smart on aid. *New York Times*, 18.05.11: A27.

Kucklick, C. (2012). Viel hilft viel. Oder nicht. *Geo. Die Welt mit anderen Augen sehen*, *5*, 98-112.

Lee, B. (2000). Theories of evaluation. In Evaluationsforschung (pp. 127-164). VS Verlag für Sozialwissenschaften, Wiesbaden.

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives*, *25*(3), 17-38.

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

Mertens, D. M. (2000). Institutionalizing evaluation in the United States of America. In *Evaluationsforschung* (pp. 41-56). VS Verlag für Sozialwissenschaften, Wiesbaden.

Miguel, E., & Kremer, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159-217.

Miles, E. L., Andresen, S., Carlin, E. M., Skjærseth, J. B., Underdal, A., & Wettestad, J. (2001). *Environmental regime effectiveness: confronting theory with evidence*. MIT Press.

Mookherjee, D. (2005). Is there too little theory in development economics today?. *Economic and Political Weekly*, *40*(40), 4328-4333.

Mosley, P. (1987). *Overseas aid: its defence and reform*. Wheatsheaf Books.

Moyo, D. (2009). *Dead aid: Why aid is not working and how there is a better way for Africa*. Macmillan.

Neudeck, R. (2010). *Die Kraft Afrikas: warum der Kontinent noch nicht verloren ist*. CH Beck.

Nohr, S. & Schmidt, M. (2012). Experimentelle Evaluierungsmethoden im Praxistest: Mehrwert, aber keine Allzweckwaffe. *KfW Entwicklungspolitik Kompakt*, *9.*

Nuscheler, F. (2008). INEF-Report 93/2008. *Die umstrittene Wirksamkeit der Entwicklungszusammenarbeit, Universität Duisburg-Essen (Institut für Entwicklung und Frieden)*.

Oberthür, S., & Stokke, O. S. (Eds.). (2011). *Managing institutional complexity: regime interplay and global environmental change*. MIT Press.

Olken, B. (2009). Comment on "The Policy Irrelevance of the Economics of Education: Is 'Normative as Positive' Just Useless, or Worse?" by Lant Pritchett. *What works in Development*, 165-9.

Paffenholz, T. (2006). *Promotion de la paix et coopération internationale: histoire, concept et pratique* (No. 25-2, pp. 19-45). Institut de hautes études internationales et du développement.

OECD. (2002), Evaluation and Aid Effectiveness No. 6 - Glossary of Key Terms in Evaluation and Results Based Management.

OECD. (2005). Paris declaration on aid effectiveness.

Ogden, T. N. (Ed.). (2017). *Experimental Conversations: Perspectives on Randomized Trials in Development Economics*. MIT Press.

Parker, I. (2010). The Poverty Lab: Transforming development economics, one experiment at a time. *New Yorker*, *17*, 79-89.

Pearson, F., Lounsbery, M. O., & Talentino, A. K. (2012). How Effective is International Military Intervention? The Evolution of Motives, Forms and Outcomes. *Studying''Effectiveness' in International Relations: A guide for students and scholars*, 123.

Peltzer, R. (2012). Reden macht klüger als zählen. *Welt-Sichten*, 7, 37-40.

Platz, T. (2012). Evaluationen in der Entwicklungszusammenarbeit von Nichtregierungsorganisationen. Unpublished. Institut für Politikwissenschaft, Universität Hamburg.

Pritchett, L. (2002). It pays to be ignorant: a simple political economy of rigorous program evaluation. *The Journal of Policy Reform*, *5*(4), 251-269.

Pritchett, L. (2009). The policy irrelevance of the economics of education: is 'normative as positive' just useless, or worse. *What works in Development*, 130-73.

Quack. M, & Sprenger. D. - Arbeitskreis Evaluation von Entwicklungspolitik DeGEval-Gesellschaft für Evaluation. (2010). *Verfahren der Wirkungsanalyse: ein Handbuch für die entwicklungspolitische Praxis*. ABI, Arnold-Bergstraesser-Inst.

Radelet, S., & Levine, R. (2008). Can we build a better mousetrap? Three new institutions designed to improve aid effectiveness. *Reinventing foreign aid*, 431-460.

Rasmussen, O. D., Malchow-Møller, N., & Barnebeck Andersen, T. (2011). Walking the talk: the need for a trial registry for development interventions. *Journal of Development Effectiveness*, *3*(4), 502-519.

Ravallion, M. (2012). *Poor, or just feeling poor? On using subjective data in measuring poverty*. The World Bank.

Rodrik, D. (2008). The new development economics: we shall experiment, but how shall we learn? Harvard Working Paper.

Roodman, D. (2007). Macro aid effectiveness research: A guide for the perplexed. *Available at SSRN 1101461*.

Rostow, W. W. (1959). The stages of economic growth. *The economic history review*, *12*(1), 1-16.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *The Lancet*, *365*(9453), 82-93.

Sachs, J. (2005). *The end of poverty: How we can make it happen in our lifetime*. Penguin UK.

Sachs, J. (2008). *Common wealth: Economics for a crowded planet*. Penguin.

Sachs, J., McArthur, J. W., Schmidt-Traub, G., Kruk, M., Bahadur, C., Faye, M., & McCord, G. (2004). Ending Africa's poverty trap. *Brookings papers on economic activity*, *2004*(1), 117-240.

Savedoff, W. (2013, July 23). Hot Topic, Cool Heads: Impact Evaluation Debated at CGD-3ie Conference. Blog post. Retrieved September 30, 2019, from https://www.cgdev.org/blog/hot-topic-cool-heads-impact-evaluation-debated-cgd-3ie-conference

Seitz, V. (2014). *Afrika wird armregiert oder Wie man Afrika wirklich helfen kann*. Deutscher Taschenbuch Verlag.

Sangmeister, H., Günther, D. V. E., Hilser, K., Schönstedt, D. V. A., Dokumentarfilm, E., Rückert, J., ... & Meng, J. (2010). *Entwicklungszusammenarbeit im 21. Jahrhundert*. Baden-Baden, Nomos Verlagsgesellschaft.

Stockmann, R. & Meyer, W. (2009). Evaluation. Eine Einführung. Stuttgart: UTB.

Stockmann, R., Menzel, U., & Nuscheler, F. (2015). *Entwicklungspolitik: Theorien–Probleme–Strategien*. Walter de Gruyter GmbH & Co KG.

Stockmann, R. (2010). 10 Thesen zur Evaluation der Wirksamkeit der Entwicklungszusammenarbeit. CEval Arbeitspapier 18, Saarbrücken: Centrum für Evaluation.

The Prize in Economic Sciences. (2019). NobelPrize.org. Nobel Media AB 2019. Mon. 14 Oct 2019. Retrieved October 20, 2019, from https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/

Wolff, J. H. (2005). *Entwicklungshilfe: ein hilfreiches Gewerbe?: Versuch einer Bilanz* (Vol. 18). Mohr Siebeck.

World Bank. (2009). Annual Review of Development Effectiveness - Achieving Sustainable Development. Washington, DC: The World Bank.

World Bank. (2011). IDA16: Delivering Development Results. Washington, DC: The World Bank.

World Bank. (2019). World Bank Open Data. Retrieved September 20, 2019, from https://data.worldbank.org/

2. The Nollywood nudge: An entertaining approach to saving[39]

2.1 Abstract

Can edu-entertainment be an effective tool to strengthen financial inclusion? In collaboration with a local NGO (*Credit Awareness*) and a Microfinance Bank (*Accion*) we explore the short- and medium-term savings decisions of a group of micro-entrepreneurs in Lagos, Nigeria by inviting business owners to one of four randomly allocated events: (i) A movie screening of *The Story of Gold* - a Nollywood[40] film encouraging entrepreneurs to save responsibly; (ii) an event where business owners are shown a "placebo" screening of a movie with no financial education content and offered "on-the-spot" microsavings accounts through *Accion*; (iii) a combined event, screening *The Story of Gold* and offering "on-the-spot" accounts; and (iv) a screening of the placebo film only as our control group. We find that entrepreneurs watching *The Story of Gold* were 5 percentage points more likely to open a savings account on the spot than those in placebo screenings, and this effect was mostly driven by male business owners. In contrast, less than 1% of entrepreneurs who were not offered "on-the-spot" signed up for a savings account after the screening. In the longer run, only moderate changes in attitudes and perceptions were found, while savings and borrowing behavior was unchanged four months after the screening. This suggests that, while influencing short-term decisions is possible, longer-run behavior is far less malleable through once-off events. This paper contributes to the literature by directly testing the importance of linking emotional stimulus to financial messages to influence short-term savings decisions and identifying the important interaction between emotional stimulus and the opportunity to act on this stimulus.

2.2 Background

Traditional rational-agent economic models rely on the assumption that people make decisions based on a rational and deliberate consideration of all costs and benefits associated with the action, conditional on available knowledge. However, low-income individuals regularly make seemingly sub-optimal financial decisions and there are strong correlations between financial knowledge, sound financial decisions and use of financial products (e.g. Hilger et al., 2003). This has led to a growing body of literature exploring the importance of providing financial education and training to individuals and entrepreneurs to effectively improve knowledge, leading to

[39] This study was published as a World Bank Policy Research Paper (Coville et al., 2019). Co-authors are Aidan Coville, Vincenzo Di Maro, and Siegfried Zottel.

[40] The Nigerian version of Hollywood.

improved financial capabilities and decisions. Despite strong correlations (e.g. Lusardi, 2007), rigorous causal impact evaluations of financial literacy training programs have shown mixed results, often with little to no effect on actual behavior (e.g. Cole et al., 2011) or showing positive impacts only through resource-intensive interventions (e.g. Bruhn et al., 2013). These limited effects could be explained by (i) only small increases in actual knowledge; or (ii) the fact that people do not fully apply this knowledge when making financial decisions such as when and how much to save. Evidence from psychology and behavioral economics highlights the fact that people act within "bounded rationality", often relying on heuristics to simplify their choices. Kahneman (2003) presents a framework that differentiates between two states that drive human decision making: intuition and reasoning. Decisions based on intuition are "fast, automatic, effortless, and often emotionally charged", whereas reasoning is "slower, effortful, and deliberately controlled". He argues that most decisions are based on intuition, where reasoning acts as a safeguard, rather than motivator, of many behaviors. This insight has important potential implications on how best to influence financial behavior. Even when people are fully aware of the most appropriate action to take, cognitive biases and heuristics may prevent this knowledge from translating into action. Thus, the traditional causal framework linking improved financial knowledge to changes in awareness, perceptions, attitudes and behavior, may underestimate important psychological barriers to financial inclusion that weaken the suggested causal chain. Acknowledgement that we base many decisions on heuristics rather than full information helps to explain why, for instance, "rule-of-thumb" approaches to financial education can be more effective at changing behavior than teaching more detailed accounting principles (Drexler et al., 2012).

This evaluation explores the effectiveness of mass- and social media delivering financial messages in order to induce behavior change beneficial to recipients. Specifically, building on the evidence that emotions and heuristics are likely to influence decisions, this study explores the effectiveness of using a Nollywood movie, *The Story of Gold*, to relay a simple message of "safe saving and responsible borrowing" through an emotionally-charged story line to a group of 2938 microentrepreneurs in Lagos, Nigeria. By intertwining the main message of responsible financial behavior into an accessible, entertaining and relatable story about twin sisters trying to succeed in business, the movie appeals to emotion, without providing specific information related to common measures of financial literacy such as understanding of interest rates and inflation. The underlying assumption is that a movie loses its entertainment value when people start explaining how to calculate risk adjusted returns to investments.

The Story of Gold is a once-off event aiming to influence transient emotions and lower transaction costs. However, responsible saving is a long-term commitment requiring continued and deliberate effort. The objective of the study was to identify whether this once-off event could spur action (in our case, opening a microsavings accounts) and serve as a catalyst to build financial capabilities through direct and continued exposure to financial institutions and products. The theory of behavioral consistency - where actions based on transient emotions have been identified to influence later decisions derived from people's desire to be consistent with previous actions - justifies the possible effectiveness of this "foot-in-the-door" hypothesis, but there is limited evidence on how this might influence savings behavior in particular.[41] Hence, shedding some light on whether and how interventions that work through affecting perception and emotions in the short term can produce change in behavior and commitment in the longer term is an important empirical topic.

The study uses a 2x2 randomized factorial design to exogenously vary (i) exposure to *The Story of Gold* and (ii) access to financial products by offering free "on-the-spot" microsavings accounts through a MFB at selected screening events. Through this framework we are able to test the relative effectiveness of (1) using "edutainment"[42] to motivate action, (2) reducing access constraints to financial products, and (3) the interaction of these two.

We find that entrepreneurs in all three treatment arms increase self-reported trust in MFBs, but the treatment arms including *The Story of Gold* had a larger effect on male self-reported trust. The combination of the movie with the presence of an MFB to help facilitate the opening of a savings account (at the time of the screening) was substantially more effective at motivating business owners to open an account than the presence of an MFB combined with a placebo screening, and this was most effective for influencing male decisions, increasing savings account sign up rates from 1% to 11%. Four months after the event we find limited or no sustained impacts on perceptions of MFBs and intention to borrow and save, and no effect on the likelihood of having a savings account (we find that many of the business owners that opened an account at the screening already had a savings account, resulting in this null effect).

This suggests that, even with relatively low-budget productions, it is possible to use entertainment to motivate action in the short term but long-term behavior is less malleable.[43] Furthermore, having a direct opportunity to act in the moment may significantly increase the impact of edutainment activities that influence transient emotions. However, care needs to be taken when developing the choice architecture designed to nudge people towards more "optimal" financial decisions, as this may induce unexpected behavior leading to further sub-optimal outcomes.

[41] More generally, this can be related to the "path dependence" principle in economics and sociology (Pierson, 2000).

[42] That is education through entertainment.

[43] This could indicate that commitment savings account might be necessary to solidify longer term behavior.

The rest of the paper is structured as follows: in Section 2.3 we explain our rationale to test edu-entertainment – in contrast to more standard financial education programs – as a means to change savings behavior. In Section 2.4 we describe the interventions and section 2.5 and 2.6 provide an overview of the identification strategy, sampling, baseline balance and attrition. Section 2.7 presents the econometric framework for analysis. Section 2.8 presents results, with robustness checks included in Section 2.9. We provide a discussion and conclude in Section 2.10.

2.3 A "nudge" for better savings outcomes?

This section explains the reasoning behind this paper's approach to test entertainment media to nudge savings behavior. It first presents the state of poor financial literacy and access to finance in Nigeria (2.3.1). We then show (2.3.2) that traditional financial education programs have mostly failed to deliver results to ameliorate this condition. We argue in (2.3.3) that the psychological biases might partly cause this inefficient savings behavior, and that they cannot be overcome by learning about the right way to do things alone. We show how to make existing biases work in favor of sound financial decision making, "work[ing] around human nature to help people save as they aspire to" (Karlan et al., 2013).[44] We then present how edu-entertainment has previously been used to aim at these biases to transform behavior (2.3.4). Lastly, we briefly describe *Nollywood* and its potential to serve as a vehicle to spread messages broadly (2.3.5).

2.3.1 Financial Literacy & Access to Finance in Nigeria

Although improvements have been registered in the last 3 years, 46% of the Nigerian population remains financially excluded, with no access to formal or informal financial services[45]. This compares unfavorably to countries such as Kenya and Botswana (33%), while in South Africa only one quarter of the population is financially excluded. Only 25% of people have a formal savings account, excluding 66 million adults (Makanjee and Ladipo, 2011; EFina, 2012). The use of Microfinance bank (MFB) accounts is even less widespread with only 4.6% of the adult population having a savings account with an MFB. This lack of access is not derived from a lack of interest or demand. According to recent survey results, almost 75% of the unbanked population in Nigeria report that they would like to have a bank account and over 80% of the population receives financial advice from family and friends. In theory, saving helps individuals and businesses by enabling consumption smoothing for volatile incomes, serving as insurance for the poor, growing investments, and allowing better access to micro-finance (e.g. Deaton, 1989; Karlan et al., 2013).

[44] See for example Datta and Mullainathan, 2003, and Sunstein and Thaler, 2003, for discussions on "libertarian paternalism".

[45] Results presented here are based on a recent nationally representative survey of 20,000 consumers conducted by EFinA in 2010.

However, "(…) very few people possess the extensive financial knowledge conducive to making and executing complex plans." (Lusardi and Mitchell, 2013). But knowledge and acting on knowledge are two different concepts and individuals often make poor financial decisions even when better options are readily available (Willis, 2011; Pathak et al., 2011), and even when they express the desire to act differently (Thaler and Benartzi, 2004). Building financial capacity in Nigeria represents a big step in helping consumers to acquire the skills and knowledge to be capable, confident, and self-reliant when making financial decisions. Evidence on the best way to build this capacity is, however, lacking. It is within this context the World Bank has worked closely with the Central Bank of Nigeria to develop and implement the World Bank-funded Micro, Small and Medium Enterprises (MSME) project[46] to test innovative consumer education programs such as the one evaluated here.

2.3.2 Financial education and business training programs

In order to improve financial decision-making, a common strategy is to offer financial or business training. Evidence on the impact of these programs is mixed. While financial literacy is correlated with household well-being (Mulaj and Jack, 2012) and less financial decision-making errors (Lusardi and Tufano, 2009; Stango and Zinman, 2009) research results do not fully support a causal chain leading from financial education to higher financial literacy, and subsequently improved behavior (Duflo and Saez, 2003; Willis, 2011).[47] Financial literacy may therefore be a secondary or even tertiary determinant of individual financial behavior (Cole and Fernando, 2008). Intensity, exposure, quality, and training content also vary widely (Drexler et al., 2012). Willis (2011, p. 431) argues that effective financial education would need to be “extensive, intensive, frequent, mandatory, and provided at the point of decision-making, in a one-on-one setting, with the content personalized for each consumer.” Also, participation levels for voluntary financial education programs are “extremely low”, even for very short courses (Bruhn et al., 2013). This presents concern for the power of the analysis, but, more broadly, not attending the courses might be an expression of economically optimal behavior by the potential recipient, reflecting the poor perceived efficacy of these programs.[48] The poor results of traditional education programs made us

[46] The project financed the production of the film, but Credit Awareness was responsible for both overseeing this production and the subsequent rollout.

[47] An increase in knowledge does not necessarily change attitudes and habits, also among more educated populations (Thaler and Bernartzi, 2004).

[48] In their literature review, McKenzie and Woodruff (2012) conclude that many impact evaluations of training programs are inconclusive due to technical shortcomings such as heterogeneity in length, content and types of firms participating. Many studies are underpowered, with hurried follow-up surveys (within one year of the training) covering small sample sizes, making it difficult (or impossible) to detect long-term effects. They also suffer from attrition and measurement problems of relevant business indicators.

think about alternative interventions such as making use of existing behavioral biases to change detrimental behavior.

2.3.3 Bounded rationality

A large body of literature from the fields of psychology and behavioral economics attempts to shed light on the fact that individuals often make irrational decisions or “mistakes” (being limited by "bounded rationality"), even when they know better. To present a framework of this bounded rationality, Kahneman (2003) introduces the “Architecture of Cognition”, distinguishing two models of thinking and deciding, broadly (and metaphorically) summarized as *intuition* – “System 1” – and *reasoning* – “System 2”:

> The operations of System 1 are fast, automatic, effortless, associative, and often emotionally charged; they are also governed by habit, and therefore difficult to control or modify. The operations of System 2 are slower, serial, effortful, and deliberately controlled; they are also relatively flexible and potentially rule-governed (…). (Kahneman, 2003, p. 1451-2)

The two systems can provide crucial insights on how to influence financial decision-making. If System 1 mainly drives financial behavior (intuition), models aiming to affect behavior through System 2 (reasoning) such as information campaigns or business training, assuming a “rational agent of economic theory” (Kahneman, 2003), might prove to be ineffective (which is supported by some evidence, see e.g. Cole et al., 2007).[49]

2.3.3.1 Accessing System 1

References (such as expectations, emotional and motivational arousal and other phenomena) can increase the accessibility of thoughts which are important for decision-making (Andrade and Ariely, 2009). Loewenstein and Lerner (2003) argue that even small "primers" can influence behavior, even when this "priming" is unnoticeable by the stimulated individual.[50] In the field of marketing, Bertrand et al. (2010) for example find that “persuasive” advertising can play a significant role in decision making, even if the content of the advertising is not directly related to the product being sold. There are different kinds of references applicable to our setting:

[49] Kahneman e.g. argues that the assumption that deciders evaluate outcomes by the utility of final asset positions is "easily" proven to be wrong.

[50] Willis (2011, p. 430): “Decisions can be strongly affected by even transitory emotions related to nothing more than the weather.”

2.3.3.2 The “affect heuristic”

People tend to base decisions that are being taken *now* on past decisions (unconsciously), shortcutting the thought-intensive System 2 process of deliberately evaluating the pros and cons of the respective decision at hand. They also base decisions on whether they *like* something, rather than carefully evaluating benefits and disadvantages (Slavic et al., 2003), answering a difficult question (What are the pros and cons?) by answering the easier question instead (How do I feel about it?) – a cognitive shortcut, where intuition (which resembles perception) acts as a substitute for reasoning (Kahneman, 2003). Advertising professionals often make use of these phenomena by focusing on conveying a good feeling of their product to their audience rather than stressing the beneficial effects of a purchase.

2.3.3.3 Behavioral consistency

Another important heuristic is the tendency to behave consistently with previous decision-making (Cialdini et al., 1995). Although the incidental effect of emotions might be short-lived, the influence of mild incidental emotions can live longer than the emotional experience itself (Andrade and Ariely, 2009). Goldberg et al. (1999) for example, illustrate the effects of an anger-inducing film on subsequent – unrelated – actions. Decisions based on a short-lived incidental emotion can develop the foundation for future choices and hence outlive the original cause (the emotion) for the behavior (Andrade and Ariely, 2009). Retrospectively, people tend to identify their past choice as an expression of their past preference (Schwarz and Clore, 1983), whereas in reality thoughts and actions are rather intuitive most of the time (as argued in Kahneman, 2003). In this manner, initial emotions serve as an “anchor” for later decisions (Tversky and Kahneman 1974), reinforcing behavioral consistency.[51] Similarly, hypothetical commitment carries over to real decisions if they are presented later (Ariely et al., 2003).[52]

2.3.3.4 Knowledge and trust

An initial reference or action can have longer lasting effects by fostering cooperative behavior based on knowledge and trust in the institution generated through repeated interaction (Mailath and Samuelson 2011). Once the initial burden of interacting in a new environment is overcome, subsequent interactions might become easier, as benefits become more salient. Following this rationale, exposure to media that induces emotions can trigger an initial action, providing a "foot-in-the-door”, which may influence later actions (Freedman and Fraser, 1966).

[51] The so-called "sunk-cost-fallacy" or the "endowment effect" are related concepts: People have a hard time to correct previous actions by realizing financial losses, consequentially making things worse (Thaler, 1980; Arkes and Blumer, 1985).

[52] Other relevant studies on past decisions affecting the present: Ottati and Isbell, 1996; Pocheptsova and Novemsky, 2010

2.3.3.5 Emotions and decisions: Gender differentials

A sizeable body of research looks into the question of whether emotions show differential gender effects on risk preferences, social preferences, and competitive preferences. Harshman and Paivio (1987) review evidence on studies showing that women experience emotions more strongly than men. Women are often more risk averse (Sunden and Surette, 1998; Croson and Gneezy, 2009) and tend to save more conservatively than men (Hinz et al., 1997).[53] However, Finucane et al. (2000) find gender differences only for whites ("white male effect"), which hints at cultural biases causing gender differences. Brought together, the literature suggests that gender differentials tend to be context (and culture) specific with few clear and unambiguous traits across population groups and activities.

2.3.4 Edutainment & behavior change

Drawing from the abovementioned studies and findings, the question arises whether (i) commercial entertainment media could be used to combine information (education) delivery with (ii) behavioral treatment arms, such as nudges, varying choice architecture, and/or emotional stimulation. Could combining the two perhaps help improve literacy levels and at the same time overcome some of the psychological barriers that stimulate bad behavior? While commercial media has for a long time been associated with effective changes in social behavior (both positive and negative) it has rarely been used in the field of finance. In other sectors, such as health and education, these tools have been used with success for a long time. For instance, as Brazil's Rede Globo network grew through the 1970s and 1980s, women also began having fewer children, experiencing the same decrease in fertility as with two extra years of education (La Ferrara et al., 2012).

While using mass media to transmit educational messages is not a novel approach, using edutainment to improve financial capabilities is less explored. The telenovela "Nuestro Barrio" is a prominent example from the U.S. aimed at Hispanic immigrants, where research found that it successfully conveyed the importance of formal bank accounts to the largely under-banked community (Spader et al., 2007). Most recently, a World Bank supported study evaluated the impact of a South African soap opera with financial messages ("Scandal!"). The study made use of an encouragement design to compare outcomes between a randomly selected group that watched Scandal and another group that watched a "placebo" show without financial education content. Watching Scandal resulted in higher financial knowledge scores, increased borrowing from formal sources, and decreased the likelihood of entering into hire purchase agreements (Berg and Zia, 2013). In

[53] Inability to determine who makes the financial decisions in a household is a potential problem for the validity of these results.

Ethiopia, a study showed that simple documentaries of relatively successful individuals from the same region affected both viewers' investment in their children's education and other future-oriented behaviors (Bernard et al., 2015; see also Bernard et al., 2014).

Edutainment, as an alternative to more formal classroom learning has the potential to be distributed more widely at lower marginal costs, and may appeal to a broader base, reaching out to people that may not otherwise be interested in in the topic. By creating emotional connections to the characters and the storyline, the process is believed to help internalize and operationalize the learning. Since this is a relatively new approach in the field of finance, there is a need for rigorous evaluation of these programs to assess the extent to which entertainment media is indeed effective in changing individuals' financial behavior. In particular, one question is about the role of edutainment through a once-off event (as is the case for *The Story of Gold*) as opposed to continued exposure to the message (like in the case of the soap operas mentioned above) that could make the emotional connections much stronger.

2.3.5 Nollywood

Movies from the Nigerian film industry penetrate almost all households in Nigeria – and across much of Africa, making them the ideal platform to deliver edutainment content. Although producing relatively low budget films, Nollywood is now the second largest movie industry in the world in terms of productions, only trailing India's Bollywood with an output of about 200 films every month. The industry is also the second largest employer in Nigeria, after the government. Films are largely made for home consumption rather than for the bigger cinema screenings. The stories told put fundamental human emotions and strong narratives front and center: Love, hate, envy, upward mobility, urban culture, and witchcraft. Due to their ubiquity, movies have the potential to reach large audiences with ease, surpassing traditional ways of conveying messages. Even politicians have understood the potential of these movies, posing with their stars at rallies and events. President Jonathan recently announced to support a N3 billion facility to support the Nigerian movie industry (Vanguard, 2013). With financial and political backing, together with large demand, Nollywood provides a unique opportunity to disseminate knowledge and build a culture of responsible financial decision-making, reaching out to the otherwise marginalized communities.

2.3.6 Application

Under the assumption that System 1 is a driver of many financial decisions and accessibility and "narrow framing" (Kahneman and Lovallo, 1993) and references are indeed important, *the Story* *of Gold* was developed to place more weight on intuition than reasoning to influence decision making.[54]

The movie seeks to address System 1 in order to encourage behavior change by promoting the take up and use of savings accounts in the short-term and encourage sustained use by building experience (offering a foot in the door) and promoting longer-term behavioral consistency with the original action. Thus, while the Nollywood movie could possibly also augment knowledge and awareness that in turn leads to better reasoning, the main intention of using the movie is to target business owners' intuitive behavior by influencing emotions, making relevant thoughts more accessible, especially when coupled with the immediate availability to sign up for savings accounts after the screening (reduction of transaction costs).

2.4 Description of the intervention

The Story of Gold is a feature-length Nollywood movie produced and distributed by Credit Awareness, a local NGO promoting "safe savings and responsible borrowing". It tells the story of identical twin sisters in Nigeria. Although identical in appearance, the decisions they make when faced with different financial choices affect their lives as well as those around them and ultimately lead them down different paths, one making sound financial decisions and succeeding in business and the other falling into a debt trap. The movie aims to impress upon low income individuals with limited formal education the importance of saving with a formal financial institution and borrowing responsibly. Focusing on this simple message and highlighting the repercussions of poor financial decisions, *The Story of Gold* focuses on the heuristic and emotional elements of human decisions to promote a stronger savings culture, facilitated by Credit Awareness. A partner microfinance bank (in this case *Accion*) participated in selected screening events and briefly presented their main savings and borrowing products after the show. They then provided all the necessary paperwork for participants to open a "Brighta Purse" business savings account on the spot if they were interested in doing so. The micro-savings account is geared towards micro-entrepreneurs as an entry savings and transaction account, requiring no initiation fees (although a minimum balance of 500 Naira is needed - one third of average daily profits from our sample of entrepreneurs). Interest on this savings account is then a function of the amount of savings held. If entrepreneurs expressed interest in opening an account but did not have the opening balance on hand, they could sign up their names and contact details and follow up with *Accion* at a later date to confirm the account opening. In this case, the combined intervention aimed at simultaneously encouraging people to save through the movie's message while reducing access barriers

[54] Kahneman (2003) stresses the point that preferences of System 1 are shaped by emotions of the moment and need not be internally coherent or reasonable. The preferences of System 1 and 2 therefore do not have to be consistent.

almost to zero with the presence of the MFB at the screening events. The hypothesis was that the movie would serve to inform, but also motivate business owners to act, and open a new savings account. The motivational effect of the movie was expected to wear off soon after the screening and giving business owners the opportunity to act in the moment, may increase the potential for this short-term motivation to translate into action. By overcoming these barriers to formal financial participation, the study could then explore whether this engagement resulted in longer-term interactions, leading to improved use of financial products over time.

While Credit Awareness plans to roll out the screening events across the country, the evaluation focused on a series of early pilot screenings to test the modality and learn before scale up. The pilot screenings were conducted at local community halls in the Ikotun region of Lagos – home to a sprawling street market. The typical screening event would be held in a hall, with local traders invited to attend. The event lasted approximately 3 hours, starting with a brief introduction, the screening of the movie and an open discussion after the event to reflect on the story's core messages. This would be followed by the engagement with the MFB. For the purpose of the evaluation two extra elements were included to the standard Credit Awareness model: (i) to ensure compliance with the assignment strategy each participant received a personalized invitation with a photograph to confirm their identity; and (ii) to improve participation rates, a lottery was held at the end of the event where participants could win spot prizes.

2.5 Sampling and identification strategy

Two community halls large enough to hold 200 people were identified in the Ikotun area of Lagos. A radius of 2 kilometers was used to set the boundaries to ensure that all participants could easily access the halls without needing to use public transport. A census of the area was then taken in July 2012, together with a short baseline listing questionnaire used to stratify the sample on whether they had a savings account, whether they kept financial records and if their store was in the main (official) market area, or in the surrounding Lagos streets. In total 2938 micro-entrepreneurs were recorded with geo-positioning and photographs to confirm identity in follow up interactions and verify intervention compliance (see Annex for an example of the invitation created from this information to verify identity at the event). The criterion used for selection into the sample was being the owner/operator of a business operating within the study area. These businesses were then randomized into one of 5 groups: (i) pure control [PC]; (ii) placebo screening [C]; (iii) Story of Gold Screening [MOVIE]; (iv) placebo screening plus presence of MFB [MFB]; and (v) Story of Gold plus presence of MFB [MOVIE/MFB].

The PC group was not invited to attend any screening. The other four groups were invited to attend one of 8 screenings (2 per group). Invitations were delivered one week before the screening and two screenings took place every Thursday during September 2012 for 4 weeks. Invitations to each screening were identical and events were held at the same time each week (8am – 11am), chosen because the cleaning of the market took place at this time, ensuring low opportunity costs to participation since businesses were not allowed to trade during this time. This uniformity of invitations and event dates was used to minimize the possibility of differential take up across screening events.

In C screenings, people were shown a Nollywood movie that had no financial messages associated with it but were given a brief talk after the event about the importance of hygiene in markets to provide quality products and services. This was done explicitly to control for the "event effect" of having received a personalized invitation and participation in a big screening event possibly confounding results, and also to create a comparable group of compliers in both treatment and control groups to simplify the analysis. The standard Credit Awareness program (screening *The Story of Gold* and interacting with an MFB) was split in order to differentiate the impact of the movie from the increased access of financial products coming from the MFB's presence. As such, a 2x2 factorial design was implemented for the treatment arms in order to detect the differential impact of each component and the interaction effect relative to C.

In total, 1261 people (60% of those invited) attended the movie screenings, where a short questionnaire was administered at the end of the event to measure perceptions and attitudes about savings, borrowing and MFBs. Administrative records were kept at the MFB and MOVIE/MFB events to record the people that (i) engaged with *Accion* to open an account at a later stage and (ii) actually opened an account at the event.

Four months later, in February 2013, a follow up survey was conducted on all baseline respondents to collect longer-term data on attitudes, intentions and behaviors with respect to saving and borrowing activities to assess the longevity of any impacts identified at the screenings.

2.6 Outcome-measures, baseline balance and attrition

The main outcome measures are aligned with the essential messages of the Nollywood movie. They can be divided into four categories that capture (i) perceptions of MFBs, (ii) perceptions of women, (iii) intentions to save or borrow, and (iv) savings and borrowing behavior.

Regarding the perceptions of MFBs, the survey asked the micro-entrepreneurs if they agree or disagree with statements such as, "I would trust an MFB to keep my money save", "MFBs treat people with respect", "If I apply to an MFB for a loan, my application will be accepted". Since the movie focused on female entrepreneurs as the main protagonists, we also explore self-reported perceptions of female business competence and access to financial opportunities. Questions designed to explore perceptions of women as business owners or financial decision makers ask respondents if they agree or disagree with statements like "Women can run businesses just as well as men", "Women make better financial decisions than men", "It is easier for men to receive loans than for women". The intention to save or borrow questions capture whether respondents agree with statements such as "I plan to apply for a loan in the next 6 months" or "I will save some money next month". Self-reported savings and borrowing behavior is captured through responses to questions such as "I saved money last month", the amount of total savings relative to the monthly income earned, savings kept at MFBs, savings at commercial banks, outstanding loans from commercial banks, MFBs, suppliers, money lenders, or family/friends. Actual savings behavior is measured through administrative records of those who engaged with representatives of *Accion* to open an account, and those who actually opened an account at the screening event.

Neither financial knowledge, nor basic numeracy skills were specifically addressed in the movie's storyline. Nevertheless, the survey also included 6 quiz-like questions with true and false choices to assess respondents' understanding of basic financial concepts as well as their numeracy skills. The underlying motivation for including these questions is that economic models of savings and investment choice consider both as indispensable for good financial decision taking (Lusardi and Mitchell, 2013). In particular, respondents were required to do simple divisions, to perform basic calculations related to interest rates, to identify the better bargain among two different savings and loan products, and to demonstrate their understanding of how inflation affects their savings. Lastly, one question aimed to evaluate respondent's know how needed to successfully interact with financial institutions (awareness of required documentation for being able to open an account).

Since single questions provide a rather incomplete picture of respondents' levels of financial knowledge, an arithmetic financial knowledge score ranging from 0-6 was calculated by summing up the correct answers to these 6 questions.

To reflect the level of difficulty associated with each question, an alternative financial knowledge score has been developed, which weights every question with the inverse of the proportion of respondents who was able to provide a correct answer. Therewith, larger weights are given to questions that fewer people answered correctly.

2.6.1 Baseline balance

Table 2.1 reports summary statistics for the entire sample, as well as for each of the 5 assignment groups for all exogenous variables including information from the baseline listing, and time-invariant variables measured at follow up. Results are thus reported on balance for business owners that were included in both the baseline and follow up survey (n=2357). The micro-entrepreneurs comprising the total sample are on average 38 years old, predominantly female (71%), married (84%), Christians (64%), are able to speak English (70%), completed high school as their highest level of education (50%), and live in households with an average size of 4.5 individuals. They are experienced in running a business (on average around 11 years of experience), and more than half of the sample (57%) already holds a savings account.

Given that treatment was randomly assigned, the 5 assignment groups are expected to have similar characteristics. Columns (4), (6), (8) and (10) in Table 2.1 show the mean baseline characteristics of all micro-entrepreneurs surveyed at the baseline by treatment group (including the pure control). Columns (5), (7), (9) and (11) report the p-values of the t-test for equality of each of these mean baseline characteristics against those in the (placebo) control group. No characteristics are significantly different from the control (placebo) group at the 5% level for the three treatments, except for the proportion of Igbo business owners in the MOVIE/MFB group. The expectation of balance on observable baseline characteristics also holds across between treatment groups, which supports our claim that the randomization worked well. We see for the Pure Control group, however, that 3 of the 26 characteristics are significantly different at a 5% level (we would expect significant difference in one of every twenty measures by chance). Particularly concerning is that there is imbalance on having a savings account (56% in placebo control group; 63% in the pure control group). This is likely to have been driven by differential non-response at follow up, where we find higher non-response rates in the pure control. We also explore balance across treatment groups for male and female business owners separately (Tables 2.15, 2.16, 2.17, 2.18) and find similar results.

Table 2.2 reports the mean characteristics of those who were assigned to a screening event (Column 1) which excludes individuals in the pure control group, and details observable differences of those who attended (Column 2) with those who did not (Column 3). As indicated in Column 4, the selection into screenings is strongly correlated with more educated micro-entrepreneurs, who are more likely to speak English, enjoy higher access to financial products, and are more likely to keep financial records for their business. This selection process may be explained by the way the screening events were framed: business owners were told that they were invited to a “business development” event and the invitation was in English (see Annex 1 for an example of the invitation). Since a major aim of edutainment is to reach out to the “bottom of the pyramid”, future edutainment activities may want to consider framing the event less as business development and more as entertainment, as well as promoting and designing it in a way that language is not perceived as a barrier to attendance. Overall participation rates are reasonably high (60%) when compared to other financial literacy programs, but it is clear that non-participants present a target group that potentially has the most marginal added value to participation but are at the same time the most difficult group to entice into these types of events.

Although there is strong evidence of self-selection into screening events, Table 2.3 shows that the drivers of this selection across screening events appear to be the same. For those that participated, we see balance across observable characteristics, which is in line with the fact that all screening events were marketed in the same way with the same characteristics. This balance of selection across events supports the possibility of comparing attendees against each other, rather than needing to rely on the intention to treat estimates.

2.6.2 Attrition

The attrition rate in this study is 21.1%, which is relatively high compared to other household surveys (e.g. EFInA 2010 had an attrition rate of 6%), but within reason when compared to enterprise surveys. Intensive efforts were made to reach all respondents which were listed at the baseline, but around 12% could not be contacted again, some refused to be re-interviewed (2.9%), and very few (0.3%) were unable to participate (e.g. for health related reasons). This attrition rate also includes former micro entrepreneurs (5.7%), who may not be considered as being eligible anymore, because they shut down their business between the baseline listing and the end-line survey. If former micro business owners are not taken into account, the attrition rate reduces to 16.3%. There is some evidence for selective attrition for the pure control group, but good balance between the placebo and three treatment arms.[55]

2.7 Model specifications

In this study we effectively have three treatment arms: Movie, MFB, and Movie/MFB. Given that the intervention assignment was randomly allocated, we can measure the causal impact of these interventions through a simple linear regression that identifies the average treatment effect (ATE) using the intention-to-treat estimator (ITT):

$$Y_i = \alpha + \sum_{j=1}^{3} \gamma_j T_{ij} + X_i + \varepsilon_i \qquad [1]$$

Where Y_i is the outcome interest for participant *i*, , and T_{ij} is the treatment status for person *i* with regard to treatment *j.* Treatment *j*= {1,2,3}, for each of the three treatment groups. X_j is a vector of exogenous control variables collected at baseline or time-invariant variables collected in the endline survey[56]. We run the same regression without controls and find point estimates to be unchanged in the analysis, consistent with the balanced nature of the selected control variables, and as such we report the adjusted results in the paper.

[55] See a detailed analysis of attrition in the Annex.

[56] The control variables included in the analysis are: business owner age, marital status, ethnicity, ability to speak English, education level, household size, religion, business experience, number of employees at baseline, whether they had a savings account or kept financial records at baselines, and whether they operated in the main market area or in the outskirts (geographically defined through GPS).

Since we are particularly interested in gender differentials, our second specification explores the impact heterogeneity by gender.

$$Y_i = \alpha + \beta G_i + \sum_{j=1}^{3}(\gamma_j + G\delta_j)T_{ij} + X_i + \varepsilon_i \quad [2]$$

Here G_i = 1 if male, 0 if female. The regression results presented in the tables generated from the analysis include the effect of treatment *j* on females (γ_j), the additional impact for males (δ_j) and the overall gender differential G_j. Each table of results presents results from Equation 1 first, followed by gender-disaggregated results from Equation 2.

In Section 3 we see that overall selection into the movie screening is such that those that attended the events were slightly different to those that did not attend the events. However, we find that this selection pattern is the same across all screening events (based on balance of observable characteristics) and, importantly, there are no differential selection patterns between 3 treatment arms and placebo screening C. In this case we run a restricted analysis on those business owners that actually attended the event. Relying on the balance across an extensive set of baseline variables and the manner in which the events were implemented (randomized invitations at the individual level), we reasonably expect this comparison to provide an unbiased estimate of the Average Treatment Effect on the Treated (ATET) – the impact for those that actually attended the event, using Equation 1 and 2 with the restricted sample of 1261 participants.

We acknowledge that, if there are large positive spillovers, this may result in a downward bias of the estimate of impact. As such, the survey included control "clusters" that were created through geographic discontinuities, where a self-contained cluster meant that all businesses within the cluster were at least 20 meters away from the next closest business outside of the cluster[57]. This sampling method creates a "pure" control group less exposed to treatment neighbors, thus exogenously varying the level of intensity of treatment in any particular area of the market, theoretically allowing us to explore spillovers. We see, however, in the pure control group that we experience differential attrition resulting in an imbalance based on baseline observable variables. As such, we exclude this group from analysis in this paper. In the following section we present results using Equation 1 with the restricted sample of business owners that actually attended a screening, using the placebo group as our control comparison.

[57] We use the rule of 20m for businesses outside of the main market area. Density is too high for businesses inside the main market area, in which case we use a 5m rule.

2.8 Results

2.8.1 Exposure

Administrative records were kept on who participated in the screenings, using the personalized invitations to verify details and treatment status, which was a requirement for entry into the movie screening. The screenings were secured and private with complete control over the entrance and exit of the events. Although participation rates averaged around 60%, contamination was very low as a result of this process. Table 2.6 highlights this fact, where less than 1% of invited guests went to a different screening to the one they had been assigned to, strengthening the justification to use Equation 1 and 2 with our restricted sample to measure the ATET.

In the follow up survey we asked for self-reported exposure, partly to confirm attendance, but also to understand whether people could remember the main activities and messages from the events – presented as a summary in Table 2.7. While people have no problem recalling the screening, they express some confusion about the details of the event. We find that 95% of people recall receiving an invitation and 96% of the people that were recorded through administrative records as attending the event confirmed that they had attended. When asked specifically about whether they saw the *Story of Gold*, 90% in Movie and 93% in Movie/MFB acknowledged that they had done so, while 77% and 82% respectively could recall the main message of the movie without prompting. However C and MFB groups also reported having seen the movie, although at significantly lower levels (59% and 58% respectively). Since the movie was tightly controlled, and not released to the public, this suggests a potential confusion between *The Story of Gold* and the placebo movie screening – possibly confounded by the fact that neighboring business may have seen and mentioned something about the movie.

Recall of *Accion* presence was much lower. We find significant increases in recall for MFB and MFB/Movie compared to Movie and Control as to be expected, but the proportions are still low. Only 16% of MFB attendees and 17% of Movie/MFB attendees recalled *Accion*'s presence at the event. We also asked a falsification question to assess the level to which respondents may have been adjusting their answers to respond positively to the interview. We find that only 1% of people responded positively to a question asking whether a certain MFB (Jaiz Bank), that is only based in Abuja, had visited them (an impossibility), and this is similar across treatment arms, suggesting that positive response bias does not seem to be a problem in our case. Since the interventions were monitored carefully and *Accion* was indeed present at these events, this contrast between *Accion* and *Story of Gold* recall highlights the differential salience of each of the interventions.

2.8.2 Financial Literacy

The quiz questions test basic numeracy and financial concepts. Since the movie screening aimed to influence emotions and perceptions rather than formal financial literacy, we expected these indicators to show balance across groups, which they do. Aggregating the questions into a single index, we find two things (see Table 2.8): (i) scores are very similar across all groups and (ii) the aggregate scores are relatively high, with the weighted and arithmetic scores yielding similar results, perhaps reflecting a lack of variation and cognitive separating ability of the set of questions. However, when exploring the covariates associated with these financial literacy scores, we find strong relationships between the overall score and (i) whether business owners had a savings account at baseline and (ii) whether they had any schooling, supporting the assertion that the indices are informative in distinguishing between financial literacy levels, and the similarities in scores across groups reflects balance induced by the randomization.

2.8.3 Perceptions

We find increases in self-reported trust and perceptions of MFBs directly after the screening events; however, when asked the same questions in the follow up survey, many of the initial differences reduce or disappear[58]. While males are influenced most strongly by the movie stimulus in the short run, differentials in self-reported trust only sustain for females in the longer run.

Table 2.9 presents the results from the screening and endline surveys. While the movie on its own has some impact on whether people report that they would trust an MFB to keep their money when people were asked this question at the screening, the presence of *Accion* seems to have a much larger effect than the movie, and there is no additivity of the interventions (although both are significant and positive). In the second follow up survey, we see that the differential between control and treatment group trust declines; however, it is the movie treatment arms that sustain results, where the impact on MFB reduces to insignificance. This sustained impact is almost entirely driven by females, even though males were most affected by the movie in the short run. A supporting question identifying positive perceptions of MFBs ("MFBs treat people with respect") shows similar results, with larger impacts for males in the short run, followed by some limited, but sustained differences for females in the longer run, even when male differentials disappear. This significant impact is only found in the combined Movie/MFB arm.

[58] Direct comparison between the two follow up surveys should be handled carefully. Although the questions asked were identical, the response method varied across data collection activities. In the immediate follow up the question responses were yes/no, and the questionnaire was self-administered. In the 4-month follow up survey the questionnaire was administered by an interviewer and response options were: strongly agree; agree; disagree; strongly disagree.

We also explore perceptions of ease in obtaining a loan and riskiness of doing so. Both the movie and MFB treatments have a significant positive effect on business owners perception of how likely it is that they may receive a loan if they applied for one in the short run (this falls away completely in the longer run), but none of the interventions have any impact on beliefs of the risk in taking out a loan[59].

2.8.4 Intentions

We tested business owners' intentions about their saving and borrowing plans, once again through the screening questionnaire and in the follow up, with results presented in Table 2.11. Here there is mixed evidence, with some impact on borrowing intentions, but no changes on what are already very high intentions to save. Intention to save is almost universal – 90% at the screening and 95% in the follow up respondents indicated that they planned to save some money in the following month. When we compare this to actual saving in the past month (65% in the endline survey – Table 2.13) it is clear that there is a disconnect between intentions and behavior, with many more business owners planning to save, but not necessarily following through with these plans, reinforcing the possibility that various frictions may be reducing people's ability to translate intention into action. The reason for this disconnect could be manifold: (i) hyperbolic discounting; (ii) lack of disposable funds; (iii) overconfidence or (iv) limited access to financial products, and we cannot necessarily disentangle all of these factors; however, we do see that the interventions provided have little influence on what are already very strong self-reported intentions to save, suggesting that this is not likely the channel through which any behavior change occurs.

2.8.5 Savings Behavior

At screening events with MFBs present, business owners were able to discuss savings opportunities with the MFB and sign up for a savings account on the spot if they were interested. Participants had two options when expressing interest in opening up an account with the MFB: (i) business owners would meet with the MFB and sign up for a follow up visit to open an account; or (ii) business owners would sign up for an account on the spot. Table 2.12 reports on the data collected at the two types of screening events (MFB; MOVIE/MFB) showing that people were more likely to express interest in opening an account by visiting the MFB stand directly after the event in the MFB group (13% vs. 8%). However, differentiating this visit into each of the two options available (signing up on the spot, or agreeing to a follow up visit to sign up for an account) we find substantial differences. The majority of people in the MFB group that visited the MFB stand opted for a

[59] We also explore perceptions of female business owners, see Table 2.10.

follow up visit rather than signing up on the spot. However, the Movie/MFB combination event was substantially more effective at incentivizing on-the-spot savings account sign ups at the event, and this effect was strongest for male participants. The MOVIE/MFB combination event motivated 7% of participants to open an account on the spot (compared to 2% in the MFB group), but this effect was substantially different between male and female participants (5% of females and 11% of males). The overall difference is statistically significant, but the gender-disaggregated differences are only significant for males.

Although the MFB event was moderately successful in encouraging people to visit their stand and agree to a follow up visit (11%), on further inspection we find that none of the people in this category actually followed up after the event (Table 2.13). In fact the only people that followed up with an MFB after the screening came from Movie, where the MFB had not been present. Although a small fraction (2% for both males and females), this is the only group with a statistically significant increase. The results provide the following insights: (i) reducing access barriers to virtually zero (MFB condition) increases engagement with the MFB and reported interest in opening an account, but has only a modest effect on actual sign up rates; (ii) even without having an immediate call to action (the ability to open an account on the spot) *The Story of Gold* has some (although very limited) impact on short-term behavior, inducing 2% of participants to follow up with an MFB afterwards (Movie condition); but (iii) combining the reduced access constraint with the movie designed to promote savings (Movie/MFB) provides the strongest incentive to open a savings account, mostly driven by male participant choices. The evaluation design helps to deconstruct some of the potential barriers to demand for a savings account and identifies that an educational event attached to an emotional stimulus can be an effective tool to increase take up, but only when combined with an intervention that allows for immediate action. However, this tells us little about savings behavior after the event.

Despite the strong impacts observed, important concerns arise from the follow up findings. Firstly, we find that 67% of all participants that opened a savings account at the event reported having a savings account at baseline (significantly higher than the average for our sample). While there may be rational reasons to hold multiple accounts (or to change accounts), the finding reinforces the fact that the intervention may be inducing action only in a sub-population that has lower marginal gains in doing so when compared to the unbanked target population. The second related concern is that in the follow up we find no distinguishable difference in whether respondents have a savings account, which is not surprising given that the majority of those induced to open an account already had one prior to the screenings. More concerning, however, we find that males in the Movie/MFB group report having been *less* likely to save some money in the month prior to the follow up survey and show no differences in saving amounts relative to their income.

While it is not clear what may be driving this result, it is possible that the event, while successfully motivating business owners to act in the moment and put money in a new savings account, only served to displace future savings, with no net gain.

2.8.6 Borrowing Behavior

For borrowing behavior, we rely only on self-reported responses in the follow up survey. The movie message centered on "responsible borrowing", highlighting the problems with relying on moneylenders, and we reflect on this through two particular indicators: (i) borrowing rate in last 4 months and (ii) the source of borrowing. In particular, we were interested in identifying whether business owners used formal or informal sources for financing. We find firstly that borrowing rates are substantial – about half of all business owners reported taking out a loan in the past four months, and half of those that took a loan did so from an informal source. The interventions have no effect on borrowing rates (although there is a reduction in all treatment groups, this is not significant). Similarly we find (see results in Table 2.14) little to no evidence on changes in the form of lending, although females in the Movie/MFB group reduce informal lending by 14 percentage points, which is borderline significant. Interestingly there seems to be more congruency between intentions to borrow and actual borrowing than for savings intentions and behavior. While 54% of people mentioned that they were planning to take out a loan in the next 6 months immediately after the screening, we find 4 months later that 51% of people did so. This contrasts sharply with the intended savings (90%) and actual savings rates (60%) which seems to confirm that, in terms of saving behavior, there are several additional barriers at play in addition to those that the interventions address directly.

2.9 Robustness checks

Our results show a significant effect of Movie/MFB on motivating business owners to open a savings account, but with little to no evidence of longer term impact on a broad range of savings and borrowing perceptions and behavior. A null effect could be a result of (i) limited power, driven by sample sizes too small to detect true impacts; (ii) spillovers improving outcomes for the control group; or (iii) selection bias resulting from the control group participants having different participation decisions to our treatment groups.

Power is of concern when we measure heterogeneous impacts by gender, given that only 28% of our sample is male. We run each of the regressions reported in this paper for the entire sample (without differentiating by gender) and continue to find mostly null to low effects on our outcomes of interest in the 4-month follow up[60]. Here our sample is substantial, and power is less of a concern. However, in most cases the point estimate of the effects is so small that the interpretation of the results would not change even in cases we were to have enough sample power to estimate these small changes.

The study was originally designed to account directly for potential spillovers, given that all participants came from the same market area and interaction between participants was expected. The pure control group was generated using cluster-randomization to address this; however, as mentioned previously, we are unable to use this group due to selective attrition and cannot rule out potential spillovers. However, given that we see the strongest effects of the intervention being in the immediate term, and given the nature of the program (increasing short-term motivation rather than focusing on financial content), it seems somewhat unlikely that secondhand information passed from treatment to control business owners is likely to be a serious concern.

Our restricted regression analysis used throughout the paper, effectively reports on the average treatment effect on the treated, without reference to the intention-to-treat (ITT) results which limits the scope of interpretation to effects on those that were actually convinced to attend the event. We run ITT regressions, including all business owners invited to the screening events on outcomes that were recorded at the endline, but do not report these results here. Unsurprisingly (see discussion above on why we can rely on the treatment effect on the treated in this context), the null effects remain, and our outcomes where impacts were found mostly remain significant, albeit with lower point estimates for impact[61].

Finally, reflection on the savings account take up rates on which we find significant impacts is required. Why is it that males react most strongly to the screening event in the short run? This could reflect the fact that male emotions are affected more than females, inducing action, but may equally reflect the possibility that females have added constraints beyond motivation that affect take up, such as low liquidity or limited autonomy in financial decision making. The literature has found that females often make decisions jointly with their spouse or other counterpart, when compared to male business owners. However, we find that business autonomy is balanced across gender in our sample with 92% of males and females reporting that they make business decisions on their own. We do find, however, that business revenues and profits across gender differ significantly, with males having nearly twice the yearly profits of females. However, selection equations regressing profits and revenue with the likelihood of opening an account show no relationship. Furthermore, we find that intermediate outcomes such as increased self-reported trust in

[60] As expected, we do find cases where significant results in the gender-disaggregated analysis becomes non-significant in the pooled specifications, particularly when male and female effect coefficients have opposite signs.

[61] Informal lending is no longer significant.

MFBs are substantially stronger for males than females. This suggests that, rather than females facing added constraints that the screening event does not overcome, the events have a differential effect on perceptions by gender that seems to be driving the differential take up of savings accounts at the event.

2.10 Discussion and Conclusion

The primary role of the evaluation was to explore the use of a new medium to transmit financial messages, focusing on the use of heuristics and emotions to spur action in the short run with the intention of getting business owners a foot in the door to use financial products more regularly, learning and building experience thereafter. The second objective was to identify how access to financial products and motivation interact to induce action, and whether choice architecture can be effectively utilized to promote welfare-enhancing financial decisions.

The results from the evaluation are mixed and warrant further discussion on three issues of importance for policy dialogue: (i) the ability of edutainment to reach out to the targeted population; (ii) the role of choice architecture on influencing short-term decisions; and (iii) ensuring sustained behavior change.

Recent evidence has highlighted the challenges to encouraging people to attend voluntary financial literacy workshops and other training programs (Bruhn and McKenzie, 2013). Low take up rates are common, and this is especially true for interventions targeting business owners. Business owners may be making a rational decision to avoid the training because of low perceived benefits. Using edutainment to transmit financial messages is a new approach that has the potential benefit of being more inclusive, lowering barriers to participation. Response rates in this study of approximately 60% reflect that, even though these events are able to reach out to the majority of potential participants, this is far from universal and more effort is needed to find ways to market these events to have more mass appeal. In particular, the least educated people with lowest access to financial products were the ones that selected out of the screening events, highlighting the difficulty of reaching out to this sub-population.

The study identifies a strong interaction between offering a stimulus (the movie) together with a direct outlet (the presence of the MFBs) for acting on this motivation. This result is not surprising, and replicates what is well known among marketers in a development setting. However, applying choice architecture to a development setting requires careful attention to the potential unexpected outcomes that may result. In our case, the once-off screening was effective at encouraging people to open new accounts, but on closer inspection, nearly two thirds of these people already had savings accounts, possibly limiting the potential marginal impact of the work. This highlights the importance of testing potential interventions at a pilot level, measuring and understanding the determinants of take up before scaling up.

While the intervention was able to influence decisions in the short run, people make financial decisions on a daily basis, and more sustained behavior change is critical in the context of saving. Our limited longer-term impacts emphasize this point. The ability to spur people into action through the use of edutainment may have more development impact for activities that are beneficial as once-off actions, particularly given the intervention's relatively low cost and simple logistics. Examples of where these types of interventions could work in other development areas could include, for instance, encouraging people to test themselves at mobile clinics for HIV/AIDS or taking vaccinations, where one-time actions of groups of people at once can have important private and public benefits. This approach could also be tailored to more sustained financial behavior change if coupled with commitment savings accounts – where decisions taken in the moment have a more binding effect in the longer-run (Ashraf et al., 2006). However, take up of financial instruments tells us little about how this increased exposure may strengthen financial capabilities – responsible use of these instruments and financial decision making more generally. The literature has traditionally explored the direction for strengthening financial capabilities as going from education to better financial decision making and increased use of financial products. There is less understanding of how a "learning-by-doing" approach – focusing on providing access to financial instruments and exploring how this translates into experiential learning and ultimately improved decision making. While we have seen that nudges can be developed to help overcome the access constraint, it is still unclear as to whether this can be effectively translated into strengthened financial capabilities in the longer run.

2.11 References

Andrade, E. B., & Ariely, D. (2009). The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes*, *109*(1), 1-8.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73-106.

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational behavior and human decision processes*, *35*(1), 124-140.

Ashraf, N., Fink, G., & Weil, D. N. (2010). *Evaluating the effects of large scale health interventions in developing countries: The zambian malaria initiative* (No. w16069). National Bureau of Economic Research.

Ashraf, N., Karlan, D., & Yin, W. (2006). Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *The Quarterly Journal of Economics*, *121*(2), 635-672.

Bernard, T., Dercon, S., Orkin, K., & Taffesse, A. (2014). *The future in mind: Aspirations and forward-looking behaviour in rural Ethiopia*. London: Centre for Economic Policy Research.

Bernard, T., Dercon, S., Orkin, K., & Seyoum Taffesse, A. (2015). Will video kill the radio star? Assessing the potential of targeted exposure to role models through video. *The World Bank Economic Review*, *29*(suppl_1), S226-S237.

Berg, G., & Zia, B. (2013). Harnessing emotional connections to improve financial decisions: evaluating the impact of financial education in mainstream media. *World Bank Policy Research Working Paper*, (6407).

Berge, L. I., Bjorvatn, K., & Tungodden, B. (2011). Human and financial capital for microenterprise development: Evidence from a field and lab experiment. *NHH Dept. of Economics Discussion Paper*, (1).

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., & Zinman, J. (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, *125*(1), 263-306.

Bruhn, M., Legovini, A., & Zia, B. (2012). Financial Literacy for High School Students and Their Parents: Evidence from Brazil. *World Bank Working Paper*.

Bruhn, M., Ibarra, G. L., & McKenzie, D. (2013). *Why is voluntary financial education so unpopular? Experimental evidence from Mexico* (No. 6439). The World Bank.

Bruhn, M., & Zia, B. (2011). Stimulating managerial capital in emerging markets: The impact of business and financial literacy for young entrepreneurs. *World Bank Policy Research Working Paper Series, Vol*.

Cardoso, A. R., & Verner, D. (2006). School drop-out and push-out factors in Brazil: The role of early parenthood, child labor, and poverty.

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, *69*, 318-318.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (1983). Applied multiple regression/correlation analysis for the behavioral sciences.

Cole, S., & Fernando, N. (2008). Assessing the importance of financial literacy. *ADB Finance for the Poor*, *9*(2).

Cole, S. A., Sampson, T. A., & Zia, B. H. (2009). *Financial literacy, financial decisions, and the demand for financial services: evidence from India and Indonesia*. Harvard Business School.

Cole, S. A., & Shastry, G. K. (2008). *If You are So Smart, why Aren't You Rich?: The Effects of Education, Financial Literacy and Cognitive Ability on Financial Market Participation*. Harvard Business School.

Cole, S., Sampson, T., & Zia, B. (2011). Prices or knowledge? What drives demand for financial services in emerging markets?. *The journal of finance*, *66*(6), 1933-1967.

Coville, A., Di Maro, V., Dunsch, F. A., & Zottel, S. (2019). The Nollywood Nudge. World Bank.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 448-474.

Datta, S., & Mullainathan, S. (2012). Behavioral Design.

Deaton, A. (1989). *Saving and liquidity constraints* (No. w3196). National Bureau of Economic Research.

Drexler, A., Fischer, G., & Schoar A. (2012). Keeping it Simple: *Financial Literacy and Rules of Thumb*. Mimeo. LSE

Duflo, E., & Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly Journal of Economics*, *118*(3), 815-842.

EFinA. (2012). EFInA Access to Financial Services in Nigeria 2012 Survey – Key Findings. Accessed October 7, 2019, from http://83.222.230.50/assets/ResearchDocuments/2013-Documents/EFInA-Access-to-Financial-Services-in-Nigeria-2012-surveyKey-Findings2.pdf

Finucane, M. L., Slovic, P., Mertz, C. K., Flynn, J., & Satterfield, T. A. (2000). Gender, race, and perceived risk: The 'white male' effect. *Health, risk & society*, *2*(2), 159-172.

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal ol Personality and Social Psychology*, *4*(2), 195-202.

Gibson, J., McKenzie, D., & Zia, B. (2012). The impact of financial literacy training for migrants. *World Bank Policy Research Working Paper*, (6073).

Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology*, *29*(56), 781-795.

Harshman, Richard A. & Paivio Allan (1987), "Paradoxical" sex differences in self-reported imagery. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 41(3), 287-302.

Hilgert, M. A., Hogarth, J. M., & Beverly, S. G. (2003). Household financial management: The connection between knowledge and behavior. *Fed. Res. Bull.*, *89*, 309.

Hinz, R. P., McCarthy, D. D., & Turner, J. A. (1997). Are women conservative investors? Gender differences in participant-directed pension investments. *Positioning pensions for the twenty-first century*, *91*, 103.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American economic review*, *93*(5), 1449-1475.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management science*, *39*(1), 17-31.

Karlan, D., Ratan, A. L., & Zinman, J. (2013). Savings by and for the Poor: A Research Review and Agenda (No. 1027). Economic Growth Center (Yale University) Discussion Paper No. 1027.

Kennedy, M. G., O'Leary, A., Beck, V., Pollard, K., & Simpson, P. (2004). Increases in Calls to the CDC National STD and AIDS Hotline Following AIDS-Related Episodes in a Soap Opera. *Journal of Communication*, *54*(2), 287-301.

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The Foreign-Language Effect Thinking in a Foreign Tongue Reduces Decision Biases. *Psychological science*, *23*(6), 661-668.

Klinger, B., & Schündeln, M. (2011). Can entrepreneurial activity be taught? Quasi-experimental evidence from Central America. *World Development*, *39*(9), 1592-1610.

La Ferrara, E., Chong, A., & Duryea, S. (2012). Soap Operas and Fertility: Evidence from Brazil. *American Economic Journal: Applied Economics*, *4*(4), 1-31.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. *Handbook of affective science*, *619*(642), 3.

Lusardi, A. (2007). *Household saving behavior: the role of literacy, information and financial education programs* (No. 2007, 28). CFS working paper.

Lusardi, A., & Mitchell, O. S. (2007). Baby boomer retirement security: The roles of planning, financial literacy, and housing wealth. *Journal of monetary Economics*, *54*(1), 205-224.

Lusardi, A., & Mitchell, O. S. (2013). *The Economic Importance of Financial Literacy: Theory and Evidence* (No. w18952). National Bureau of Economic Research.

Lusardi, A., & Tufano, P. (2009). *Debt literacy, financial experiences, and overindebtedness* (No. w14808). National Bureau of Economic Research.

Makanjee, M. & Ladipo, M. (2011, November 16). From Data to Action: Using Finscope in Nigeria. Blog post. Retrieved October 7, 2019, from http://www.cgap.org/blog/data-action-using-finscope-nigeria

Mailath, G. J., & Samuelson, L. (2011). Repeated games and reputations: long-run relationships. *OUP Catalogue*.

McKenzie, D., & Woodruff, C. (2012). What are we learning from business training and entrepreneurship evaluations around the developing world?.

McKnight, P. E., McKnight, K. M., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Mulaj, F., & Jack, W. (2012). *Evaluating the efficacy of mass media and social marketing campaigns in changing consumer financial behavior* (No. 73924). The World Bank.

Ottati, V. C., & Isbell, L. M. (1996). Effects of mood during exposure to target information on subsequently reported judgments: An on-line model of misattribution and correction. *Journal of Personality and Social Psychology*, *71*, 39-53.

Pathak, P., Holmes, J., & Zimmerman, J. (2011). Accelerating financial capability among youth: Nudging new thinking. *New America Foundation*.

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American political science review*, 251-267.

Pocheptsova, A., & Novemsky, N. (2010). When do incidental mood effects last? Lay beliefs versus actual effects. *Journal of Consumer Research*, *36*(6), 992-1001.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology*, *45*(3), 513.

Servon, L. J., & Kaestner, R. (2008). Consumer Financial Literacy and the Impact of Online Banking on the Financial Behavior of Lower-Income Bank Customers. *Journal of consumer affairs*, *42*(2), 271-305.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, *177*(3), 1333-1352.

Spader, J., Ratcliffe, J., Montoya, J., & Skillern, P. (2009). The bold and the bankable: How the Nuestro Barrio telenovela reaches Latino immigrants with financial education. *Journal of Consumer Affairs*, *43*(1), 56-79.

Stango, V., & Zinman, J. (2009). Exponential growth bias and household finance. *The Journal of Finance*, *64*(6), 2807-2849.

Sunden, A. E., & Surette, B. J. (1998). Gender differences in the allocation of assets in retirement savings plans. *The American Economic Review*, *88*(2), 207-211.

Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. The University of Chicago Law Review, 1159-1202.

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters*, *8*(3), 201-207.

Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of political Economy*, *112*(S1), S164-S187.

Tufano, P., Flacke, T., & Maynard, N. W. (2010). Better Financial Decision Making among Low-Income and Minority Groups.

Tversky, A., & Kahneman, D. (1974). Heuristics and biases: Judgement under uncertainty. *Science*, *185*, 1124-1130.

Vanguard. (2013, March 14). Beyond Project Nollywood. Blog post. Accessed 7 October, 2019, from https://www.vanguardngr.com/2013/03/beyond-project-nollywood/

Vohs, K. D., Baumeister, R. F., & Loewenstein, G. (2007). *Do emotions help or hurt decision making?: a hedgefoxian perspective*. Russell Sage Foundation Publications.

Willis, L. E. (2011). The Financial Education Fallacy. *American Economic Review*, *101*(3), 429-34.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of personality and social psychology*, *60*(2), 181-192.

2.12 Annexes

Table 2.1: Baseline Balance

	Total sample		Control	Movie		MFB		Movie + MFB		Pure control	
Variable	N	Mean	Mean	Mean	P-value	Mean	P-value	Mean	P-value	Mean	P-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Personal characteristics											
Age of respondent	2314	37.76	37.90	37.52	0.553	37.89	0.996	37.31	0.339	38.44	0.427
Gender (male)	2358	0.29	0.26	0.30	0.173	0.30	0.220	0.29	0.371	0.31	0.138
Married	2357	0.84	0.85	0.82	0.211	0.86	0.557	0.82	0.206	0.86	0.845
Widowed	2357	0.02	0.02	0.03	**0.094***	0.01	0.284	0.02	0.984	0.03	0.264
Single	2357	0.14	0.13	0.15	0.551	0.13	0.795	0.16	0.190	0.12	0.494
Muslim	2356	0.36	0.35	0.40	0.136	0.35	0.793	0.36	0.717	0.33	0.421
Christian	2356	0.64	0.64	0.60	0.154	0.65	0.958	0.63	0.621	0.67	0.387
Can speak English	2346	0.70	0.70	0.67	0.321	0.72	0.450	0.71	0.636	0.73	0.382
Igbo	2356	0.20	0.17	0.17	0.925	0.21	0.141	0.22	0.104	0.24	**0.012****
Yoruba	2356	0.75	0.78	0.78	0.873	0.75	0.219	0.72	**0.035****	0.71	**0.025***
Other ethnicitiy	2356	0.05	0.05	0.05	0.635	0.04	0.777	0.06	0.242	0.04	0.839
Education											
No completed school education	2356	0.07	0.06	0.07	0.421	0.08	0.180	0.08	0.297	0.08	0.347
Primary school education	2356	0.22	0.24	0.24	0.968	0.21	0.164	0.21	0.209	0.19	**0.067***
High school diploma	2356	0.50	0.49	0.48	0.749	0.50	0.754	0.51	0.527	0.53	0.329
Diploma	2356	0.10	0.11	0.10	0.512	0.11	0.825	0.09	0.276	0.11	0.945
Graduate school	2356	0.10	0.09	0.10	0.866	0.10	0.626	0.11	0.425	0.09	0.916
Household characteristics											
Household (HH) size	2343	4.53	4.58	4.57	0.902	4.43	0.168	4.48	0.395	4.61	0.825
Number of children below 12 in HH	2311	1.33	1.38	1.29	0.230	1.30	0.311	1.25	**0.080***	1.44	0.524
Number of dependents in HH	2322	2.44	2.45	2.39	0.671	2.41	0.769	2.41	0.747	2.57	0.385
Number of dependents outside the	2213	1.55	1.50	1.53	0.843	1.53	0.827	1.54	0.784	1.66	0.330
Business characteristics											
Months in operation	2310	97.40	98.69	97.58	0.847	96.98	0.771	101.02	0.698	91.03	0.218
Has a savings account	2350	0.57	0.56	0.57	0.732	0.54	0.624	0.57	0.753	0.63	**0.035****
Keeps written financial records	2340	0.37	0.36	0.35	0.684	0.37	0.708	0.38	0.619	0.40	0.315
Operating inside main market	2324	0.25	0.24	0.26	0.500	0.24	0.985	0.26	0.535	0.27	0.287
Number of employees	2352	1.44	1.57	1.46	0.345	1.40	0.169	1.39	0.161	1.36	0.168
Business experience in years	2350	10.75	10.84	10.77	0.892	10.78	0.907	10.48	0.497	10.97	0.834

*** p<0.01, ** p<0.05, * p<0.1

Table 2.2: Selection into screenings

Variable	Total		Participated in screening		Did not participate		
	N	**Mean**	**N**	**Mean**	**N**	**Mean**	**P-value**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Personal characteristics							
Age of respondent	1946	37.63	1242	38.26	704	36.52	**0.000*** **
Gender (male)	1984	0.29	1260	0.28	724	0.30	0.368
Married	1983	0.84	1259	0.85	724	0.82	**0.054***
Widowed	1983	0.02	1259	0.02	724	0.01	**0.031****
Single	1983	0.14	1259	0.13	724	0.17	**0.004*** **
Muslim	1983	0.36	1260	0.35	723	0.39	0.112
Christian	1983	0.63	1260	0.64	723	0.61	0.111
Can speak English	1974	0.70	1255	0.72	719	0.66	**0.005*** **
Igbo	1982	0.19	1260	0.20	722	0.18	0.149
Yoruba	1982	0.75	1260	0.75	722	0.75	0.965
Other ethnicitiy	1982	0.05	1260	0.04	722	0.07	**0.012****
Education							
No completed school education	1983	0.07	1260	0.06	723	0.10	**0.006*** **
Primary school education	1983	0.22	1260	0.22	723	0.24	0.386
High school diploma	1983	0.50	1260	0.50	723	0.49	0.843
Diploma	1983	0.10	1260	0.11	723	0.09	0.137
Graduate school	1983	0.10	1260	0.11	723	0.09	0.101
Household characteristics							
Household (HH) size	1972	4.51	1251	4.52	721	4.51	0.873
Number of children below 12 in HH	1948	1.30	1234	1.31	714	1.29	0.761
Number of dependents in HH	1954	2.41	1241	2.47	713	2.31	**0.090***
Number of dependents outside the HH	1862	1.53	1179	1.52	683	1.54	0.882
Business characteristics							
Months in operation	1947	98.59	1235	98.76	712	98.30	0.917
Has a savings account	1977	0.56	1260	0.59	717	0.52	**0.002*** **
Keeps written financial records	1968	0.37	1254	0.39	714	0.32	**0.002*** **
Operating inside main market	1979	0.25	1260	0.28	719	0.20	**0.000*** **
Number of employees	1980	1.45	1259	1.45	721	1.45	0.987
Business experience in years	1977	10.70	1256	10.88	721	10.40	0.218

*** p<0.01, ** p<0.05, * p<0.1

Table 2.3: Balance across screening participants

	Total		Control		Movie			MFB			Movie + MFB		
Variable	**N**	**Mean**	**N**	**Mean**	**N**	**Mean**	**P-value**	**N**	**Mean**	**P-value**	**N**	**Mean**	**P-value**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Personal characteristics													
Age of respondent	1243	38.27	309	38.13	327	38.46	0.79	287	37.92	0.81	307	38.52	0.60
Gender (male)	1261	0.28	313	0.25	333	0.26	0.78	292	0.30	0.21	310	0.30	0.19
Married	1260	0.85	312	0.84	333	0.84	0.87	292	0.87	0.41	310	0.85	0.67
Widowed	1260	0.02	312	0.02	333	0.04	0.15	292	0.01	0.24	310	0.02	0.56
Single	1260	0.13	312	0.13	333	0.12	0.55	292	0.12	0.68	310	0.13	0.83
Muslim	1261	0.35	313	0.34	333	0.40	**0.080***	292	0.35	0.79	310	0.32	0.71
Christian	1261	0.64	313	0.66	333	0.60	**0.096***	292	0.65	0.72	310	0.67	0.78
Can speak English	1256	0.72	311	0.71	331	0.70	0.76	292	0.71	0.97	309	0.77	0.13
Igbo	1261	0.20	313	0.19	333	0.19	0.94	292	0.23	0.30	310	0.22	0.40
Yoruba	1261	0.75	313	0.78	333	0.77	0.88	292	0.74	0.34	310	0.73	0.15
Other ethnicitiy	1261	0.04	313	0.04	333	0.04	0.62	292	0.03	0.95	310	0.06	0.18
Education													
No completed school education	1261	0.06	313	0.05	333	0.08	0.26	292	0.05	0.98	310	0.06	0.85
Primary school education	1261	0.22	313	0.24	333	0.23	0.96	292	0.22	0.61	310	0.19	0.15
High school diploma	1261	0.50	313	0.50	333	0.47	0.57	292	0.51	0.77	310	0.52	0.57
Diploma	1261	0.11	313	0.11	333	0.11	0.94	292	0.12	0.76	310	0.11	0.92
Graduate school	1261	0.11	313	0.11	333	0.10	0.74	292	0.10	0.91	310	0.13	0.44
Household characteristics													
Household (HH) size	1252	4.52	311	4.49	332	4.63	0.40	289	4.37	0.36	307	4.54	0.72
Number of children below 12 in HH	1235	1.31	307	1.39	331	1.31	0.34	285	1.27	0.19	299	1.26	0.20
Number of dependents in HH	1242	2.47	306	2.44	331	2.51	0.69	287	2.40	0.85	305	2.51	0.67
Number of dependents outside the	1180	1.52	297	1.50	308	1.47	0.94	274	1.65	0.43	288	1.51	0.93
Business characteristics													
Months in operation	1420	97.23	350	96.94	369	101.37	0.47	334	96.54	0.95	352	95.03	0.76
Has a savings account	1448	0.59	356	0.58	378	0.61	0.30	343	0.58	0.80	356	0.60	0.48
Keeps written financial records	1442	0.39	355	0.40	377	0.36	0.36	341	0.40	0.89	354	0.42	0.48
Operating inside main market	1448	0.27	356	0.27	378	0.26	0.80	343	0.26	0.86	356	0.28	0.73
Number of employees	1448	1.52	356	1.58	378	1.56	0.89	343	1.47	0.54	355	1.48	0.53
Business experience in years	1257	10.89	312	10.95	330	11.02	0.96	292	10.45	0.45	310	11.03	0.92

*** p<0.01, ** p<0.05, * p<0.1

Table 2.4: Attrition in Endline Survey

Dependent Variable: Interviewed in Endline Survey	
	(1)
Movie	-0.014
	(0.02)
MFB	-0.032
	(0.02)
Movie + MFB	-0.021
	(0.02)
Pure Control	-0.069**
	(0.02)
N. of Obs.	2437
R-squared	0
P-value of F model	0.6

* p<0.05, ** p<0.01, *** p<0.001

Table 2.5: Item Non-response across screening participants

Variable	Total sample		Control		Movie		MFB		Movie + MFB	
	Have Item (in %)	INR (in %)	Have Item (in %)	INR (in %)	Have Item (in %)	INR (in %)	Have Item (in %)	INR (in %)	Have Item (in %)	INR (in %)
Knowledge										
Simple Division	100	7.21	100	5.75	100	8.62	100	7.11	100	6.63
Inflation	100	2.37	100	2.18	100	2.40	100	1.46	100	2.21
Necessary documentation	100	3.77	100	3.57	100	3.21	100	3.97	100	3.61
Better savings product	100	1.74	100	2.18	100	1.60	100	1.67	100	2.01
Interest rate	100	4.07	100	4.37	100	5.21	100	3.56	100	3.61
Better loan product	100	2.67	100	3.37	100	2.40	100	2.30	100	3.61
Perceptions										
MFB will accept loan application (screening)	52	0.00	62	0.00	65	0.00	59	0.00	59	0.00
MFB will accept loan application (endline)	100	19.34	100	19.05	100	19.24	100	20.50	100	18.27
Taking a loan is too risky (screening)	52	0.00	61	0.00	66	0.00	60	0.00	60	0.00
Taking a loan is too risky (endline)	100	4.03	100	2.98	100	3.41	100	3.41	100	4.62
Trust in MFBs (screening)	52	0.00	61	0.00	66	0.00	59	0.00	61	0.00
Trust in MFBs (endline)	100	9.88	100	8.53	100	10.62	100	12.13	100	8.63
MFBs treat people with respect (screening)	50	0.68	59	0.00	63	0.00	56	2.60	60	0.33
MFBs treat people with respect (endline)	100	20.23	100	19.44	100	19.44	100	21.34	100	19.88
Perceptions about women										
Women can run businesses as well as men	100	0.81	100	0.60	100	0.20	100	0.63	100	1.41
Easier for men to receive loans than for women	100	9.88	100	9.52	100	9.62	100	9.62	100	9.04
Women make better financial decisions than men	100	2.50	100	2.38	100	2.00	100	2.72	100	2.61
Intentions										
Plan to apply for loan in next 6 months (screening)	52	0.16	62	0.00	67	0.00	59	0.71	61	0.00
Plan to apply for loan in next 6 months (endline)	100	4.66	100	3.17	100	5.21	100	4.18	100	4.62
Will save money next month (screening)	52	0.00	62	0.00	66	0.00	59	0.00	61	0.00
Will save money next month (endline)	100	4.24	100	3.77	100	4.21	100	4.81	100	3.82
Savings behavior										
Opened account on day of screening	1	0.00	0	0.00	0	0.00	1	0.00	5	0.00
Follow-up with Accion	1	0.00	0	0.00	0	0.00	7	0.00	0	0.00
Plan to follow up with Accion	5	0.00	5	0.00	6	0.00	6	0.00	7	0.00
Saved money last month	100	0.47	100	0.79	100	0.40	100	0.00	100	0.60
Savings relative to income	100	8.57	100	9.13	100	8.22	100	8.79	100	8.84
Savings at MFB	25	0.00	23	0.00	29	0.00	26	0.00	20	0.00
Savings at commercial bank	25	0.00	23	0.00	29	0.00	26	0.00	20	0.00
Borrowing behavior										
Outstanding mortgage loan	100	0.38	100	0.79	100	0.20	100	0.00	100	0.60
Outstanding loan at commercial bank	100	0.30	100	0.20	100	0.20	100	0.00	100	0.80
Outstanding loan at MFB	100	0.25	100	0.20	100	0.40	100	0.00	100	0.60
Loan from money lenders	100	0.30	100	0.00	100	0.20	100	0.21	100	0.60
Supplier credit	100	0.25	100	0.00	100	0.20	100	0.21	100	0.40
Loan from family/friends	100	0.25	100	0.00	100	0.00	100	0.00	100	0.60

Table 2.6: Compliance Table

	Did not	Attended the following screening			
Treatment Assignment	attend	Placebo	Movie	MFB	Movie +MFB
Pure Control	99.0%	0.0%	0.2%	0.4%	0.4%
Control/Placebo	41.0%	57.9%	1.0%	0.2%	0.0%
Movie	38.0%	0.2%	61.5%	0.3%	0.0%
MFB	42.6%	0.3%	0.5%	56.6%	0.0%
MFB + Movie	41.1%	0.0%	0.2%	0.5%	58.3%

Table 2.7: Self-reported exposure to interventions

Exposure variables	Remembered receiving an invitation		Attended the event		Remembered seeing a movie called The Story of Gold		Remembered attending an event in one of the community halls where Accion presented		Correctly identified the message of the movie	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Movie only	0.01	-0.00	0.04	-0.01	0.23***	0.30***	0.02	0.02	0.26***	0.36***
	(0.014)	(0.011)	(0.029)	(0.016)	(0.031)	(0.031)	(0.019)	(0.027)	(0.030)	(0.035)
MFB	-0.00	-0.01	0.01	-0.00	-0.01	-0.01	0.04**	0.06**	-0.02	-0.04
	(0.014)	(0.011)	(0.030)	(0.016)	(0.031)	(0.033)	(0.019)	(0.028)	(0.030)	(0.036)
MFB + Movie	-0.00	-0.00	0.00	0.01	0.21***	0.33***	0.04**	0.07**	0.26***	0.41***
	(0.014)	(0.011)	(0.029)	(0.016)	(0.031)	(0.032)	(0.019)	(0.028)	(0.030)	(0.036)
Observations	*1,976*	*1,259*	*1,975*	*1,259*	*1,974*	*1,258*	*1,974*	*1,259*	*1,979*	*1,261*
R-squared	*0.00*	*0.00*	*0.00*	*0.00*	*0.05*	*0.14*	*0.00*	*0.01*	*0.08*	*0.18*
Controls	*NO*	*NO*	*NO*	*NO*	*NO*	*NO*	*NO*	*NO*	*NO*	*NO*
Restricted Sample	*NO*	*YES*	*NO*	*YES*	*NO*	*YES*	*NO*	*YES*	*NO*	*YES*
Control Mean:	*0.948*	*0.984*	*0.673*	*0.958*	*0.404*	*0.593*	*0.0734*	*0.102*	*0.286*	*0.419*

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 2.8: Financial Literacy Indices

Financial literacy scores		Arithmetic FL Score	Weighted FL Score
		(1)	(2)
Treatments			
Movie		-0.11	-0.14
		(0.075)	(0.112)
MFB		0.04	0.10
		(0.078)	(0.115)
Movie + MFB		-0.05	-0.04
		(0.077)	(0.114)
Gender disaggregated interaction effects (female base)			
Movie		-0.11	-0.12
		(0.088)	(0.130)
MFB		0.10	0.14
		(0.092)	(0.136)
Movie + MFB		-0.09	-0.10
		(0.091)	(0.134)
Gender disaggregated interaction effects (male interaction)			
Male		0.11	0.18
		(0.131)	(0.193)
Male*Movie		-0.03	-0.07
		(0.172)	(0.254)
Male*MFB		-0.18	-0.14
		(0.176)	(0.261)
Male*(Movie + MFB)		0.12	0.19
		(0.173)	(0.257)
p-values	$\delta_1 + \gamma_1 \neq 0$	0.36	0.38
for F-tests	$\delta_2 + \gamma_2 \neq 0$	0.57	0.98
	$\delta_3 + \gamma_3 \neq 0$	0.82	0.65
Observations		1,261	1,254
R-squared		0.14	0.12
Controls		YES	YES
Restricted Model		YES	YES
Control Mean:		5.262	7.556

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 2.9: Perceptions of Microfinance Banks

Trust in MFBs	If I apply to an MFB for a loan my application will be accepted		Taking a loan is too risky for me		I would trust an MFB to keep my money			MFBs treat people with respect		
	Screening	Endline	Screening	Endline	Screening	Endline (agree strongly)	Endline (agree & agree strongly)	Screening	Endline (agree strongly)	Endline (agree & agree strongly)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatments										
Movie	0.06**	0.04	-0.01	0.01	0.15***	0.08**	0.01	-0.05	0.03	0.01
	(0.026)	(0.033)	(0.038)	(0.039)	(0.034)	(0.038)	(0.033)	(0.031)	(0.038)	(0.034)
MFB	0.10***	-0.01	-0.02	0.01	0.26***	0.05	0.01	0.01	0.06	0.02
	(0.027)	(0.034)	(0.039)	(0.041)	(0.035)	(0.039)	(0.034)	(0.032)	(0.040)	(0.035)
Movie + MFB	0.08***	0.05	-0.02	0.01	0.27***	0.08**	0.06*	0.10***	0.10**	0.06*
	(0.027)	(0.034)	(0.039)	(0.040)	(0.034)	(0.039)	(0.033)	(0.031)	(0.039)	(0.035)
Gender disaggregated interaction effects (female base)										
Movie	0.04	0.00	-0.01	-0.01	0.08**	0.06	0.01	-0.08**	-0.00	-0.01
	(0.031)	(0.038)	(0.044)	(0.046)	(0.039)	(0.044)	(0.038)	(0.036)	(0.044)	(0.040)
MFB	0.10***	-0.03	-0.03	-0.01	0.25***	0.07	0.03	-0.01	0.07	0.01
	(0.032)	(0.040)	(0.046)	(0.048)	(0.041)	(0.046)	(0.040)	(0.038)	(0.047)	(0.042)
Movie + MFB	0.08**	0.05	-0.02	-0.02	0.22***	0.12***	0.05	0.07*	0.13***	0.05
	(0.032)	(0.040)	(0.045)	(0.047)	(0.040)	(0.046)	(0.039)	(0.037)	(0.046)	(0.041)
Gender disaggregated interaction effects (male interaction)										
Male	-0.03	-0.02	-0.08	-0.10	-0.17***	0.01	-0.00	-0.09	0.03	-0.02
	(0.046)	(0.057)	(0.066)	(0.068)	(0.058)	(0.066)	(0.057)	(0.054)	(0.066)	(0.059)
Male*Movie	0.06	0.13*	-0.01	0.10	0.28***	0.04	0.03	0.14*	0.11	0.08
	(0.061)	(0.075)	(0.087)	(0.090)	(0.077)	(0.087)	(0.075)	(0.071)	(0.087)	(0.078)
Male*MFB	-0.02	0.07	0.03	0.08	0.04	-0.06	-0.06	0.08	-0.01	0.03
	(0.062)	(0.077)	(0.089)	(0.092)	(0.078)	(0.089)	(0.076)	(0.074)	(0.089)	(0.080)
Male*(Movie + MFB)	0.01	0.00	-0.02	0.11	0.19**	-0.14*	0.02	0.11	-0.09	0.03
	(0.061)	(0.076)	(0.087)	(0.090)	(0.077)	(0.087)	(0.075)	(0.071)	(0.088)	(0.079)
p-values $\delta_1 + \gamma_1 \neq 0$	0.05	0.04	0.81	0.21	0	0.18	0.55	0.36	0.17	0.28
for F-tests $\delta_2 + \gamma_2 \neq 0$	0.11	0.56	0.98	0.14	0	0.93	0.63	0.27	0.48	0.61
$\delta_3 + \gamma_3 \neq 0$	0.1	0.42	0.58	0.25	0	0.77	0.23	0	0.66	0.21
Observations	*1,215*	*1,261*	*1,223*	*1,261*	*1,226*	*1,261*	*1,261*	*1,174*	*1,261*	*1,261*
R-squared	*0.04*	*0.05*	*0.05*	*0.04*	*0.11*	*0.05*	*0.05*	*0.06*	*0.05*	*0.05*
Controls	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*
Restricted Model	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*
Control Mean:	*0.820*	*0.754*	*0.356*	*0.495*	*0.586*	*0.581*	*0.757*	*0.808*	*0.559*	*0.722*

Standard errors in parentheses

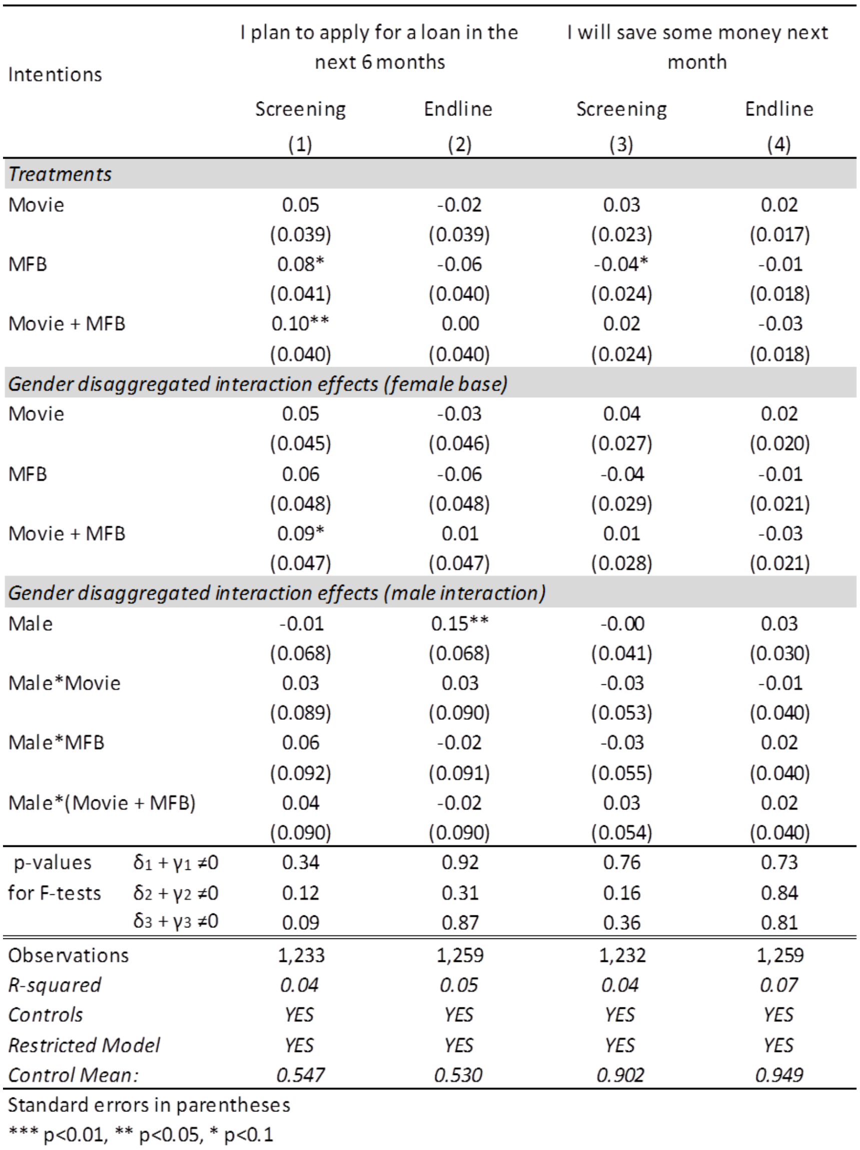
*** p<0.01, ** p<0.05, * p<0.1

Table 2.10: Perception of Female Financial Performance

Perceptions of women at endline	Women can run businesses just as well as men	It is easier for men to receive loans than women	Women make better financial decisions than men
	(1)	(2)	(3)
Treatments			
Movie	-0.00	0.07*	0.05
	(0.020)	(0.038)	(0.030)
MFB	0.00	0.07*	0.04
	(0.020)	(0.039)	(0.031)
Movie + MFB	0.00	0.07*	0.06*
	(0.020)	(0.039)	(0.031)
Gender disaggregated interaction effects (female base)			
Movie	-0.01	0.04	-0.02
	(0.023)	(0.044)	(0.035)
MFB	0.01	0.05	0.01
	(0.024)	(0.046)	(0.037)
Movie + MFB	0.00	0.03	0.01
	(0.024)	(0.046)	(0.036)
Gender disaggregated interaction effects (male interaction)			
Male	-0.13***	0.09	-0.48***
	(0.034)	(0.066)	(0.052)
Male*Movie	0.04	0.13	0.25***
	(0.045)	(0.087)	(0.069)
Male*MFB	-0.04	0.06	0.15**
	(0.046)	(0.088)	(0.071)
Male*(Movie + MFB)	-0.01	0.16*	0.19***
	(0.045)	(0.087)	(0.070)
p-values $\delta_1 + \gamma_1 \neq 0$	0.55	0.03	0
for F-tests $\delta_2 + \gamma_2 \neq 0$	0.54	0.13	0.01
$\delta_3 + \gamma_3 \neq 0$	0.88	0.01	0
Observations	*1,261*	*1,261*	*1,261*
R-squared	*0.09*	*0.08*	*0.19*
Controls	*YES*	*YES*	*YES*
Restricted Model	*YES*	*YES*	*YES*
Control Mean:	*0.936*	*0.342*	*0.751*

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

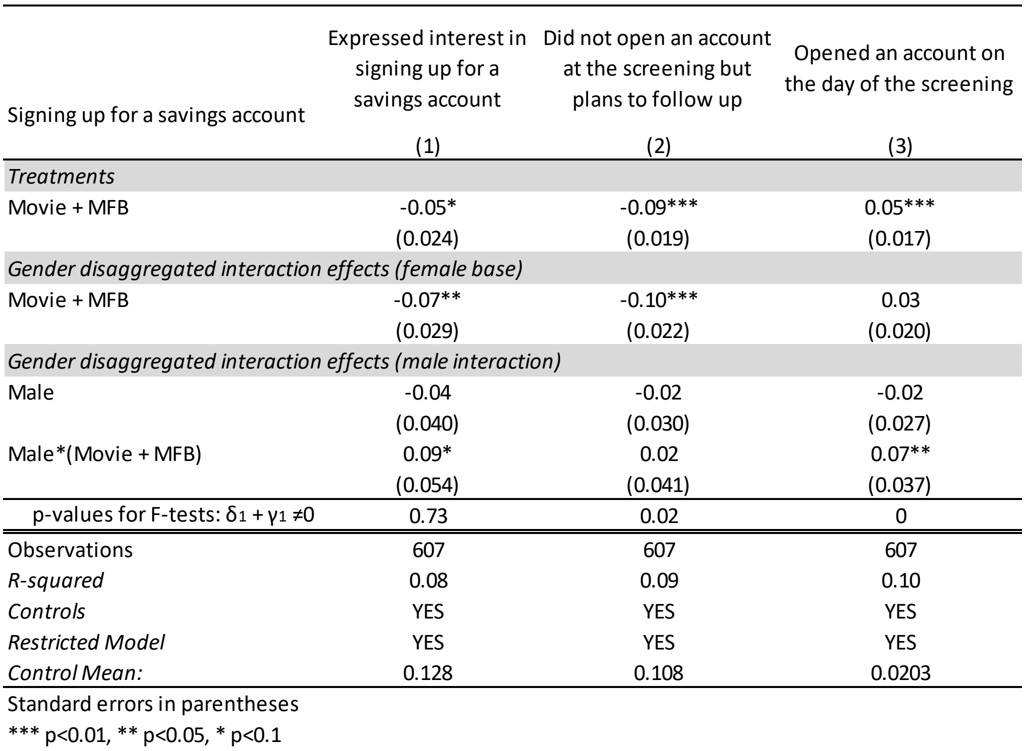
Table 2.11: Intentions



Intentions		I plan to apply for a loan in the next 6 months		I will save some money next month	
		Screening	Endline	Screening	Endline
		(1)	(2)	(3)	(4)
Treatments					
Movie		0.05	-0.02	0.03	0.02
		(0.039)	(0.039)	(0.023)	(0.017)
MFB		0.08*	-0.06	-0.04*	-0.01
		(0.041)	(0.040)	(0.024)	(0.018)
Movie + MFB		0.10**	0.00	0.02	-0.03
		(0.040)	(0.040)	(0.024)	(0.018)
Gender disaggregated interaction effects (female base)					
Movie		0.05	-0.03	0.04	0.02
		(0.045)	(0.046)	(0.027)	(0.020)
MFB		0.06	-0.06	-0.04	-0.01
		(0.048)	(0.048)	(0.029)	(0.021)
Movie + MFB		0.09*	0.01	0.01	-0.03
		(0.047)	(0.047)	(0.028)	(0.021)
Gender disaggregated interaction effects (male interaction)					
Male		-0.01	0.15**	-0.00	0.03
		(0.068)	(0.068)	(0.041)	(0.030)
Male*Movie		0.03	0.03	-0.03	-0.01
		(0.089)	(0.090)	(0.053)	(0.040)
Male*MFB		0.06	-0.02	-0.03	0.02
		(0.092)	(0.091)	(0.055)	(0.040)
Male*(Movie + MFB)		0.04	-0.02	0.03	0.02
		(0.090)	(0.090)	(0.054)	(0.040)
p-values	$\delta_1 + \gamma_1 \neq 0$	0.34	0.92	0.76	0.73
for F-tests	$\delta_2 + \gamma_2 \neq 0$	0.12	0.31	0.16	0.84
	$\delta_3 + \gamma_3 \neq 0$	0.09	0.87	0.36	0.81
Observations		1,233	1,259	1,232	1,259
R-squared		*0.04*	*0.05*	*0.04*	*0.07*
Controls		*YES*	*YES*	*YES*	*YES*
Restricted Model		*YES*	*YES*	*YES*	*YES*
Control Mean:		*0.547*	*0.530*	*0.902*	*0.949*

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.12: Saving account sign up rates



Signing up for a savings account	Expressed interest in signing up for a savings account	Did not open an account at the screening but plans to follow up	Opened an account on the day of the screening
	(1)	(2)	(3)
Treatments			
Movie + MFB	-0.05*	-0.09***	0.05***
	(0.024)	(0.019)	(0.017)
Gender disaggregated interaction effects (female base)			
Movie + MFB	-0.07**	-0.10***	0.03
	(0.029)	(0.022)	(0.020)
Gender disaggregated interaction effects (male interaction)			
Male	-0.04	-0.02	-0.02
	(0.040)	(0.030)	(0.027)
Male*(Movie + MFB)	0.09*	0.02	0.07**
	(0.054)	(0.041)	(0.037)
p-values for F-tests: $\delta_1 + \gamma_1 \neq 0$	0.73	0.02	0
Observations	607	607	607
R-squared	0.08	0.09	0.10
Controls	YES	YES	YES
Restricted Model	YES	YES	YES
Control Mean:	0.128	0.108	0.0203

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

*Note that in this table the treatment being considered is Movie/MFB and the relevant comparison (control) is the MFB only group.

Table 2.13: Savings Behavior

Savings Behavior	Followed up with an MFB after the event	Currently has any form of formal savings account	I saved some money last month	Do you currently have savings of less than or equal to 1 month of income?
	(1)	(2)	(3)	(4)
Treatments				
Movie	0.02***	-0.01	0.02	0.01
	(0.006)	(0.029)	(0.037)	(0.039)
MFB	0.00	-0.04	0.01	0.07*
	(0.006)	(0.030)	(0.038)	(0.040)
Movie + MFB	0.00	-0.04	-0.04	0.02
	(0.006)	(0.030)	(0.038)	(0.040)
Gender disaggregated interaction effects (female interaction)				
Movie	0.02***	0.02	0.05	0.03
	(0.007)	(0.034)	(0.043)	(0.045)
MFB	0.00	-0.05	0.02	0.05
	(0.008)	(0.035)	(0.045)	-0.047
Movie + MFB	0.00	-0.04	0.01	0.03
	(0.008)	(0.035)	(0.044)	(0.047)
Gender disaggregated interaction effects (male interaction)				
Male	0.00	0.07	0.03	0.05
	(0.011)	(0.050)	(0.064)	(0.067)
Male*Movie	-0.02*	-0.09	-0.10	-0.05
	(0.014)	(0.066)	(0.084)	(0.089)
Male*MFB	-0.01	0.02	-0.05	0.05
	(0.015)	(0.068)	(0.086)	(0.091)
Male*(Movie + MFB)	-0.01	-0.01	-0.17**	-0.02
	(0.014)	(0.067)	(0.085)	(0.089)
p-values $\delta_1 + \gamma_1 \neq 0$	0.92	0.18	0.43	0.73
for F-tests $\delta_2 + \gamma_2 \neq 0$	0.82	0.64	0.76	0.17
$\delta_3 + \gamma_3 \neq 0$	0.78	0.38	0.03	0.93
Observations	*1,261*	*1,261*	1,256	*1,261*
R-squared	0.03	*0.34*	0.08	*0.05*
Controls	YES	*YES*	YES	*YES*
Restricted Model	YES	*YES*	YES	*YES*
Control Mean:	0	*0.738*	0.650	*0.415*

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.14: Borrowing Behavior

Borrowing behavior		Taken out a loan in the last 4 months (1)	The loan was from an informal source (2)
Treatments			
Movie		-0.06	-0.02
		(0.039)	(0.070)
MFB		-0.07*	0.07
		(0.040)	(0.070)
Movie + MFB		-0.06	-0.08
		(0.040)	(0.069)
Gender disaggregated interaction effects (female interaction)			
Movie		-0.06	-0.07
		(0.045)	(0.081)
MFB		-0.06	0.05
		(0.047)	(0.081)
Movie + MFB		-0.05	-0.14*
		(0.047)	(0.081)
Gender disaggregated interaction effects (male interaction)			
Male		0.01	-0.11
		(0.067)	(0.121)
Male*Movie		0.01	0.19
		(0.089)	(0.166)
Male*MFB		-0.03	0.11
		(0.091)	(0.161)
Male*(Movie + MFB)		-0.01	0.21
		(0.089)	(0.159)
p-values	$\delta_1 + \gamma_1 \neq 0$	0.5	0.47
for F-tests	$\delta_2 + \gamma_2 \neq 0$	0.25	0.27
	$\delta_3 + \gamma_3 \neq 0$	0.36	0.61
Observations		*1,261*	*410*
R-squared		*0.06*	*0.11*
Controls		*YES*	*YES*
Restricted Model		*YES*	*YES*
Control Mean:		*0.508*	*0.470*

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.15: Descriptive statistics (female)

	Total sample		Control	Movie		MFB		Movie + MFB		Pure control		Means by gender	
Variable	N	Mean	Mean	Mean	P-value	Mean	P-value	Mean	P-value	Mean	P-value	Male	Female
Personal characteristics													
Age of respondent	1642	38.16	38.59	38.38	0.783	38.13	0.547	37.34	0.081*	38.48	0.900	36.79	38.16
Married	1674	0.89	0.89	0.88	0.682	0.90	0.676	0.87	0.428	0.90	0.647	0.72	0.89
Widowed	1674	0.03	0.02	0.04	0.103	0.01	0.321	0.02	0.929	0.03	0.465	0.00	0.03
Single	1674	0.08	0.09	0.07	0.563	0.09	0.998	0.10	0.410	0.07	0.343	0.27	0.08
Muslim	1674	0.34	0.34	0.38	0.341	0.32	0.470	0.37	0.484	0.29	0.192	0.40	0.34
Christian	1674	0.66	0.66	0.62	0.341	0.68	0.524	0.63	0.436	0.71	0.192	0.59	0.66
Can speak English	1667	0.68	0.66	0.64	0.448	0.68	0.561	0.69	0.468	0.72	0.169	0.77	0.68
Igbo	1674	0.17	0.15	0.15	0.971	0.19	0.130	0.18	0.271	0.20	0.050*	0.28	0.17
Yoruba	1674	0.78	0.81	0.80	0.746	0.76	0.086*	0.77	0.178	0.75	0.058*	0.67	0.78
Other ethnicitiy	1674	0.05	0.04	0.05	0.588	0.05	0.512	0.05	0.503	0.05	0.842	0.05	0.05
Education													
No completed school education	1673	0.08	0.08	0.08	0.689	0.09	0.415	0.08	0.856	0.07	0.939	0.06	0.08
Primary school education	1673	0.23	0.25	0.25	0.995	0.20	0.091*	0.22	0.341	0.20	0.107	0.20	0.23
High school diploma	1673	0.48	0.46	0.45	0.936	0.50	0.253	0.48	0.607	0.51	0.178	0.56	0.48
Diploma	1673	0.11	0.12	0.09	0.366	0.11	0.819	0.10	0.625	0.13	0.543	0.09	0.11
Graduate school	1673	0.10	0.10	0.11	0.762	0.10	0.956	0.12	0.486	0.08	0.441	0.09	0.10
Household characteristics													
Household (HH) size	1665	4.63	4.73	4.71	0.861	4.51	0.060*	4.53	0.103	4.68	0.703	4.29	4.63
Number of children below 12 in HH	1644	1.35	1.38	1.38	0.970	1.33	0.591	1.26	0.193	1.41	0.743	1.27	1.35
Number of dependents in HH	1647	2.28	2.32	2.24	0.560	2.25	0.616	2.27	0.716	2.32	0.993	2.82	2.28
Number of dependents outside the HH	1572	1.41	1.23	1.38	0.291	1.58	0.024	1.33	0.477	1.58	0.030**	1.88	1.41
Business characteristics													
Months in operation	1632	96.50	95.77	94.33	0.828	99.27	0.616	101.12	0.447	90.33	0.452	99.58	96.50
Has a savings account	1668	0.54	0.51	0.55	0.307	0.51	0.913	0.54	0.438	0.64	0.001***	0.64	0.54
Keeps written financial records	1662	0.37	0.36	0.36	0.902	0.35	0.968	0.38	0.458	0.39	0.363	0.38	0.37
Operating inside main market	1648	0.30	0.27	0.32	0.147	0.28	0.679	0.29	0.470	0.33	0.082*	0.14	0.30
Number of employees	1672	1.27	1.38	1.31	0.551	1.29	0.482	1.28	0.435	1.02	0.004***	1.86	1.27
Business experience in years	1667	10.49	10.89	10.86	0.956	10.51	0.553	10.08	0.190	10.04	0.219	11.37	10.49

*** p<0.01, ** p<0.05, * p<0.1

Table 2.16: Descriptive statistics (male)

	Total sample		Control	Movie		MFB		Movie + MFB		Pure control		Means by gender	
Variable	**N**	**Mean**	**Mean**	**Mean**	**P-value**	**Mean**	**P-value**	**P-value**	**P-value**	**Mean**	**P-value**	**Male**	**Female**
Personal characteristics													
Age of respondent	672	36.79	35.97	35.53	0.715	37.34	0.231	37.23	0.296	38.35	**0.054***	36.79	38.16
Married	683	0.72	0.73	0.68	0.315	0.77	0.446	0.69	0.438	0.75	0.817	0.72	0.89
Widowed	683	0.00	0.00	0.01	0.349	0.00		0.00		0.02	0.129	0.00	0.03
Single	683	0.27	0.27	0.31	0.376	0.23	0.446	0.31	0.438	0.23	0.585	0.27	0.08
Muslim	682	0.40	0.38	0.45	0.250	0.41	0.617	0.35	0.648	0.40	0.734	0.40	0.34
Christian	682	0.59	0.61	0.55	0.307	0.57	0.469	0.63	0.732	0.60	0.828	0.59	0.66
Can speak English	679	0.77	0.80	0.74	0.318	0.81	0.810	0.77	0.600	0.75	0.375	0.77	0.68
Igbo	682	0.28	0.26	0.23	0.638	0.27	0.824	0.31	0.297	0.33	0.210	0.28	0.17
Yoruba	682	0.67	0.69	0.71	0.662	0.72	0.628	0.60	0.129	0.63	0.367	0.67	0.78
Other ethnicitiy	682	0.05	0.05	0.05	0.991	0.01	**0.072***	0.08	0.315	0.03	0.490	0.05	0.05
Education													
No completed school education	683	0.06	0.02	0.05	0.187	0.06	0.101	0.08	**0.041****	0.09	**0.024****	0.06	0.08
Primary school education	683	0.20	0.21	0.21	0.980	0.22	0.901	0.17	0.435	0.17	0.451	0.20	0.23
High school diploma	683	0.56	0.59	0.55	0.456	0.51	0.165	0.60	0.860	0.56	0.587	0.56	0.48
Diploma	683	0.09	0.10	0.11	0.822	0.10	0.998	0.06	0.186	0.07	0.418	0.09	0.11
Graduate school	683	0.09	0.08	0.07	0.939	0.11	0.299	0.09	0.652	0.11	0.316	0.09	0.10
Household characteristics													
Household (HH) size	678	4.29	4.15	4.23	0.713	4.26	0.620	4.38	0.369	4.45	0.242	4.29	4.63
Number of children below 12 in HH	667	1.27	1.39	1.09	**0.040****	1.24	0.331	1.20	0.238	1.50	0.557	1.27	1.35
Number of dependents in HH	675	2.82	2.79	2.75	0.870	2.78	0.961	2.74	0.849	3.12	0.249	2.82	2.28
Number of dependents outside the HH	641	1.88	2.25	1.86	0.259	1.41	**0.015****	2.06	0.598	1.82	0.258	1.88	1.41
Business characteristics													
Months in operation	678	99.58	106.76	105.08	0.885	91.69	0.175	100.78	0.607	92.53	0.245	99.58	96.50
Has a savings account	682	0.64	0.70	0.62	0.176	0.61	0.144	0.64	0.347	0.62	0.189	0.64	0.54
Keeps written financial records	678	0.38	0.38	0.32	0.342	0.42	0.506	0.36	0.796	0.40	0.693	0.38	0.37
Operating inside main market	676	0.14	0.16	0.12	0.344	0.14	0.674	0.17	0.845	0.14	0.643	0.14	0.30
Number of employees	680	1.86	2.10	1.80	0.309	1.65	0.135	1.67	0.162	2.15	0.901	1.86	1.27
Business experience in years	683	11.37	10.70	10.57	0.883	11.42	0.481	11.48	0.448	13.04	**0.033****	11.37	10.49

*** p<0.01, ** p<0.05, * p<0.1

Table 2.17: Balance across screening participants (female)

	Total		**Control**		**Movie**			**MFB**			**Movie + MFB**		
Variable	**N**	**Mean**	**N**	**Mean**	**N**	**Mean**	**P-value**	**N**	**Mean**	**P-value**	**N**	**Mean**	**P-value**
Personal characteristics													
Age of respondent	894	38.73	231	39.11	240	39.10	0.892	201	38.11	0.313	214	38.44	0.466
Married	908	0.89	233	0.89	245	0.89	0.887	205	0.89	0.887	217	0.89	0.961
Widowed	908	0.03	233	0.03	245	0.05	0.200	205	0.01	0.282	217	0.02	0.641
Single	908	0.08	233	0.08	245	0.06	0.410	205	0.09	0.680	217	0.09	0.830
Muslim	909	0.34	234	0.34	245	0.37	0.438	205	0.31	0.572	217	0.33	0.871
Christian	909	0.66	234	0.66	245	0.63	0.438	205	0.68	0.648	217	0.66	0.953
Can speak English	904	0.69	232	0.69	243	0.65	0.416	205	0.66	0.496	216	0.75	0.178
Igbo	909	0.17	234	0.15	245	0.16	0.826	205	0.21	0.128	217	0.16	0.844
Yoruba	909	0.78	234	0.81	245	0.80	0.696	205	0.74	**0.096***	217	0.78	0.539
Other ethnicitiy	909	0.05	234	0.04	245	0.04	0.704	205	0.05	0.597	217	0.06	0.403
Education													
No completed school education	909	0.07	234	0.06	245	0.09	0.275	205	0.05	0.645	217	0.06	0.995
Primary school education	909	0.22	234	0.25	245	0.24	0.920	205	0.21	0.473	217	0.19	0.192
High school diploma	909	0.47	234	0.45	245	0.45	0.954	205	0.51	0.252	217	0.48	0.606
Diploma	909	0.12	234	0.13	245	0.10	0.398	205	0.13	0.966	217	0.12	0.775
Graduate school	909	0.11	234	0.11	245	0.10	0.785	205	0.09	0.623	217	0.14	0.318
Household characteristics													
Household (HH) size	904	4.57	233	4.63	244	4.70	0.822	203	4.38	0.088	216	4.47	0.299
Number of children below 12 in HH	893	1.32	230	1.39	244	1.35	0.622	200	1.28	0.282	211	1.22	0.122
Number of dependents in HH	896	2.28	229	2.30	243	2.31	0.969	201	2.23	0.761	215	2.29	0.996
Number of dependents outside the HH	851	1.36	222	1.15	225	1.32	0.264	194	1.66	**0.010****	202	1.40	0.151
Business characteristics													
Months in operation	885	96.58	228	94.36	238	100.43	0.444	197	97.27	0.758	215	94.48	0.973
Has a savings account	908	0.55	234	0.53	245	0.58	0.215	205	0.52	0.939	217	0.58	0.262
Keeps written financial records	902	0.38	233	0.37	244	0.36	0.854	203	0.38	0.746	215	0.42	0.257
Operating inside main market	908	0.32	234	0.30	245	0.33	0.354	205	0.31	0.694	217	0.33	0.416
Number of employees	908	1.31	234	1.35	245	1.27	0.545	205	1.28	0.704	216	1.37	0.892
Business experience in years	905	10.85	233	10.87	242	11.40	0.616	205	9.98	0.258	217	10.94	0.952

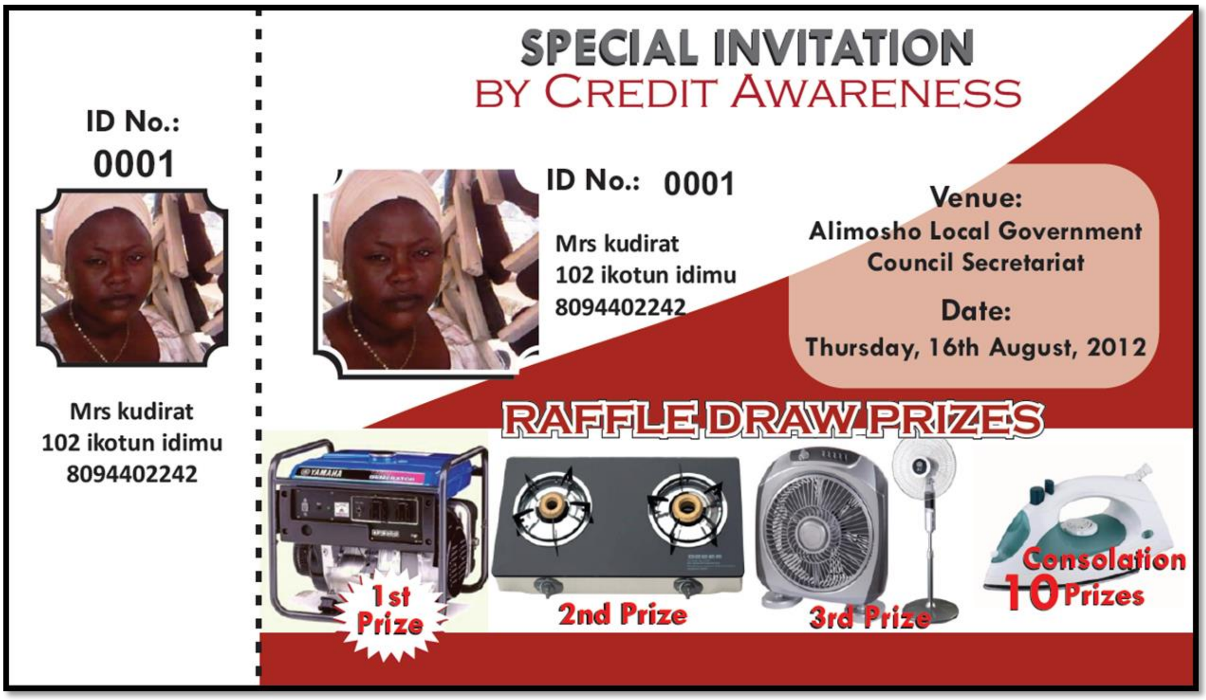
*** p<0.01, ** p<0.05, * p<0.1

Table 2.18: Balance across screening participants (male)

	Total		Control		Movie			MFB			Movie + MFB		
Variable	**N**	**Mean**	**N**	**Mean**	**N**	**Mean**	**P-value**	**N**	**Mean**	**P-value**	**N**	**Mean**	**P-value**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Personal characteristics													
Age of respondent	349	37.08	78	35.23	87	36.68	0.376	86	37.47	**0.095***	93	38.71	**0.013****
Married	352	0.75	79	0.71	88	0.72	0.998	87	0.80	0.152	93	0.77	0.331
Widowed	352	0.00	79	0.00	88	0.01	0.339	87	0.00		93	0.00	
Single	352	0.24	79	0.29	88	0.27	0.872	87	0.20	0.152	93	0.23	0.331
Muslim	352	0.38	79	0.33	88	0.48	**0.038****	87	0.43	0.205	93	0.30	0.695
Christian	352	0.61	79	0.66	88	0.52	**0.057***	87	0.56	0.213	93	0.69	0.678
English	352	0.81	79	0.77	88	0.83	0.399	87	0.83	0.374	93	0.81	0.584
Igbo	352	0.29	79	0.29	88	0.26	0.615	87	0.25	0.582	93	0.34	0.461
Yoruba	352	0.68	79	0.68	88	0.70	0.721	87	0.75	0.367	93	0.59	0.214
Other ethnicitiy	352	0.03	79	0.03	88	0.03	0.721	87	0.00	0.137	93	0.06	0.226
Education													
No completed school education	352	0.04	79	0.03	88	0.03	0.721	87	0.06	0.306	93	0.04	0.531
Primary school education	352	0.20	79	0.20	88	0.22	0.770	87	0.22	0.804	93	0.17	0.611
High school diploma	352	0.57	79	0.62	88	0.53	0.277	87	0.51	0.139	93	0.61	0.922
Diploma	352	0.09	79	0.05	88	0.13	0.162	87	0.09	0.307	93	0.08	0.513
Graduate school	352	0.10	79	0.10	88	0.09	0.863	87	0.13	0.613	93	0.10	0.922
Household characteristics													
Household (HH) size	348	4.39	78	4.06	88	4.43	0.207	86	4.34	0.356	91	4.70	**0.038****
Number of children below 12 in HH	342	1.30	77	1.38	87	1.20	0.301	85	1.24	0.467	88	1.36	0.949
Number of dependents in HH	346	2.95	77	2.87	88	3.07	0.555	86	2.80	0.839	90	3.02	0.646
Number of dependents outside the HH	329	1.92	75	2.52	83	1.87	0.200	80	1.63	**0.065***	86	1.78	0.114
Business characteristics													
Months in operation	350	104.27	79	109.81	87	107.10	0.943	87	91.21	0.208	92	109.64	0.991
Has a savings account	352	0.67	79	0.70	88	0.68	0.895	87	0.62	0.309	93	0.71	0.848
Keeps written financial records	352	0.41	79	0.42	88	0.33	0.297	87	0.47	0.491	93	0.43	0.871
Operating inside main market	352	0.17	79	0.18	88	0.11	0.193	87	0.16	0.781	93	0.22	0.537
Number of employees	352	1.81	79	2.05	88	1.90	0.700	87	1.68	0.399	93	1.66	0.352
Business experience in years	352	10.98	79	11.18	88	9.99	0.241	87	11.55	0.761	93	11.24	0.963

*** p<0.01, ** p<0.05, * p<0.1

Figure 2.1 – Invitation Card



ID No.:
0001

Mrs kudirat
102 ikotun idimu
8094402242

SPECIAL INVITATION
BY CREDIT AWARENESS

ID No.: 0001

Mrs kudirat
102 ikotun idimu
8094402242

Venue:
Alimosho Local Government
Council Secretariat

Date:
Thursday, 16th August, 2012

RAFFLE DRAW PRIZES

1st Prize
2nd Prize
3rd Prize
Consolation 10 Prizes

Annex – Notes on Attrition

Attrition is largest in the pure control group (25.5%) when compared to the control and treatment groups (20.2%). Table 2.4 suggests a random pattern of attrition for the 3 treatment arms when compared to the placebo control group, but a large and significant differential attrition in the pure control group. This differential attrition is reinforced by the balance results from 2.1, and may be resulting from the fact that pure control business owners were only contacted at baseline and follow up, whereas all other groups had another intermediate contact to receive the screening invitation making them (i) more aware of the activities and (ii) easier to track. Given the significantly lower response rate in the pure control group, we subsequently analyze treatment effects by comparing the placebo screening group with the different treatment arms.

When data are analyzed by simply excluding respondents with missing values for any relevant outcome measures (item non-response, or INR), this could again cause biased results if missingness is systematically related to a respondent's potential outcomes. Table 2.5 presents INR rates for main outcomes measures across different treatment and control groups. For instance, for the question of basic understanding of inflation, it can be seen that 100 percent of the surveyed micro-entrepreneurs are asked this question (column 1) and that 2.37 percent of those who are asked do not give a response (column 2). Overall, the data in Table 2.5 indicates that INR for main outcome measures is not a critical issue (most of the times INR rates are <5%) and non-differential across treatment and control groups. Interestingly, INR is the lowest for measures of intentions, savings and borrowing behavior, whereas highest INR rates (between 10 and 20 %) can be observed for questions related to perceptions about MFBs, possibly reflecting cases where business owners have not interacted with MFBs and therefore have not been able to form an opinion. Table 2.5 also reveals a striking increase in INR for the questions of perceptions about MFBs at the end-line survey relatively to the data that was collected shortly after the screening. This increase does not interact with a particular treatment status and may owe to different modes of interviews and the design of the questionnaires: While the short survey right after the screenings was self-administered by attendees, the end-line survey was conducted face-to-face. To avoid unit non-response and potential measurement errors, the self-administered questionnaire was designed to be as simple as possible and only asked dichotomous - Yes or No - type of questions with no explicit "Don't know" or "Refusal" choices. This means that direct comparison over time (eg. through a difference-in-difference approach) would present challenges; however, similar response patterns across treatment groups support the idea that responses are at least internally consistent.

Given the rather low INR rates for most outcome measures and the fact that they are indistinguishable across control and treatment groups, we take no specific measures to address this type of missingness. Nevertheless, we do account for missing data on covariates: In the regression analysis, coefficients of predictors of interest are adjusted by using a procedure advocated by Cohen and Cohen (1985), whereby measures with missing values are replaced by zero and a dummy variable indicating such missing values is included. The logic behind this approach is that the dummy variables adjust the parameters for theoretically relevant predictors by removing variance which can be attributed to missing data that is lurking in the dependent variable (McKnight et al 2007). This also avoids losses in sample size during regression analysis in cases where observations would otherwise be dropped due to missing covariate responses.

3. Management, Supervision, and Health Care: A Field Experiment[62]

3.1 Abstract

If health service delivery is poorly managed, then increases in inputs or ability may not translate into gains in quality. However, little is known about how to increase managerial capital to generate persistent improvements in quality. We present results from a randomized field experiment in 80 primary health care centers (PHCs) in Nigeria to evaluate the effects of a health care management consulting intervention. One set of PHCs received a detailed improvement plan and nine months of implementation support (full intervention), another set received only a general training session, an overall assessment and a report with improvement advice (light intervention), and a third set of facilities served as a control group. In the short term, the full intervention had large and significant effects on the adoption of several practices under the direct control of the PHC staff, as well as some intermediate outcomes. Virtually no effects remained one year after the intervention concluded. The light intervention showed no consistent effects at either point. We conclude that sustained supervision is crucial for achieving persistent improvements in contexts where the lack of external competition fails to create incentives for the adoption of effective managerial practices.

[62] This study was published as a National Bureau of Economic Research (NBER) and Institute for Labor Economics (IZA) working paper (Dunsch et al., 2017). Co-authors are David Evans, Mario Macis, and Ezinne Eze-Ajoku.

3.2 Introduction

In recent years, improving the quality of health care provision – beyond merely making it available – has become a higher priority for the World Health Organization (WHO) and other health agencies (WHO 2006; Institute of Medicine, 2001; Das et al. 2008). Recent research suggests that improvements in outcomes may not always require significant infusions of additional resources. In wealthy economies, a wide dispersion in health outcomes remains after controlling for access, spending and other structural aspects of quality (Chandra et al. 2013; Skinner 2011). The idea that improvements in health care quality can be achieved without increasing the recurrent resources employed can be particularly appealing to resource-constrained developing countries. At the same time, a recent and growing literature suggests that managerial and organizational practices matter greatly for organizational productivity and outcomes (Bloom et al., 2012; Bloom et al., 2013), including in the health care sector (Bloom et al., 2014), and that differences in management practices across organizations and countries account for a large share of the dispersion in productivity not explained by the quantity and quality of the inputs used. In fact, Brun, Karlan, and Schoar (2013) suggest that the lack of managerial and organizational capital may be a key constraint to productivity growth in developing countries. If so, then simply increasing the quantity of inputs may not translate into improved quality of health care: Das and Hammer (2014) find "no correlation between structural inputs and practice quality" across a number of studies, and Das et al. (2012) find that differences in levels of medical training of caregivers account for small or no differences in the quality of provided care. Improving the management of health facilities holds the promise of improving the quality of care and increasing the returns to other inputs.

The empirical literature on the role of "managerial" or "organizational" capital on the quality of health care delivery in developing countries is still scarce and, to our knowledge, limited to hospitals (Bloom et al., 2014). However, the typical first point of access to care in developing countries is primary health centers (PHCs). The expansion of PHCs has been a crucial component of many developing countries' strategies to expand access to care to their populations, especially in rural areas. However, despite the expansion of PHCs, the quality of health care delivery in developing countries remains low (Das and Hammer, 2014; Strasser et al., 2016).

In this paper, we present results from a randomized field experiment conducted in partnership with the Nigerian Federal Ministry of Health (FMOH) to evaluate the effects of a health care management consulting program for public PHCs in six Nigerian states. The FMOH contracted SafeCare, an international agency that specializes in health care quality standards and patient safety in developing-country contexts, to (i) provide a general training session to representatives from the PHCs, (ii) conduct baseline quality assessments at each PHC accompanied by a brief report, (iii) assist the PHCs' staff in formulating improvement plans, and (iv) provide periodical feedback and support toward implementation of the plans for the duration of nine months. The assessment and plans focused on a set of organizational and managerial practices that comprise basic international standards for running primary health care facilities, ranging from the management of human resources, information, and risk, to the organization of the pharmacy and management of the drug inventory.[63]

An independent evaluation of the SafeCare intervention is policy-relevant in its own right, as many countries across Africa – including Ghana, Kenya, Namibia, Tanzania and Uganda – are working with this agency to improve standards of care at primary and secondary health care facilities (SafeCare, 2017). However, our experimental design allows us to go beyond a simple program evaluation. In particular, we distinguish between different mechanisms through which a management consulting intervention can affect practices and outcomes.

Of the 80 facilities included in the study, 24 were randomly assigned to receive the full treatment described above; 24 to receive a light, information-only treatment consisting of (i) a general training session to PHC representatives and (ii) a baseline quality assessment and a brief report highlighting basic improvement areas and actions, but without a detailed improvement plan and without any additional feedback and support; and 32 to a control group. Comparing the full and the light interventions allows us to identify whether the main barriers to improving practices and quality of care are information constraints or implementation constraints. In the first case, the staff lacks knowledge of the appropriate or recommended organizational and operational practices, and providing that information (the light treatment) should improve practices. On the other hand, if the principal barrier to improvement is an implementation constraint – i.e., the staff lack the capacity to implement the changes, whether because of a lack of management ability or a lack of attention due to competing tasks – then information plus continued coaching and monitoring have the potential to improve practices.

To distinguish between management ability and attention, in addition to collecting data periodically during the implementation phase and immediately after its completion, we gathered data one year after the end of the intervention. Results from this long-term evaluation reveal whether the intervention had lasting effects and – importantly – demonstrate the relative importance of implementation support versus monitoring. Persistent impacts would suggest that initial implementation support improved management ability, which endured beyond the period of support. Short-run impacts with no long-run impacts would suggest that a lack of consistent attention to quality improvements is the binding constraint and that ongoing monitoring is key to sustained

[63] SafeCare is an agency created as part of a collaboration between the Joint Commission International based in the US, PharmAccess Foundation of the Netherlands, and the Council for Health Service Accreditation of Southern Africa established to "address issues of poor and limited health care delivered in developing countries."

improvements. Testing the effects of monitoring is particularly important in this public sector context where the lack of competition implies that incentives to adopt superior organizational practices are essentially non-existent for the facilities' officers-in-charge.

Although the ultimate objective of better standards of care is to improve health outcomes, the scale of this program was insufficient to allow us to detect meaningful changes in outcomes such as infant or maternal mortality or infections. Because the focus of the intervention was to improve practices, our main outcome variables of interest relate to the adoption by the PHCs of the recommended organizational standards. We also measured several intermediate outcomes that should be affected by the improved practices, and that are demonstrated to impact health outcomes in other contexts. One organizational standard was to organize drugs and vaccines in the drug storeroom by type, using labels and ordering them by expiration date. An organized pharmacy should reduce the likelihood of stock-outs, improving the PHCs' ability to provide patients with essential drugs and vaccines, thereby improving the recovery chances of sick patients and immunization rates. In our study, we observe how the pharmacy is organized (practice adoption) and stock-outs of essential drugs and vaccines (intermediate outcomes), which are necessary conditions for improvements in actual health outcomes. Another intermediate outcome is the observed cleanliness of the PHC. Finally, we also measured patient experience and satisfaction through patient exit interviews. One of the Nigerian government's goals with this intervention was to encourage more people to seek care in the public PHCs: Higher patient satisfaction might improve the PHC's reputation in the community and thus contribute to increased access.

The full intervention had large and significant effects on the adoption of several organizational practices that were under the direct control of the facilities' staff. These included practices that required a minimal, one-time effort exertion such as displaying posters with hand-washing guidelines or having clearly marked waste bins for different types of waste, but also practices that required moderate and sustained effort such as labeling and organizing drugs in the pharmacy by expiration date or making hand-washing supplies consistently available in the consulting room and in other key areas of the facility. We also detected economically and statistically significant effects on some intermediate outcomes, including cleanliness of toilets and waiting rooms. In contrast, the light intervention had no systematic effects; in most cases, the estimated coefficients were both economically and statistically insignificant, indicating no meaningful differences with the control group.

Because we are considering many outcomes, we perform corrections for Multiple Hypothesis Testing (Anderson 2008; List et al., 2016). Specifically, we combine outcomes into broad indices (z-scores), thereby reducing the number of tests being performed, and we also compute Family

Wise Error Rate-adjusted and False Discovery Rate-adjusted p-values of the individual outcome estimated coefficients. The results are robust to these corrections.

When we measured practices and intermediate outcomes one year after the end of the intervention, however, we found that almost all of these effects had disappeared. Taken together, the two treatments and the short-term and long-term effects indicate that, first, information alone on what practices should be adopted is not sufficient; results are obtained only when detailed information on what changes need to occur is combined with sustained implementation support and monitoring. Second, the lack of long-term effects – despite the fact that about 70% of the core staff who were employed in the PHCs at the time of the intervention were still present one year later – suggests that monitoring during implementation played a crucial role. Third, the results are also informative about the nature of "adjustment costs," which have been emphasized as a reason why organizations are often reluctant to adopt new, more efficient practices (Bloom et al., 2013): the intervention failed to produce sustainable changes, but it did result in measurable changes in practices during the "implementation support" phase; this suggests that adjustment costs might be best viewed and modeled as variable costs rather than one-time fixed costs.

Finally, we found no effects on practices that required substantial additional effort on the part of staff, infrastructure investments, or support from and coordination with government agencies (e.g., consistent access to power). This is not surprising, but it underlines the fact that improved management and organizational processes are insufficient to solve the major infrastructural constraints faced in many PHCs around the world. A lack of incentives may also contribute to explain the absence of effects for organizational practices requiring considerable additional and continued effort on the part of the staff.

Our study makes several contributions. Our main contribution is to the literature on the adoption of organizational and managerial practices with what we believe is the first evidence from primary health care facilities in a developing country context. Although in recent years evidence has accumulated indicating that management practices have important effects on productivity, the mechanisms through which superior practices are adopted and the barriers to their adoption are still poorly understood. The profit motive can explain why managers in market contexts adopt better practices upon learning about them (Bloom et al. 2013). However, in many contexts the profit motive is absent. In the health care sector in particular, public providers play a central role in many countries, often with limited competition from private providers. Our experiment demonstrates whether and how better managerial and organizational practices can be adopted by staff in public health care facilities. Specifically, our design distinguishes between the effects of information, implementation support, and supervision on the adoption of practices in the short term and in the long term. Moreover, we advance the empirical literature on health care quality in developing countries, by providing evidence on the effects of a policy-relevant intervention that several governments, particularly in African countries, are adopting to achieve improvements. Previous literature on improving health care quality examines non-managerial policies—including legal mandates, accreditation and administrative regulations, professional oversight, national and local guidelines, information sharing, and incentive provision—with mixed results (Peabody et al., 2006). Even when existing studies report positive results of interventions aimed at improving organizational and individual performance in adopting standards, they have significant design limitations, often focusing on longitudinal change without a credible control group.[64] This makes interpretation of the results problematic. Moreover, the interventions typically have multiple components without a design that allows for the effects of the various components to be assessed separately. In contrast with the existing health care management literature, the randomized-controlled nature of our study allows clearer causal inferences, and our experimental and data collection design allow us to distinguish the effects of different components of the intervention.

The remainder of the paper is structured as follows. In Section 2 we describe the context and provide details on the SafeCare program. In Section 3 we describe our experimental design and research questions, and in Section 4 we discuss the data and estimation strategy. We present the results in Section 5, where we also perform various corrections for Multiple Hypotheses Testing. In Section 6 we offer our conclusions and discuss policy implications.

3.3 The Nigerian Context and the Program

3.3.1 Health and Health Care in Nigeria

Nigeria has a population of almost 186 million and a per capita income of US$2,178 ($5,867 when adjusted for purchasing power parity). The country's total health expenditures amount to 3.7 percent of GDP. Life expectancy at birth is 53 years. Even though life expectancy has increased in the past decade, it is still 12 years shorter than the average among countries in the same income group (World Bank 2016).[65] The main causes of death in Nigeria are lower respiratory infections

[64] For instance, Berwick (2004) reports on a successful intervention in Peru aimed at improving tuberculosis care by adopting standard practices such as treatment planning, systematic drug supply management, and maintenance of registries. Chakraborti et al. (2000) studied the effect of information, feedback and monitoring on private practitioners' case-management skills for treating sick children in rural India, finding large positive effects on a number of standard procedures.

[65] Nigeria is classified as a "lower middle income" country by the World Bank. Life expectancy at birth for all lower middle income countries is 67.4 years.

(14%), HIV/AIDS (10.4%), malaria (8.7%), diarrheal diseases (6.3%), pre-term birth complications (4.7%) and birth asphyxia and birth trauma (4.3%) (WHO 2015).

In 2013, the under-five mortality rate was about 120 per 1000 live births (WHO 2015). About a quarter of all under-five deaths are accounted for by deaths of newborn babies. The leading cause of under-5 death is malaria (21%), followed by acute respiratory infections (15%), prematurity (12%), birth asphyxia (10%), diarrhea (10%), and neonatal sepsis (5%). In the same year, maternal mortality was 560 deaths per 100,000 live births. Many deaths could be prevented by simple, essential interventions reaching women and children on time, for example with antenatal care, vaccination, and timely diagnosis of treatable infectious diseases such as malaria, pneumonia, diarrhea, and measles. Improved quality of health care delivery at primary health care facilities is one important vehicle to achieve better health outcomes (WHO, 2006).

Nigeria's large population means that it accounts for a large share of total deaths in the African continent and worldwide. For example, in 2013 Nigeria alone accounted for about 14 percent of the total number of maternal deaths, 13 percent of under-five deaths and 10 percent of neonatal deaths worldwide (You et al., 2013). Thus, even small reductions in mortality rates through improvements in the quality of health services could result in large reductions in the absolute number of lives saved. For example, a one percent reduction in the under-five mortality rate would save the lives of about 8,000 children under the age of five every year in Nigeria.

The intervention we evaluate in this paper was part of a broader set of actions implemented by the Nigerian government between 2011 and 2015 with the overarching goal of improving health care access and quality. The Health Strategy and Delivery Foundation (HSDF), a not-for-profit organization, partnered with the FMOH to develop a National Framework for Quality Improvement.[66] The FMOH partnered with the World Bank in the assessment of quality of service across primary, secondary and tertiary facilities nationwide. In addition, the FMOH set an agenda to improve the delivery of primary health care services around the country through its Subsidy Reinvestment and Empowerment Program – Maternal and Child Health component (SURE-P MCH), by improving staffing and upgrading primary health care facilities and increasing usage of MCH services through a conditional cash transfer incentive scheme. Quality improvement of PHCs was part of the national quality strategy across primary, secondary, and tertiary care facilities. Thus, in 2013, the FMOH implemented a management intervention to build local capacity and improve quality of care through the organization SafeCare, in partnership with HSDF.

[66] The HSDF was formerly known as the Saving One Million Lives Initiative.

3.3.2 The SafeCare Program

Formed in 2001, SafeCare is an agency specializing in producing and assessing quality standards specific to resource-constrained public and private health care facilities of all kinds. These include tertiary (teaching) hospitals, referral hospitals, district hospitals, primary health centers (as in our case), basic health centers, and health shops or nurse-driven clinics. SafeCare also offers technical assistance, or consulting, with a focus on building knowledge to guide and facilitate the adoption of quality standards.

The SafeCare standards are grouped in 13 “service elements” in four broad areas: health care organization management, care of patients, specialized services, and ancillary services (Table 3.1). The service elements encompass the entire range of clinical services, including management of human resources, information and risk, logistics and management of medication, and laboratory and facility services, among others. For each service element, SafeCare has developed a set of indicators for specific standards or managerial/organizational practices or actions. The SafeCare standards were accredited by the International Society for Quality in Health Care in March of 2013 (Joint International Commission, 2013). The full set of standards can be found in SafeCare (2015).

The SafeCare program consists of the following five components:

1. General training session: SafeCare conducted an initial 2-day general training session attended by one point person from each PHC. The attendees were trained in standard best practices according to the SafeCare model.
2. Baseline assessment and gap analysis: SafeCare personnel visit each PHC and make a detailed assessment. Specifically, for each of 823 standards in health care organization management, care of patients, specialized services, and ancillary services, SafeCare gives a score to the facility ranging from 5 points (“not compliant, very serious”) to 100 points (“compliant”).[67]
3. Initial feedback: Based on the outcome of the assessment, SafeCare provides a summary of the main gaps that were identified in the facility, highlighting areas where the facility needs to improve. The feedback is communicated to the PHC point person and the PHC’s “officer in charge” (OIC or the “in charge” for short).

[67] The full scoring scale is as follows: 100 if “compliant”, 75 if “partially compliant – mild”, 65 if “partially compliant – moderate”, 55 if “partially compliant – serious”, 45 if “partially compliant – very serious”, 35 if “non compliant – mild”, 25 if “not compliant – moderate”, 15 if “not compliant – serious”, 5 if “not compliant – very serious” (SafeCare 2014).

4. Improvement Plan: In consultation with the facility's staff and personnel from the Federal Ministry of Health, the SafeCare consultants formulate a detailed "quality improvement plan" (QIP) for each PHC. Appendix Table A3.1 lists the standards and actions that were recommended by SafeCare.
5. Implementation Assistance and Feedback: SafeCare personnel provide both remote and in-person assistance and feedback to the PHC staff towards the implementation of the plan. The in-person visits by SafeCare personnel occur every other week for nine months from the introduction of the plan. A staff member of the FMOH also accompanies SafeCare personnel; the staff visited each facility once a week to monitor progress and assist the PHC's staff in the implementation of the improvement plan.

3.4 Experimental Design

3.4.1 Experimental design

To evaluate the program's effects, the assignment of PHCs to the treatment was randomized, and independent data collection took place.[68] The randomized controlled trial involved a total of 80 PHCs, located in 20 hospital catchment areas in 6 states. These facilities were randomly assigned to one of the following experimental conditions:

- Treatment A: The full SafeCare program as described in Section 2, including the general 2-day training session, the initial assessment and feedback, the quality improvement plan, and the implementation support and monitoring for nine months.
- Treatment B: A light version of the SafeCare program, including the general 2-day training session, the baseline assessment and initial feedback, but without improvement plan or implementation support.
- Control: Facilities in the control group did not receive any treatment.

Poor quality of health service delivery could be due to the PHC staff's lack of management training, which would imply that the staff is unaware of the recommended practices (standards) to organize a health care facility. Another possibility is that the staff is aware of how the facilities should be managed and organized, but they lack the capacity (either skill or attention) to implement the practices or to put in place the processes necessary for the practices to be adopted. Treatment A provides both information about what should be done and for the implementation of the practices, whereas Treatment B only provides facilities with information, but not with

[68] Ugo et al. (2016) performed a before-after comparison using the SafeCare assessments and without a control group.

implementation support. Therefore, comparison of the full and the light interventions allows us to identify whether the main barriers to improving practices and quality of care are information constraints or implementation constraints. The implementation assistance includes periodic visits to the PHCs by both SafeCare personnel and by FMOH staff. Thus, this component of the program contains both implementation support and monitoring. Both elements could potentially lead to better outcomes, but through different mechanisms: the implementation support is a form of training, and the monitoring could induce the staff at the PHC to exert additional effort to implement the plan, either because regular monitoring visits keep attention on quality improvements, or out of a concern that failure to do so might be penalized by the FMOH financially or with dismissal.[69] To distinguish between these two channels, in addition to collecting data during and immediately after the intervention, we collected data one year after the end of the intervention. If any process and outcome improvements associated with Treatment A (if any) are simply due to the periodical monitoring, then they are more likely to depreciate once the monitoring ceases; if, however, the improvements are mainly due to the assistance component, then we expect them to be more likely to persist over time.

3.4.2 Selection of states and PHCs

The FMOH selected six states for the intervention in order to achieve representation from each of Nigeria's 6 geopolitical zones: Niger (North Central zone), Bauchi (North East), Kebbi (North West), Anambra (South East), Cross River (South South), and Ekiti (South West). The PHCs selected to receive the intervention, 80 facilities in total, were all facilities included in the SURE-P subsidy program in these states (described in section 3.3).

3.4.3 Baseline PHCs characteristics in participating and non-participating facilities

Although the random assignment of facilities to experimental conditions, coupled with the fact that facilities could not opt out of the intervention, ensures the internal validity of our comparisons, how representative are our participating facilities of primary health care facilities in Nigeria? Facility characteristics are not available for the universe of PHCs in Nigeria; however, our baseline data do provide us with rich data on a number of characteristics of all 474 PHCs that were included in the nationwide subsidies program (SURE-P) described in section 2.1, 80 of which were located in the six states that constitute our study's sample. The comparisons presented in Table 3.2 reveal that on most dimensions, the participating PHCs are similar to the remaining 394 non-participating PHCs. For example, the average number of staff members qualified

[69] There were no formally stated or directly enforced consequences for failure to implement the quality improvements, but attention from superiors can still induce a concern for consequences. Qualitative evidence from Zambia shows that with regular and thorough supervision visits to health centers, health workers "feel pressured to improve performance and also take pride in their recognized accomplishments" (Evans, 2018).

as midwives or nurses is 2.5 in participating facilities and 2.7 in non-participating facilities; 73 percent of the participating PHCs and 74 percent of the non-participating ones have at least one midwife per shift; participating facilities have on average 2.8 beds while non-participating facilities have 3.2 beds; the average total number of health workers is 12.3 in participating facilities and 12.4 in non-participating facilities; 50 percent of the participating PHCs and 58 percent of the non-participating PHCs had developed a "facility workplan" for the current year (prior to the intervention); and both groups of facilities are located on average around 20 km from the referral hospital. Participating and non-participating facilities differ substantially, on average, on some dimensions including the number of registered cases of antenatal care (49 versus 71 cases per month) and the number of deliveries (9 versus 30 deliveries per month), which are explained by the presence of several larger facilities among the non-participating ones.

3.4.4 Assignment of PHCs to treatment and control conditions

Twenty-four of the 80 PHCs were randomly assigned to Treatment A, and 24 were assigned to Treatment B. The number of facilities assigned to Treatments A and B were constrained by FMOH budget limitations. The remaining 32 facilities were assigned to the control condition. For the random assignment, we stratified by state and SURE-P intervention.[70] Table 3.3 shows the distribution of facilities across experimental conditions by state, and Figure 3.1 shows a map with the 6 states and the location of the study's PHCs by experimental condition.

3.5 Data, Baseline comparisons and Estimation methods

3.5.1 Data sources

We use data from existing PHC-level surveys as well as data that we collected specifically for the purposes of this study. There is no facility-level attrition, since all 48 PHCs assigned to the two treatment groups participated in the program and were surveyed.

Baseline data: Baseline pre-intervention data stems from two sources, the Service Delivery Indicators (SDI) from August 2013, and a World Bank data collection exercise that covered all of Nigeria's 500 SURE-P PHCs in September/October 2013. The SDI include data from a facilities questionnaire with general facility information, infrastructure, and availability of equipment,

[70] The randomization of PHCs into the two treatment groups and the control group followed these steps: (1) We assigned a random number to each of the 80 PHCs in our population; (2) These numbers were ranked in ascending order; (3) We ranked these numbers within each hospital cluster; (4) The PHC with the highest random number in each was assigned to Treatment A, the second highest number was assigned to Treatment B, and the third highest number was assigned to the control group. This created groups of 20 for each treatment arm; (5) lastly, the 20 PHCs with the fourth highest numbers were ranked again. Then, the 4 highest numbers were allocated to Treatment A, numbers 5-8 went to Treatment B, and the rest were assigned to the control group. Each hospital cluster was within a single state and SURE-P intervention group. The SURE-P intervention groups included monetary incentives for midwives, non-monetary incentives for midwives, a combination, and a control group.

materials, drugs, and supplies.[71] From the SURE-P baseline data collection, we use information on facility characteristics and staffing details (e.g., number of doctors, nurses, and community-health workers). The SDI and SURE-P data are used to make baseline comparisons and randomization checks, and also as controls in some of the regressions in Section 5 below.

Follow-up data: We implemented six rounds of monthly data collection, the first about two months since the start of the SafeCare program (June 2014), and the last one about one year after its conclusion. This repeated data collection over the course of the intervention improves the statistical power of our tests for actions and outcomes that are not strongly autocorrelated (McKenzie, 2012). Our data collection instrument included three parts. First, we administered a questionnaire to each "officer-in-charge" of the PHC – usually the senior clinic staff member – to collect detailed information on facility practices, staff, inputs, challenges and so on. Second, we employed a facility observation module to check for available infrastructure and equipment, and stockouts of drugs and vaccines. More details on these data will be provided below. Third, we conducted monthly patient exit interviews with about three patients per PHC right after their consultation – with spatial separation from the PHC to ensure confidentiality – to inquire about demographics (e.g., wealth, education, family size), satisfaction with the services rendered, and perceptions about the quality of care. The data collection was carried out by a professional survey firm independent of SafeCare or the Nigerian government. The enumerator visits occurred on dates that were not communicated to the PHCs in advance, and the data were collected electronically using tablets.[72] Questions were read directly from the devices and responses were recorded.

3.5.2 Randomization checks

Consistent with our random assignment of PHCs to experimental conditions, comparisons between the treatment groups show balance at baseline. Formal tests shown in Table 3.4 indicate balance on a number of PHC-level characteristics. With only some exceptions, differences across experimental conditions along a number of facility-level variables tend to be small, and t-tests indicate that they are not statistically significant. Taking into account the relatively small sample size of our treatment groups, (N_A = 24, N_B = 24, C = 32), we performed permutation tests in addition to the standard t-tests (Butar and Park 2008). Specifically, we computed Fisher's exact tests and Wilcoxon ranksum tests with 1,000 permutations. The results again show that the differences

[71] The 5 modules of the SDI are: a. Facility questionnaire: General facility information, infrastructure, availability of equipment, materials, drugs, and supplies. b. Staff roster: Part A: List of all health workers by cadre type; Part B: Administered to 10 randomly selected health workers to measure absenteeism. c. Clinical knowledge assessment: Clinical knowledge using 5 medical vignettes + 2 vignettes for maternal & newborn complications. d. Public expenditure module: Collects receipts and spending (monetary and in-kind) by health facilities. e. Exit module: User satisfaction, socio-demographic characteristics & payments. The SDI data collection included 79 of the 80 clinics in this evaluation. One clinic in Anambra was omitted in the data collection.

[72] The data collection employed Asus Google Nexus 7 tablets with the software "SurveyCTO."

across experimental conditions are in most cases not statistically significant (Table 3.4). This indicates that our randomization has succeeded in creating comparable treatment and control groups.

3.5.3 Estimation methods

We estimate pooled-OLS and ANCOVA models with dummies for each wave of data collection:

$$Y_{i,t} = \beta_0 + \beta_1 T_A + \beta_2 T_B + \beta_3 Y_0 + X_i + \delta_t + \varepsilon_{i,t}$$

$Y_{i,t}$ are the outcome variables (described in the next section), and T_A and T_B indicate whether clinic i is in treatment group A (full treatment) or B (light treatment). Y_0 is the SURE-P or SDI baseline value if available, and δ_t designates survey round fixed effects. X_i designates the stratification dummies including state dummies and SURE-P intervention status.

3.6 Outcome Variables and Results

3.6.1 Outcome variables

The goal of the SafeCare program was to assist the PHCs in adopting a set of organizational practices. The full set of SafeCare standards includes more than 800 indicators. Taken together, these indicators define the "standard" according to which primary health care facilities in resource-restricted settings should be managed. In coordination with the FMOH, we have selected a subset of 75 outcome indicators. We did so prior to the intervention, with the agreement that the research team would collect data on these outcomes independently of the consultants or the government. Our aim was to select a broad range of outcomes in critical managerial and organizational areas and with varying degrees of ease of implementation. In fact, the "standards" (both the full set and the subset on which we focus) vary in whether they are under the control of PHCs' staff, and in the amount of effort required to achieving them.

To organize the analysis, the selected outcomes were classified into three groups: "Within PHC control/Low effort", "Within PHC control/Moderate effort" and "Outside PHC control/High effort". The "Within PHC control/Low effort" outcomes are fully within the control of the PHC staff and require no or minimal additional resources and effort – e.g., displaying posters in the waiting area with hand washing guidelines, malaria symptoms, or a charter of patient rights. The "Within PHC control/Moderate effort" outcomes can be implemented with higher and more sustained effort on the part of staff, but still without any additional support from the local or central government – e.g., ensuring the presence of hand washing materials and keeping the facility clean. Finally, the "Outside PHC control/High effort" outcomes include outcomes that require either substantial additional effort on the part of the staff or significant infrastructure support from the government. For example, one of the SafeCare standards prescribes that each PHC should have uninterrupted access to electricity; however, whether any given PHC is connected to the national power grid is outside the control of local PHC management. Of the 75 selected outcomes, 18 were classified as "Within PHC control/Low effort", 37 indicators were classified as "Within PHC control/Moderate effort", and 20 were classified as "outside PHC control/High effort". The full list of outcomes and their classification are provided in the appendix.

At the time when we selected and classified the outcome variables, we did not yet have access to the Quality Improvement Plans (QIPs) that the 24 Treatment A facilities had received. When we received access to the detailed QIPs, we matched the actions in the QIPs to the variables that we used in our data collection. The actions in the QIPs are fairly broad in their formulation (see the examples in Figure 3.2), and therefore in most cases there were multiple variables from our surveys that would match with an individual QIP action. However, for other QIP actions, there were no variables in our surveys that matched. In total, we matched 46 variables from our surveys to the QIP actions. The FMOH and representatives from the PHCs involved in Treatment A determined who at the PHC was responsible for implementing the suggested improvements. 30 QIP actions were directed at the PHC's officer-in-charge, 7 others were directed at the local government or the federal (SURE-P) program managers, and 9 were aimed at both levels.[73] Changes to be implemented by the federal or local government would be harder (or even impossible) to change by the local staff of the PHC. When we compare our Low/Moderate/High effort-classification with the QIP actions for the variables that could be matched, we observe a large overlap in the classifications, as a large majority of the variables we classified as "Within PHC control/Low effort" or "Within PHC control/Moderate effort" were indeed marked as changes to be implemented by the PHC staff in the QIPs. Specifically, about 80% of our "Within PHC control/Low" and "Within PHC control/Moderate effort" variables were classified by the FMOH as being within the control of the facility staff, and the remaining 20% was classified as being the responsibility of both the staff and the local or federal government; and all of the outcome variables that we classified as "Outside PHC control/High effort" were classified by the FMOH as being outside the control of the PHCs' staff. It is important to note that the SafeCare intervention could in principle have effects also on "Outside PHC control/High effort" practices. In fact, the FMOH was considerably involved in the implementation of the intervention; specifically, FMOH personnel would visit Treatment A facilities periodically, providing monitoring and support during the implementation of the improvement plan.

We also classified indicators according to where they reach the clinical process. Some changes ("process" indicators) focus principally on process but only indirectly affect patient health, such

[73] A detailed list of QIP actions and their corresponding variables in our surveys can be found in Appendix table 3.

as putting up a poster with clinical information. Other changes ("intermediate outcome" indicators) may have a more direct effect on patient health, such as the cleanliness of the facilities and the availability of hand washing materials. Across our 75 measured indicators, we identified 61 that are focused on process and 14 that capture intermediate outcomes. The ultimate goal of this intervention, of course, is to actually improve health outcomes. However, as explained above, given the sample size of the evaluation, implausibly large changes in health outcomes would be required in order to emerge as statistically significant; as such, we focus on the adoption of practices and on intermediate outcomes.

3.6.2 Results

3.6.2.1 Summary of Results

Before presenting our results in detail, we summarize the findings (Table 3.5): Treatment A had a positive and statistically significant effect on 22 of the 75 indicators that we considered, whereas Treatment B had a statistically significant effect on only 3 indicators. When we divide the indicators according to the difficulty of implementation as described above, we observe that the vast majority of the statistically significant effects of Treatment A were obtained for the indicators that were classified as being "Within PHC control/Low effort" (7 out of 18 indicators, or 39%) or "Within PHC control/Moderate effort" (12 out of 37 indicators, or 32%), whereas the Treatment A had a statistically significant effect on only 3 of the 20 "Outside PHC control/High effort" indicators. As for Treatment B, we only find statistically significant differences in 8% (3 out of 37) "Within PHC control/ Moderate effort" indicators.

Looking at "process" versus "intermediate outcome" indicators, we observe that Treatment A resulted in positive, significant changes in 30% of the process indicators (18 out of 61), and 29% of the intermediate outcome indicators (4 out of 14). Treatment B, instead, resulted in significant changes in 5% of process indicators and none of the intermediate outcome indicators.

After describing our detailed regression results below, we perform two exercises to correct for Multiple Hypothesis Testing. First, we construct a small set of indices based on the classification of indicators described above, which reduce greatly the number of tests being performed. Second, we adjust the p-values on the original regressions' coefficients to account for the fact that we are testing a large number of hypotheses.

3.6.2.2 Process Indicators

Management and Leadership (Table 3.6A)

The SafeCare program emphasized certain aspects of facility management, including the need for regular communications between the health center staff. In Table 3.6A we observe that Treatment A clinics increased the likelihood of holding staff meetings in the previous month by 16 percentage points, and reported holding about 0.2 additional meetings in the previous month (marginally significant). By comparison, 67 percent of facilities in the control group reported holding a staff meeting in the last month, and the average number of meetings held in the control facilities was slightly above 1. Both these indicators were classified as “Within PHC control/Moderate effort.” PHCs in Treatment A are also 15 percentage points more likely (statistically insignificant) to report that they are “working towards quality improvement targets”. However, staff did not appear to be more likely to make suggestions for improvement to the officer-in-charge.

Treatment A clinics displayed a 64 percentage point higher likelihood than control facilities of posting an organizational chart on the wall (versus a rate of zero in the control group), an action classified as “Within PHC control/Low effort,” and a 20 percentage point higher likelihood of having a well-organized drug storage area, i.e. with drugs that are labeled and arranged by expiration date (versus a rate of zero in the control group). The latter, an action classified as “Within PHC control/Moderate effort,” is a practice recommended to reduce the likelihood of stock-outs of essential drugs and vaccines. No meaningful (statistically or economically) effects were found for Treatment B.

Patient Rights (Table 3.6B)

Treatment A led to a 63 percentage point increase in PHCs visibly posting a patient rights charter in the waiting area (versus a rate of zero in the control group). However, no effect was found for posters with clinical information, although those started from a much higher baseline of 57 percent. Both of these processes were classified as “Within PHC control/Low effort” actions. The number of ward screens in the facility – an action classified as “Outside PHC control/High effort” – increased for both treatment groups; however, the estimated effect of Treatment A is twice as large as that of Treatment B, and it is statistically significant, whereas the estimated coefficient is insignificant for Treatment B.

Risk Management, Waste Management, Sterilization and Security (Table 3.6C)

Risk management and sterilization processes are core elements of quality of care and patient safety. Treatment A led to a 34 percentage point increase (from a baseline of 16 percent) in the likelihood that facilities designate an individual responsible for infection control. Also, Treatment A facilities were 20 percentage points more likely to have guidelines on waste management, compared to a baseline of zero (significant at the ten percent level). Both these indicators were classified as "Within PHC control/Low effort."

SafeCare also emphasized the separation of medical waste from ordinary waste, as medical waste that is not properly handled and disposed of represents a high risk of infection or injury to health care personnel, as well as a lesser risk to the general public through the spread of micro-organisms from health care facilities into the environment (Windfield and Brooks, 2015). Treatment A led to a 32 percentage point increase in the adoption of clearly marked bins for different types of waste (versus a baseline of 32 percent in the control PHCs), and to a (marginally significant) 17 percentage point increase in the availability of a poster showing waste separation. However, we do not detect effects on medical and other waste actually being disposed of differently, which is a harder to change intermediate outcome indicator (classified as "Within PHC control/Moderate effort") than the relatively low effort processes of putting up posters or marking waste bins. Neither treatment increased the availability of medical gloves or sterilization equipment. We classified the availability of professional sterilization equipment as "Outside PHC control/High effort,“ because the PHCs are dependent on actions by government authorities to provide these tools.

Finally, SafeCare emphasized the importance of using different cleaning devices, such as mops, for the different areas of the clinic, for example to reduce the likelihood of spreading germs from the toilets to the waiting area. Despite this emphasis, we do not observe that the treatments increased usage of different mops, which could have been implemented with some effort ("Within PHC control/Moderate effort"). However, for the clinics that did use different mops, both treatments increased the likelihood that a color-coded system was employed to differentiate the respective mops.

Facility Management Services (Table 3.6D)

We do not observe changes in basic facility infrastructure (e.g., whether the facility has electricity interruptions or clean water available all year), which are of course "Outside PHC control/High effort" actions. So access to power and water were not affected by Treatment A or Treatment B. However, if the facility possessed a generator (which is classified as a "Outside PHC control/High effort" process indicator), Treatment A led to a 26 percentage point increase in the availability of fuel for the generator (a "Within PHC control/Moderate effort action with a baseline of 58% in control PHCs). Note that PHCs did not receive an additional discretionary budget, so additional availability of fuel may imply some community organization.

Human Resources Management (Table 3.6E)

We do not observe changes in any of the indicators related to human resources management. Because the facilities' officers-in-charge do not have resources or authority to hire extra staff or to reward staff performance, there are no differences between the numbers of clinic staff or human resource practices such as performance measurement systems or reward programs. However, some indicators that were classified as "Within PHC con troll/Low effort," namely whether the facility had a written list of all clinical staff and whether they had submitted a request for additional staff, were also unaffected by the treatment.

Primary Health Care Services (Table 3.6F)

The program showed no impacts on intermediate outcome indicators such as the number of antenatal care visits, the number of deliveries at the clinic or the number of deliveries with complications. However, Treatment A facilities are significantly more likely to report Apgar scores for newborns ("Within PHC control/Moderate effort"), an important tool, but neither treatments shows effects on the availability of a partograph ("Within the PHC control/Low effort").[74,75] The treatments also did not affect whether the clinics would keep individual case records ("Within PHC control/Moderate effort").

Critical goals of the quality improvement program were procedures that would improve hygiene and cleanliness. Evidence from other studies demonstrates that handwashing improves health (Ejemot-Nwadiaro et al. 2015; WHO, 2009) and that the provision of handwashing materials can increase handwashing (Kotch et al., 2007; Maury et al., 2000). We find that Treatment A increased the availability of hand washing facilities for patients by 18 percentage points (from a baseline of 42 percent), and both Treatment A and B increased the availability of hand washing facilities for medical personnel, although the baseline in control PHCs in this case was 84 percent. Treatment A also increased the availability of water in the consulting room and the waiting room by 28 percent and 13 percent, respectively (from a baseline of about 30 percent in both cases). We detected no effects on water availability in the bathrooms and the delivery room. All these indicators were classified as "Within PHC control/Moderate effort." Treatment A also had a large impact on the availability of a poster describing hand-washing behavior (which was a "Within PHC control/Low effort action).

3.6.2.3 Intermediate Outcomes

In the 3.7 tables we show the results of our regressions where the dependent variable measures an intermediate outcome. We have two sets of intermediate outcomes: the cleanliness of critical

[74] Apgar is a quick test performed on a baby at 1 and 5 minutes after birth. The 1-minute score determines how well the baby tolerated the birthing process. The 5-minute score tells the doctor how well the baby is doing outside the mother's womb. The Apgar test is done by a doctor, midwife, or nurse. The health care provider examines the baby's breathing effort, heart rate, muscle tone, reflexes, and skin color.

[75] The partograph is a graphical record of the course of labor. Its use can reduce the rate of maternal mortality since abnormal markers in the progress of labor can be identified early on (Asibong et al., 2014).

areas in the facility (Table 3.7A), and the availability of essential drugs and vaccines (Table 3.7B).[76] Specifically, our enumerators took pictures and evaluated the degree of cleanliness of the waiting areas, the toilets, and the bed linens stored at the facility. They also visited the drug storage area in each facility, took pictures, and checked whether unexpired essential drugs and vaccines were available.

Treatment A increased the likelihood that the waiting room is reported to be "very clean" by 13.6 percentage points and the toilets to be perceived as "very clean" by 11 percentage points (measured on a 1-5 Likert scale). Both coefficients are statistically significant at the 10 percent confidence level. We do not detect any significant impacts for Treatment B. These outcomes were classified as "Within PHC control/Moderate effort" outcomes. We detect a 9.8 percentage point increase in the probability that all essential drugs are available due to Treatment A (significant at the 10 percent level), up from a baseline of 15 percent in control facilities, but we find no effect on the availability of vaccines, although the baseline in this case was much higher (88 percent of control PHCs had all essential vaccines available). During 85% of our visits at least one essential drug was out of stock, whereas in 88% of our visits all essential vaccines were available.

3.6.2.4 Patient Experience and Satisfaction

One of the goals of the government was to increase patient satisfaction. There is evidence from elsewhere in Africa that better clinical knowledge is associated with higher levels of patient satisfaction (Leonard 2008; Evans & Tärneberg 2017). As shown in Table 3.8, we find that the treatments had no impact on measures of patient experience and satisfaction. In part, this might reflect the fact that the initial levels of patient satisfaction were high, hovering around the 90% mark.[77] The only significant result is that patients for clinics in treatment group A are slightly more inclined to report that staff spent sufficient time with them.

3.6.3 Multiple Hypothesis Testing

Because we consider a large number of indicators that are potentially affected by the treatments, we are concerned about the possibility of Type I errors (i.e., false positives). In fact, it is well known that the probability of finding a statistically significant effect when the true effect is zero increases sharply with the number of hypotheses being tested (Savin 1984). In our study, the concern is attenuated because if our findings were purely due to Type I errors we would expect a

[76] Drugs defined as essential are Misoprostol, Oxytocin, Magnesium Sulfate (MG), Zinc, Chlorhexidine, Amoxycillin, ORS, ACT, Fansidar/IPT. The essential vaccines are BCG, Penta, Polio, Measles, Yellow Fever, Hepatitis B.

[77] In these same PHCs, we find not only extremely high rates of satisfaction but also evidence of "acquiescence bias," that patients tend to agree with interviewer statements and so satisfaction may be an artifact of positively framed statements (e.g., do you agree or disagree with the statement, "You were satisfied with your service") (Dunsch et al. 2018). Evidence from a larger Nigerian sample shows similarly high levels of satisfaction (Evans & Tärneberg 2017).

roughly similar proportion of positive and significant coefficients for Treatment A and Treatment B, whereas almost all of the statistically significant effects are associated with Treatment A. Nonetheless, we perform various corrections for "multiple hypothesis testing" (MHT) as described below.

There are two main ways to deal with MHT. The first involves aggregating the outcomes into a smaller set of indicators, thereby reducing the number of tests being performed (see – for example – Kling et al., 2007). The second approach consists of applying a statistical correction to the p-values of the estimated coefficients to account for the fact that multiple tests are being performed simultaneously (Family-Wise Error Rate (FWER)-adjusted or False-Discovery Rate (FDR)-adjusted p-values; see Anderson, 2012). We follow both approaches. The first approach is useful because it allows us to answer the question "did the intervention lead to statistically significant changes overall?", which in our context is a meaningful question in particular when we consider our classification of the indicators into groups based on the ease of implementation and the process/intermediate outcome nature of the variables, as defined above. However, the second approach allows us to look at specific process and intermediate outcome indicators, which is important because different indicators vary in their potential ultimate impact on health outcomes (e.g., putting up a poster with patient rights vs. providing hand washing supplies to patients). In other words, as noted by Anderson (2012), these two approaches make different tradeoffs, with the first method reducing the number of tests while avoiding to adjust p-values (which reduces statistical power), and the second adjusting p-values without reducing the number of tests being performed; using both methods balances the tradeoffs of each of them. In total, we conduct 3 tests. Specifically, we construct indices (Kling et al., 2017; Table 3.9A), we utilize an FDR-correction approach (Benjamini et al., 2006; Table 3.9B), and a FWER-correction (List et al., 2016; Table 3.9C).

Indices (Table 3.9A): To build the indices we followed Kling et al. (2007), creating summary indices that aggregate information over several treatment effect estimates. Table 3.9A presents the outcomes grouped in indices following our earlier classification ("Within PHC control/Low effort", "Within PHC control/Moderate effort", "Outside PHC control/High effort", and "Process vs. Outcome"). After allocating each outcome variable to one index, we adjusted the signs so that a positive sign would be always associated with a better outcome for all variables. Next, we demeaned all variables and divided them by the control group's standard deviation, which converted them into normalized effect sizes.[78] Therefore, each element of the index has mean 0 and

[78] For the indices we use only 72 of the 75 indicators. Two indicators had no variation in the control group, and one indicator's coefficient is an extreme outlier ("patients' right charter visibly displayed") which would have distorted the index.

standard deviation 1 for the control group. Lastly, we regressed the index variable on the treatment status to estimate the effect.

We pooled the observations from each PHC into one observation each (column 1; N = 80). Column 2 shows the number of variables that were pooled in the respective index. Columns 4 and 5 show the coefficients for Treatments A and B, measured in standard deviation units. Row 1 shows that Treatment A had large significant effects of 1.66 standard deviation units in Treatment Group A for the "Within PHC control/Low effort" index and 1.28 standard deviation units for the "Within PHC control/Moderate effort" index. This corroborates our earlier findings (see section 5.2.1) as most of the significant effects from individual outcome indicators were found for the "Within PHC control/Low effort" and (to a lesser degree) "Within PHC control/Moderate effort" actions. There were no significant effects for the "Outside PHC control/High effort" index. We also detected strongly significantly effects of Treatment A on the "Process" index (1.45 standard deviation units), which overlaps highly with the "Within PHC control/Low effort" index. The "Outcome" index for Treatment A is marginally significant with a 0.47 standard deviation unit increase.

FDR adjustment (Table 3.9B): The false discovery rate (FDR) was developed as a middle-ground between measures that are considered too restrictive (e.g., the Bonferroni adjustment) and not controlling for multiplicity at all (Benjamini et al., 2006). The false discovery rate (FDR) designates the proportion of null-hypothesis rejections that are type I errors (Anderson 2012; Benjamini & Hochberg 1995). FDR has greater power than FWER (see below), at the cost of allowing a higher rate of type I errors. Using the two-stage step-up FDR-control procedure following Benjamini et al. (2006), we are rejecting a total of 12 of the initially 22 rejected null-hypotheses when using naïve p-values (Table 3.9B).[79] 6 of the 7 initially rejected null-hypotheses in the "Within PHC control/Low effort" group, and 6 of the 12 initially rejected null-hypotheses in the "Within PHC control/Moderate effort" group remain rejected when correcting for the FDR. The fact that a substantial portion of the significant results for Treatment A remain significant after controlling for the FDR corroborates our finding that Treatment A was effective at improving quality of care standards that are within the control of the local PHC staff.

FWER adjustment (Table 3.9C): The family-wise error rate designates the probability that at least one true null hypothesis is rejected (Holm, 1979). To control for the FWER (at 0.05 confidence level), we followed a procedure developed by List et al. (2016), which asymptotically controls for the FWER and incorporates information about the joint dependence structure of the test statistics and therefore is more powerful than the standard procedures developed by Bonferroni (1935) and Holm (1979). When controlling for the FWER, 11 null-hypotheses remain rejected. Due to the

[79] We used the "krieger" Stata command described in Newson et al. (2003), originated from Benjamini et al. (2001).

different methodology utilized, these 11 do not all coincide with the 12 null hypotheses rejected using the FDR-control.[80]

3.6.4 Long-Term Effects

One year after the intervention ended, we gathered a final round of data in order to examine whether the impacts were likely driven by improved management capacity (which would be signaled by persistent effects) or by supervision (non-persistent effects). Only 3 of our 22 rejected null-hypotheses for Treatment A are still significant in our long-run follow up data (round 7): The visible display of a patients' rights charter (β = 0.571***; SE =0.098), clearly marked waste bins of different types of waste (β = 0.308*; SE =0.111), and the availability of an organizational structure chart in the facility (β = 0.557***; SE =0.119).[81] All three of these findings could result from inaction on the part of the staff; they simply did not take down the patients' rights charter, for example. This underscores the notion that the driver of our effects in Treatment A, which was a composite of providing information and support/monitoring, was likely the regular monitoring component.

The lack of sustained effects in the long-run is not due to staff turnover. In fact, 71% of the core staff (doctors, midwives, nurses) that worked at the PHCs in round 1 were still working there through round 7, i.e. one year after the intervention ended (see table 3.10). Retention rates are similar across the three experimental conditions.

3.7 Conclusions

We conducted a randomized field experiment evaluating the effects of a health care management consulting program for primary health care centers in Nigeria. To our knowledge, this is the first randomized controlled study of the effects of management consulting on the adoption of organizational "standards" in primary health care facilities in a developing country context. Moreover, our experimental design allows us to distinguish between information effects, implementation support effects, and supervision effects.

We find that providing a detailed quality improvement plan paired with continuous monitoring and feedback increased the adoption of several standards and processes. The more intensive treatment also led to improvements in some intermediate outcomes, namely those that were within direct control of the PHC staff, such as cleanliness of toilets and waiting rooms and

[80] With the List et al FWER-correction it is not possible to include control variables, which is why the total number of rejected null-hypotheses using uncontrolled p-values is 29 in Table 3.9C, as opposed to 22 when employing control variables (and as we reported in section 5.2.1).

[81] The visible presence of hand washing supplies (β = 0.202*; SE =0.079) and clean storage of bed linens (β = 0.254*; SE =0.106) were marginally significant in the long-run follow-up, but were not significant during rounds 1-6.

availability of hand-washing equipment. These effects, however, essentially disappeared one year after the end of the intervention. Alternatively, merely presenting baseline quality assessments and summary feedback were insufficient to change health care practices.

All of the short-term effects were found for practices that were under the direct control of the PHC staff, and that required minimal or moderate additional effort. The lack of adequate infrastructure and support structures for PHC staff which our data reveal are contributing factors to poor quality of health care provision. For example, many clinics do not have access to the national power grid, and stock-outs of essential drugs are not always promptly replenished. Moreover, the PHC staff seem to lack incentives to implement process improvements that require extra effort and thus are not “free”.

These findings indicate that information alone on what practices should be adopted is not sufficient. That is, there seem to be no minimal interventions that immediately lead to the sustained adoption of modern organizational practices. We find that improvements occur when specific information on practices to be adopted is combined with implementation support. In particular, periodical monitoring of the progress appears to be important for achieving sustained improvements in contexts where the absence of external competition or managerial pay-for-performance fail to create incentives for the adoption of organizational standards. In a context where many health care facilities share the same challenges, a lower-cost alternative to the intervention here may involve a less intensive baseline evaluation but more sustained monitoring.

3.8 References

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American statistical Association*, *103*(484), 1481-1495.

Asibong, U., Okokon, I. B., Agan, T. U., Oku, A., Opiah, M., Essien, E. J., & Monjok, E. (2014). The use of the partograph in labor monitoring: a cross-sectional study among obstetric caregivers in general Hospital, Calabar, Cross River State, Nigeria. *International journal of women's health*, *6*, 873.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289-300.

Benjamini, Y., Krieger, A., & Yekutieli, D. (2001). Two-Staged Linear Step-Up FDR Controlling Procedure, Department of Statistics and Operation Research, Tel-Aviv University, and Department of Statistics, Wharton School, University of Pennsylvania. Technical Report.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 93(3), 491-507.

Berwick, D. M. (2004). Lessons from developing nations on improving health care. *Bmj*, *328*(7448), 1124-1129.

Bloom, N., Sadun, R., & Van Reenen, J. (2012). The organization of firms across countries. *The quarterly journal of economics*, *127*(4), 1663-1705.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., & Roberts, J. (2013). Does management matter? Evidence from India. *The Quarterly Journal of Economics*, *128*(1), 1-51.

Bloom, N., Sadun, R., & Van Reenen, J. (2016). *Management as a Technology?* (No. w22327). National Bureau of Economic Research.

Bloom, N., Sadun, R., & Van Reenen, J. (2014). Does management matter in healthcare. *Boston, MA: Center for Economic Performance and Harvard Business School*.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. Tipografia del Senato.

Butar, F. B., & Park, J. W. (2008). Permutation tests for comparing two populations. Journal of Mathematical Science & Mathematics Education V3, (2), 19-30.

Chakraborty, S., D'Souza, S. A., & Northrup, R. S. (2000). Improving private practitioner care of sick children: testing new approaches in rural Bihar. *Health policy and planning*, *15*(4), 400-407..

Chandra, A., Finkelstein, A., Sacarny, A., & Syverson, C. (2013). Health care Exceptionalism? Productivity and Allocation in the US Health care Sector (No. w19200). National Bureau of Economic
Research.

Das, J., & Hammer, J. (2014). Quality of primary care in low-income countries: facts and economics. *Annu. Rev. Econ.*, *6*(1), 525-553.

Das, J., Holla, A., Das, V., Mohanan, M., Tabak, D., & Chan, B. (2012). In urban and rural India, a standardized patient study showed low levels of provider training and huge quality gaps. *Health Affairs*, *31*(12), 2774-2784.

Das, J., Hammer, J., & Leonard, K. (2008). The quality of medical advice in low-income countries. The Journal of Economic Perspectives, 22(2), 93-114.

Dunsch, F. A., Evans, D. K., Eze-Ajoku, E., & Macis, M. (2017). *Management, Supervision, and Health Care: A Field Experiment* (No. w23749). National Bureau of Economic Research.

Dunsch, F., Evans, D. K., Macis, M., & Wang, Q. (2018). Bias in patient satisfaction surveys: a threat to measuring healthcare quality. *BMJ global health*, *3*(2), e000694.

Ejemot-Nwadiaro, R. I., Ehiri, J. E., Arikpo, D., Meremikwu, M. M., & Critchley, J. A. (2015). Hand washing promotion for preventing diarrhoea. The Cochrane Database of Systematic Reviews, (9), 1–95.

Evans, A. (2018). Amplifying accountability by benchmarking results at district and national levels. *Development Policy Review*, *36*(2), 221-240.

Evans, D.K., & Tärneberg, A.W. (2017). Health Care Quality and Information Failure: Evidence from Nigeria. Unpublished manuscript.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 65-70.

Institute for Health Metrics and Evaluation. (2016). Nigeria Country Profile. Accessed on 7 October, 2019 from http://www.healthdata.org/nigeria

Institute of Medicine. (2001). Crossing the Quality Chasm: A New Health System for the 21st Century. Washington D.C.: National Academy Press.

Jacob, B., Kapustin, M., & Ludwig, J. (2014). Human capital effects of anti-poverty programs: evidence from a randomized housing voucher lottery (No. w20164). National Bureau of Economic Research.

Joint Commission International. (2013). SafeCare Standards receive International ISQua Accreditation. Accessed October 6, 2019 from https://www.jointcommissioninternational.org/safecare-standards-receive-international-isqua-accreditation/

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. Econometrica, 75(1), 83-119.

Kotch, J. B., Isbell, P., Weber, D. J., Nguyen, V., Savage, E., Gunn, E., ... & Allen, J. (2007). Handwashing and diapering equipment reduces disease among children in out-of-home child care centers. *Pediatrics*, *120*(1), e29-e36.

Leonard, K. (2008). Is patient satisfaction sensitive to changes in the quality of care? An exploitation of the Hawthorne effect. Journal of health economics, 27(2), 444-459.

List, J. A., Shaikh, A. M., & Xu, Y. (2016). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 1-21.

Maury, E., Alzieu, M., Baudel, J. L., Haram, N., Barbut, F., Guidet, B., & Offenstadt, G. (2000). Availability of an alcohol solution can improve hand disinfection compliance in an intensive care unit. *American journal of respiratory and critical care medicine*, *162*(1), 324-327.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. Journal of Development Economics, 99(2), 210-221.

Newson, R., & ALSPAC Study Team. (2003). Multiple-test procedures and smile plots. Stata J, 3(2), 109-132.

Peabody, J. W., Taguiwalo, M. M., Robalino, D. A., & Frenk, J. (2006). Improving the quality of care in developing countries.

Peabody, J. W., Gertler, P. J., & Leibowitz, A. (1998). The policy implications of better structure and process on birth outcomes in Jamaica. *Health Policy*, *43*(1), 1-13.

SafeCare. (2014). Standards: SafeCare Advanced Assessment. Accessed October 30, 2015 from http://www.safe-care.org/uploads/Standards%202014%20ADVANCED.pdf

SafeCare. (2015). SafeCare Standards. Accessed October 30, 2015 from http://www.safe-care.org/index.php?page=safecare-standards.

SafeCare. (2017). Government partners. Accessed July 31, 2017 from http://www.safe-care.org/index.php?page=government-partners Accessed

Savin, N. E. (1984). Multiple hypothesis testing. *Handbook of Econometrics*, *2*, 827-879.

Strasser, R., Kam, S. M., & Regalado, S. M. (2016). Rural health care access and policy in developing countries. *Annual Review of Public Health*, *37*, 395-412.

Ugo, O., Ezinne, E. A., Modupe, O., Nicole, S., Winifred, E., & Kelechi, O. (2016). Improving quality of care in primary health-care facilities in rural Nigeria: successes and challenges. *Health services research and managerial epidemiology*, *3*, 2333392816662581.

Windfeld, E. S., & Brooks, M. S. L. (2015). Medical waste management–A review. *Journal of environmental management*, *163*, 98-108.

World Bank. (2016). World Development Indicators. Available at http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators

World Health Organization. (2006). Quality of care: a process for making strategic choices in health systems.

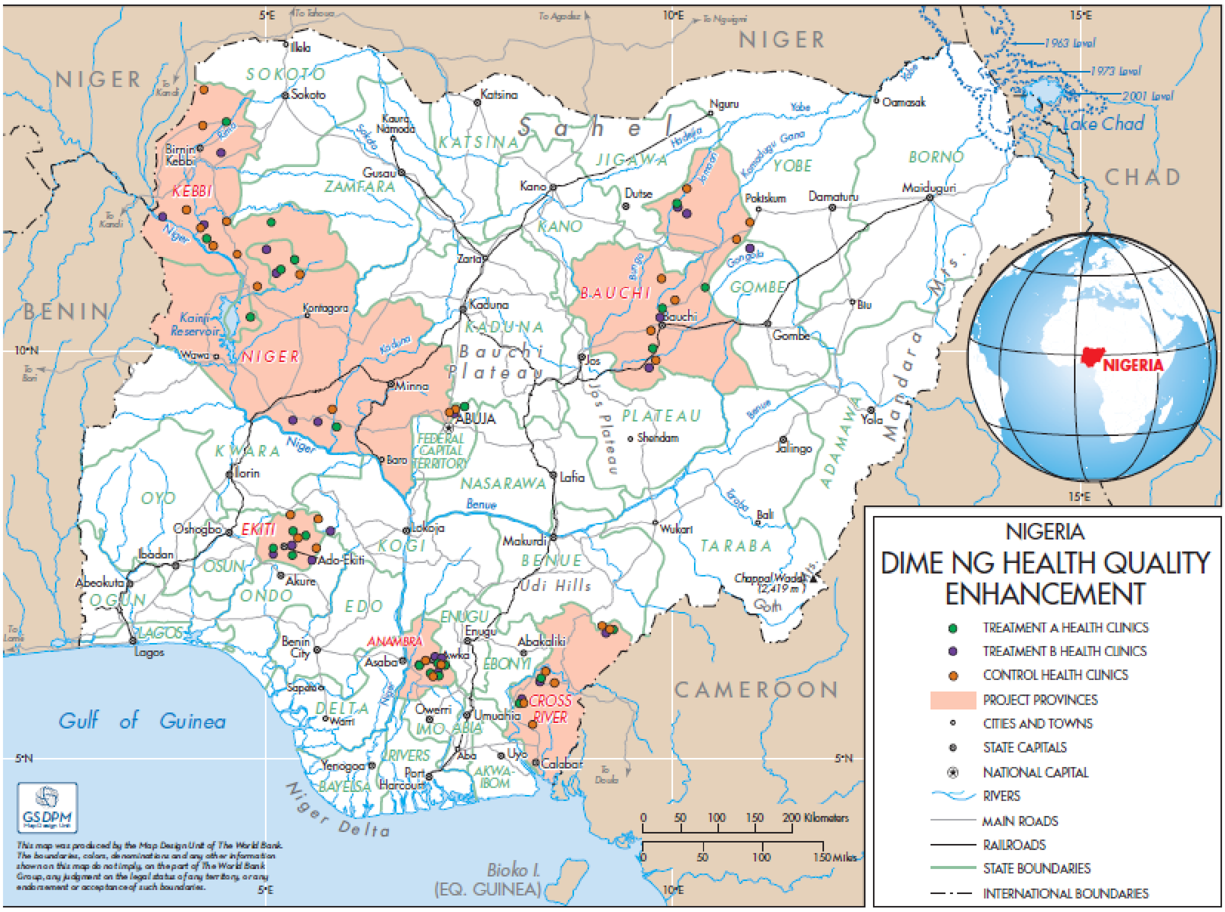
World Health Organization. (2009). WHO Guidelines on Hand Hygiene in Health Care: A Summary.

World Health Organization. (2015). Nigeria: WHO statistical profile.

You, D., Bastian, P., Wu, J., & Wardlaw, T. (2013). Levels and trends in child mortality. Report 2013. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation.

3.9 Figures and Tables

Figure 3.1: Map with Study Sites



Note: The map shows the 6 states where the intervention took place. Treatment A facilities are marked with a green dot, Treatment B facilities are marked purple, and facilities in the control group are orange.

Figure 3.2: Examples of action items from the Quality Improvement Plans and our outcome variables

QIP Example 1: Design an organizational chart or document which describes the lines of authority and account-ability from governance and within the service. (1.1.1.2)

Our variable:

- *Is an organizational structure chart available in the facility?*

QIP Example 2: Ensure the availability of safety boxes and covered dustbins in all areas of the facility for waste collection. Dustbins should have colour coded bin liners or should be painted with the respective colour codes. (5.6.2.4.; 13.3.4.2.; 13.3.4.3.)

Our variables:

- *Are there waste bins in the clinic?*
- *Are the waste bins covered?*
- *Are the waste bins for different types of waste clearly marked? (for example color coded)*

Table 3.1: SafeCare standards categories

Health care organization management	1. Management and leadership 2. Human resource management 3. Patient rights and access to care 4. Management and information 5. Risk management
Care of patients	6. Primary health care services 7. In-patient care
Specialized services	8. Operating theatre and anesthetic services 9. Laboratory services 10. Diagnostic imaging services 11. Medication management
Ancillary services	12. Facility management services 13. Support services

Note: The full list of SafeCare standards can be found from SafeCare, 2019.

Table 3.2: Comparison of participating and non-participating facilities

	(1) **Participating PHCs**	(2) **Non-Participating PHCs**	(3) **p-values**
N. of facilities	79	394	(Participating vs. Non-participating)
Facility Characteristics			
% having 24 hours shift rotation	0.86	0.88	0.67
% having at least one midwife per shift	0.73	0.74	0.97
% having a reception/registration room	0.66	0.72	0.31
number of observation beds	2.77	3.23	0.16
number of days with no electricity/light at all during the last week	4.83	4.74	0.78
distance to the referral facility/hospital (km)	19.18	20.76	0.58
% having transportation for patients	0.10	0.15	0.22
Working Conditions			
number of staff meeting in the past 12 months	8.17	9.41	0.23
% having developed a facility workplan for this year	0.50	0.58	0.21
% having a WDC supervisor	0.95	0.93	0.51
% having a patients feedback mechanism	0.63	0.68	0.47
% having a staff reward system	0.30	0.21	0.08
Human Resources			
number of staff qualified as midwife and nurse	2.54	2.67	0.69
number of staff qualified as midwife only	0.63	0.73	0.55
number of staff qualified as nurse only	0.33	0.31	0.88
number of health workers	12.25	12.35	0.93
Patients			
number of women discharges last week after having given birth	3.99	3.59	0.46
number of registered cases of antenatal care last month	40.05	35.86	0.40
number of registered cases of deliveries last month	6.92	6.54	0.68

Notes: Data are from the 2013 Nigeria SURE-P MCH facilities' survey. The universe consists of the 474 PHCs nationwide that participated in the SURE-P subsidies program (see Section 2 of the paper for details).

Table 3.3: Distribution of PHCs across experimental conditions, by State

State	Total # of PHCs	Treatment A	Treatment B	Control
Anambra	12	5	4	3
Bauchi	16	4	5	7
Cross River	12	3	3	6
Ekiti	12	4	4	4
Kebbi	16	4	4	8
Niger	12	4	4	4
Total	**80**	**24**	**24**	**32**

Notes: This table shows the number of PHCs by State and experimental conditions. A total of 80 facilities were involved in the study.

Table 3.4A: Baseline balance tests – Treatment A vs. Control

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean		**non-Permutation Tests**			**Permutation Tests**		
	Control	*Treatment A*	*T-test*	*Exact*	*Ranksum*	*T-test*	*Exact*	*Ranksum*
Respondent								
age	43.32	45.08	0.35	0.22	0.47	0.39	0.79	0.47
gender	0.48	0.63	0.31	0.41	0.30	0.32	0.69	0.32
Facility Characteristics								
% having 24 hours shift rotation	0.77	0.92	0.16	0.27	0.16	0.16	0.83	0.17
% having at least one midwife per shift	0.65	0.83	0.12	0.14	0.12	0.12	0.91	0.11
% having a reception/registration room	0.61	0.75	0.29	0.39	0.29	0.26	0.76	0.26
number of observation beds	3.13	2.30	0.15	0.57	0.22	0.14	0.45	0.22
number of days with no electricity/light at all during the last week	5.00	5.04	0.96	0.53	0.98	0.96	0.48	0.98
distance to the referral facility/hospital (km)	21.90	16.17	0.27	0.92	0.24	0.24	0.09	0.23
% having transportation for patients	0.10	0.13	0.74	1.00	0.74	0.78	0.32	0.79
Working Condition								
number of staff meeting in the past 12 months	10.77	7.61	0.27	0.19	0.41	0.22	0.81	0.42
% having developed a facility workplan for this year	0.55	0.43	0.41	0.58	0.41	0.42	0.57	0.42
% having a WDC supervisor	0.97	0.88	0.20	0.31	0.19	0.18	0.82	0.22
% having a patients feedback mechanism	0.68	0.63	0.69	0.78	0.69	0.72	0.40	0.72
% having a staff reward system	0.42	0.21	0.10	0.15	0.10	0.11	0.88	0.11
Human Resources								
number of staff qualified as midwife and nurse	1.97	3.08	0.07	0.25	0.07	0.10	0.75	0.07
number of staff qualified as midwife only	0.84	0.50	0.42	0.98	0.80	0.37	0.04	0.79
number of staff qualified as nurse only	0.45	0.38	0.74	0.31	0.95	0.69	0.71	0.93
number of health workers	14.00	12.13	0.33	0.04	0.11	0.28	0.95	0.11
Patients								
number of women discharges last week after having given birth	3.94	3.30	0.46	0.06	0.87	0.56	0.95	0.87
number of registered cases of antenatal care last month	40.38	40.95	0.95	0.18	0.92	0.95	0.84	0.91
number of registered cases of deliveries last month	6.89	7.78	0.65	0.15	0.51	0.65	0.85	0.53

Table 3.4B: Baseline balance tests – Treatment B vs. Control

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean		**non-Permutation Tests**			**Permutation Tests**		
	Control	*Treatment B*	*T-test*	*Exact*	*Ranksum*	*T-test*	*Exact*	*Ranksum*
Respondent								
age	43.32	43.48	0.94	0.80	0.56	0.95	0.24	0.55
gender	0.48	0.54	0.68	0.79	0.67	0.69	0.33	0.69
Facility Characteristics								
% having 24 hours shift rotation	0.77	0.92	0.16	0.27	0.16	0.16	0.82	0.17
% having at least one midwife per shift	0.65	0.75	0.41	0.56	0.41	0.41	0.52	0.43
% having a reception/registration room	0.61	0.63	0.93	1.00	0.93	0.94	0.24	0.94
number of observation beds	3.13	2.74	0.53	0.32	0.62	0.49	0.68	0.65
number of days with no electricity/light at all during the last week	5.00	4.42	0.44	0.48	0.30	0.43	0.53	0.31
distance to the referral facility/hospital (km)	21.90	18.65	0.55	0.90	0.41	0.52	0.12	0.42
% having transportation for patients	0.10	0.08	0.87	1.00	0.86	0.93	0.34	0.94
Working Condition								
number of staff meeting in the past 12 months	10.77	5.33	0.05	0.66	0.03	0.03	0.35	0.04
% having developed a facility workplan for this year	0.55	0.50	0.71	0.79	0.71	0.71	0.29	0.71
% having a WDC supervisor	0.97	1.00	0.38	1.00	0.38	0.61	0.43	0.40
% having a patients feedback mechanism	0.68	0.58	0.48	0.58	0.48	0.51	0.53	0.51
% having a staff reward system	0.42	0.25	0.20	0.26	0.19	0.18	0.78	0.19
Human Resources								
number of staff qualified as midwife and nurse	1.97	2.75	0.24	0.30	0.40	0.28	0.69	0.43
number of staff qualified as midwife only	0.84	0.50	0.40	0.57	0.98	0.38	0.45	0.98
number of staff qualified as nurse only	0.45	0.13	0.13	0.39	0.15	0.11	0.64	0.15
number of health workers	14.00	10.13	0.01	0.51	0.01	0.02	0.49	0.01
Patients								
number of women discharges last week after having given birth	3.94	4.71	0.51	0.16	0.57	0.48	0.85	0.57
number of registered cases of antenatal care last month	40.38	38.78	0.86	0.49	0.49	0.85	0.52	0.49
number of registered cases of deliveries last month	6.89	6.09	0.67	0.98	0.68	0.65	0.03	0.67

Table 3.4C: Baseline balance tests – Treatment A vs. Treatment B

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean		**non-Permutation Tests**			**Permutation Tests**		
	Treatment A	*Treatment B*	*T-test*	*Exact*	*Ranksum*	*T-test*	*Exact*	*Ranksum*
Panel A: Respondent								
age	45.08	43.48	0.50	0.71	0.96	0.47	0.35	0.96
gender	0.63	0.54	0.57	0.77	0.56	0.64	0.55	0.52
Panel B: Facility Characteristics								
% having 24 hours shift rotation	0.92	0.92	1.00	1.00	1.00	1.00	0.46	1.00
% having at least one midwife per shift	0.83	0.75	0.49	0.72	0.48	0.63	0.57	0.47
% having a reception/registration room	0.75	0.63	0.36	0.53	0.36	0.44	0.69	0.36
number of observation beds	2.30	2.74	0.39	0.93	0.39	0.47	0.07	0.39
number of days with no electricity/light at all during the last week	5.04	4.42	0.40	0.39	0.34	0.44	0.63	0.36
distance to the referral facility/hospital (km)	16.17	18.65	0.61	0.48	0.82	0.66	0.54	0.83
% having transportation for patients	0.13	0.08	0.65	1.00	0.64	0.81	0.51	0.76
Panel C: Working Condition								
number of staff meeting in the past 12 months	7.61	5.33	0.12	0.19	0.16	0.39	0.82	0.16
% having developed a facility workplan for this year	0.43	0.50	0.66	0.77	0.66	0.67	0.35	0.66
% having a WDC supervisor	0.88	1.00	0.08	0.23	0.08	0.10	0.98	0.10
% having a patients feedback mechanism	0.63	0.58	0.77	1.00	0.77	0.89	0.32	0.86
% having a staff reward system	0.21	0.25	0.74	1.00	0.73	0.87	0.35	0.67
Panel D: Human Resources								
number of staff qualified as midwife and nurse	3.08	2.75	0.67	0.17	0.43	0.66	0.83	0.44
number of staff qualified as midwife only	0.50	0.50	1.00	0.84	0.78	1.00	0.21	0.78
number of staff qualified as nurse only	0.38	0.13	0.15	0.32	0.13	0.31	0.72	0.14
number of health workers	12.13	10.13	0.20	0.39	0.66	0.30	0.63	0.68
Panel E: Patients								
number of women discharges last week after having given birth	3.30	4.71	0.22	0.76	0.59	0.21	0.25	0.60
number of registered cases of antenatal care last month	40.95	38.78	0.84	0.23	0.62	0.82	0.79	0.64
number of registered cases of deliveries last month	7.78	6.09	0.40	0.21	0.30	0.38	0.80	0.28

Notes for Table 3.4A-C: Nigeria SURE-P MCH Survey Data; Column (1) and (2) present the mean of the indicated group. Column (3) presents p-values from simple T-tests with null hypothesis Treatment A (mean) = Control (mean). Column (4) and (5) present p-values from Fisher's Exact Tests (Exact) and Wilcoxon Ranksum Tests. Column (6), (7) and (8) are p-values from permutated T-tests, Fisher's Exact Tests and Wilcoxon Ranksum Tests with 1000 repetitions. Permutation p-value=number of cases with absolute difference value >= |diff| (real observed one) /number of random permutations performed (reps(1000)).

Table 3.5: Summary of the Results

		Treatment A	
	No. Indicators	No. Significant	%
Results by "Within/Outside PHC control and effort"			
"Within PHC control/Low effort" Index	18	7	39%
"Within PHC control/Moderate effort" Index	37	12	32%
"Outside PHC control/High effort" Index	20	3	15%
In total	75	22	29%
Results by "Process vs Intermediate Outcome"			
"Process" Index	61	18	30%
"Intermediate Outcome" Index	14	4	29%
In total	75	22	29%

Notes: This table reports the number of positive and statistically significant (at least at the 95% confidence level) coefficients on the “Full Intervention” indicator in the OLS regressions whose results are shown in Tables 6 and 7 below.

Table 3.6A: Management & Leadership

	Sample	Model	Obs.	Ctrl Mean	Treatment A	Treatment B	Control level	Process vs. Intermediate outcome
An organizational structure chart available in the facility?	round 6	(1)	80	0.03	0.643*** [0.0977]	-0.048 [0.0477]	Within PHC control /low effort	Process
Any staff meetings held last month?	rounds 1-6	(1)	466	0.67	0.161** [0.0539]	0.026 [0.0558]	Within PHC control /moderate effort	Process
N. meetings held last month	rounds 1-6	(1)	336	1.13	0.169* [0.0732]	-0.013 [0.0687]	Within PHC control /moderate effort	Process
Have a written summary for the most recent meeting last month?	rounds 1-6	(1)	332	0.79	0.077 [0.0443]	-0.020 [0.0569]	Within PHC control /moderate effort	Process
Ever been approached by staff or approached in-charge with suggestions for PHC improvement	rounds 1 and 6	(1)	471	0.25	0.079 [0.0405]	-0.008 [0.0491]	Within PHC control /low effort	Process
Currently working towards any improvement targets?	round 6	(1)	76	0.61	0.152 [0.0997]	-0.098 [0.116]	Within PHC control /low effort	Process
Drugs and vaccines are labeled and organized by expiration date	rounds 1-6	(1)	439	0.03	0.197** [0.0673]	-0.017 [0.0259]	Within PHC control /moderate effort	Process

Notes: This table reports results from Linear Probability Models estimated with Ordinary Least Squares. The regressions include state fixed effects, survey round fixed effects, and SURE-P intervention status. Standard errors are clustered by healthcare facility in regressions including observations from multiple rounds (Model 1). Model 2 includes the baseline value of the outcome variable.

3.6B: Patient Rights

	Sample	Model	Obs.	Ctrl Mean	Treatment A	Treatment B	Within/Out-side PHC control and effort level	Process vs. Intermediate outcome
Is a patient rights charter visibly displayed?	rounds 1-6	(1)	471	0.02	0.632*** [0.0694]	-0.014 [0.0208]	Within PHC control /low effort	Process
Have you put up any posters with clinical information last month?	rounds 1-6	(1)	471	0.57	0.072 [0.0444]	0.036 [0.0364]	Within PHC control /low effort	Process
Number (out of 7) of printed medical issue guidelines available	rounds 1-6	(1)	471	1.56	0.114 [0.0739]	0.110 [0.0913]	Within PHC control /moderate effort	Process
How many ward screens are available throughout the facility?	rounds 1-6	(1)	471	1.74	0.934** [0.346]	0.414 [0.261]	Outside PHC control /high effort	Process

Notes: This table reports results from Linear Probability Models estimated with Ordinary Least Squares. The regressions include state fixed effects, survey round fixed effects, and SURE-P intervention status. Standard errors are clustered by healthcare facility in regressions including observations from multiple rounds (Model 1). Model 2 includes the baseline value of the outcome variable.

3.6C: Risk Management, Waste Management, Sterilization and Security

	Sample	Model	Obs.	Ctrl Mean	Treatment A	Treatment B	Within/Outside PHC control and effort level	Process vs. Intermediate outcome
Flammable materials are clearly labeled	round 6	(1)	80	0.56	-0.079	0.147	Within PHC control /low effort	Process
					[0.110]	[0.0913]		
Are there fire extinguishers (functional)?	rounds 1-6	(1)	468	0.52	0.187*	0.035	Outside PHC control /high effort	Process
					[0.0869]	[0.0981]		
Are there posters showing waste separation in the clinic?	round 6	(1)	80	0.00	0.174*	0.052	Within PHC control/low effort	Process
					[0.0797]	[0.0492]		
Is there a waste bin in the clinic?	rounds 1-6	(1)	471	0.98	0.004	0.005	Within PHC control /moderate effort	Process
					[0.0131]	[0.0162]		
Are the waste bins for different types of waste clearly marked?	round 6	(1)	79	0.32	0.322**	-0.052	Within PHC control/low effort	Process
					[0.106]	[0.0979]		
Medical waste and regular waste are disposed of separately	rounds 1-6	(1)	391	0.28	-0.016	0.006	Within PHC control /moderate effort	Intermediate outcome
					[0.0628]	[0.0643]		
Does this facility have any guidelines on health care waste management?	round 6	(1)	80	0.00	0.208*	0.043	Within PHC control/low effort	Process
					[0.0826]	[0.0444]		
Have you or any provider(s) received training in health care waste management?	round 6	(1)	80	0.09	0.117	0.000	Within PHC control /moderate effort	Process
					[0.106]	[0.0827]		
Are there different mops available for high and low risk areas?	round 6	(1)	80	0.63	0.074	-0.311*	Within PHC control /moderate effort	Process

					[0.129]	[0.127]		
There is color system for these mops	round 6	(1)	47	0.25	0.371**	0.428**	Within PHC control /moderate effort	Process
					[0.119]	[0.147]		
Are there medical gloves available?	round 6	(1)	80	0.97	0.020	-0.021	Within PHC control /moderate effort	Process
					[0.0306]	[0.0581]		
Is there a designated individual responsible for infection control at this facility?	round 6	(1)	80	0.16	0.337**	-0.019	Within PHC control /low effort	Process
					[0.114]	[0.0926]		
Were staff trained on disinfection techniques? (last 6 months)	round 6	(1)	74	0.03	0.176	0.041	Within PHC control /moderate effort	Process
					[0.0953]	[0.0544]		
Are there materials for sterilization of equipment	rounds 1-5	(2)	380	0.90	-0.075	-0.014	Outside PHC control /high effort	Process
					[0.0498]	[0.0593]		
IF YES, is there a functional Autoclave?	rounds 1-5	(1)	346	0.65	0.031	-0.014	Outside PHC control /high effort	Process
					[0.0872]	[0.0720]		
IF YES, is there an electric dry heat sterilizer (functional)	rounds 1-5	(1)	345	0.24	-0.010	-0.075	Outside PHC control /high effort	Process
					[0.125]	[0.110]		
Have contact phone numbers of any external security sources?	round 6	(1)	80	0.28	0.190	-0.081	Within PHC control /low effort	Process
					[0.118]	[0.101]		

Notes: This table reports results from Linear Probability Models estimated with Ordinary Least Squares. The regressions include state fixed effects, survey round fixed effects, and SURE-P intervention status. Standard errors are clustered by healthcare facility in regressions including observations from multiple rounds (Model 1). Model 2 includes the baseline value of the outcome variable.

3.6D: Facility Management Services		Model	Obs.	Ctrl Mean	Treatment A	Treatment B	Within/Outside PHC control and effort level	Process vs. Intermediate outcome
Connected to national power grid?	rounds 1-6	(1)	471	0.74	-0.026	-0.111	Outside PHC control /high effort	Process
					[0.0783]	[0.104]		
Hours connected to national power grid	rounds 1-6	(1)	345	3.34	-0.053	-0.320	Outside PHC control /high effort	Process
					[0.614]	[0.603]		
N. days without electricity interruptions in past two weeks	round 6	(1)	64	3.12	-0.389	0.138	Outside PHC control /high effort	Process
					[1.031]	[0.956]		
N. days without electricity interruptions in past two weeks	round 6	(2)	26	3.12	-0.791	2.003		
					[1.758]	[1.411]		
Have functional generator?	rounds 1-6	(1)	459	0.58	0.066	0.022	Outside PHC control /high effort	Process
					[0.0910]	[0.0906]		
Currently have fuel for the generator?	rounds 1-6	(1)	277	0.58	0.257**	-0.012	Within PHC control /moderate effort	Process
					[0.0932]	[0.117]		
N. days with access to power last week	rounds 1-6	(1)	410	3.76	0.378	0.243	Outside PHC control /high effort	Process
					[0.420]	[0.389]		
Clean water seasonal or available all year?	rounds 1 and 6	(1)	160	0.86	0.056	-0.051	Outside PHC control /high effort	Process
					[0.0650]	[0.0764]		
N. days with access to clean water last week	rounds 1-6	(1)	469	6.56	0.130	-0.149	Outside PHC control /high effort	Process
					[0.189]	[0.212]		
N. days without water supply interruptions in past two weeks	round 6	(1)	75	13.34	-0.132	-0.889	Outside PHC control /high effort	Process
					[0.781]	[1.046]		

Notes: This table reports results from Linear Probability Models estimated with Ordinary Least Squares. The regressions include state fixed effects, survey round fixed effects, and SURE-P intervention status. Standard errors are clustered by healthcare facility in regressions including observations from multiple rounds (Model 1). Model 2 includes the baseline value of the outcome variable.

3.6E: Human Resources Management

	Sample	Model	Obs.	Ctrl Mean	Treatment A	Treatment B	Within/Outside PHC control and effort level	Process vs. Intermediate outcome
Facility has written list of all clinical staff	round 1 and 6	(1)	159	0.83	0.077	-0.035	Within PHC control /low effort	Process
					[0.0618]	[0.0701]		
Facility has enough staff	round 1 and 6	(1)	129	0.14	0.038	0.029	Outside PHC control /high effort	Process
					[0.0845]	[0.0795]		
Has this facility submitted a request for additional staff?	rounds 1 and 6	(1)	144	0.68	-0.089	-0.114	Within PHC control /low effort	Process
					[0.0914]	[0.0892]		
Facility has system for measuring personnel performance	rounds 1 and 6	(1)	153	0.37	-0.029	0.029	Within PHC control /moderate effort	Process
					[0.0740]	[0.0794]		
Facility has system for rewarding personnel performance	rounds 1 and 6	(2)	154	0.54	0.022	-0.019	Outside PHC control /high effort	Process
					[0.0794]	[0.0706]		

Notes: This table reports results from Linear Probability Models estimated with Ordinary Least Squares. The regressions include state fixed effects, survey round fixed effects, and SURE-P intervention status. Standard errors are clustered by healthcare facility in regressions including observations from multiple rounds (Model 1). Model 2 includes the baseline value of the outcome variable.

3.6F: Primary Health Care Services

	Sample	Model	Obs.	Ctrl Mean	Treatment A	Treatment B	Within/Outside PHC control and effort level	Process vs. Intermediate outcome
Pregnancies, Labor and Delivery								
How many antenatal visits did this facility receive last month?	rounds 1-6	(2)	432	96.59	2.395	9.122	Outside PHC control /high effort	Intermediate outcome
					[17.67]	[28.67]		
Keep individual ANC records?	round 6	(1)	67	0.77	0.025	0.078	Within PHC control /moderate effort	Process
					[0.0976]	[0.0779]		
How many deliveries took place at this PHC in the last month?	rounds 1-6	(1)	468	18.93	1.010	5.851	Outside PHC control /high effort	Intermediate outcome
					[3.445]	[5.298]		
N. deliveries without complication/N. deliveries in the PHC	rounds 1-6	(1)	466	0.98	0.007	0.010	Outside PHC control /high effort	Intermediate outcome
					[0.00929]	[0.00808]		
Did the respondent use written records to answer the above questions?	rounds 1-6	(1)	50	0.96	0.054	0.005	Within PHC control /moderate effort	Process
					[0.0833]	[0.0429]		
Is there a partograph available in the facility?	rounds 1 and 6	(2)	151	0.27	0.137	-0.076	Within PHC control /low effort	Process
					[0.0789]	[0.0739]		

IF YES, is it posted visibly?	rounds 1 and 6	(1)	49	0.29	-0.074	0.233	Within PHC control /low effort	Process
					[0.169]	[0.222]		
Of the 10 most recent births records, how many have an "apgar" report?	rounds 1-6	(1)	468	3.55	1.407**	-0.096	Within PHC control /moderate effort	Intermediate outcome
					[0.415]	[0.451]		
Patient Records								
Do you keep individual case records?	round 6	(1)	80	0.81	0.034	0.007	Within PHC control /moderate effort	Process
					[0.0936]	[0.0970]		
Can we look at 5 records now please?	round 6	(1)	67	0.96	0.019	-0.024	Within PHC control /moderate effort	Process
					[0.0430]	[0.0716]		
Average completeness of the 5 patient records	round 6	(1)	65	0.82	0.041	-0.011	Within PHC control /moderate effort	Process
					[0.0393]	[0.0411]		
Keep files for all patients (not just selected or sporadical cases)?	round 6	(1)	67	0.69	0.047	-0.216	Within PHC control /moderate effort	Process
					[0.113]	[0.124]		
Diagnosis and Treatment of Malaria								
Do you have printed guidelines for the treatment of Malaria	rounds 1 and 6	(1)	160	0.95	-0.032	0.043	Within PHC control /moderate effort	Process
					[0.0448]	[0.0366]		
N. cases diagnosed via RDT/N. cases malaria	rounds 1-6	(1)	453	0.789	-0.0854	-0.0265	Within PHC control	Intermediate outcome
					[0.0484]	[0.0400]		

							/moderate effort	
N. cases diagnosed via lab/N. cases malaria	rounds 1-6	(1)	453	0.0461	0.0408	-0.0434*	Within PHC control /moderate effort	Intermediate outcome
					[0.0434]	[0.0197]		
Keep individual malaria records?	round 6	(1)	67	0.23	-0.011	-0.105	Within PHC control /moderate effort	Process
					[0.0923]	[0.0837]		
Hand Washing Guidelines and Equipment								
Is there a hand washing facility for patients?	rounds 1-6	(1)	471	0.42	0.178**	0.114	Within PHC control /moderate effort	Process
					[0.0619]	[0.0847]		
Is there a hand washing facility for medical personnel?	rounds 1-6	(1)	471	0.84	0.132*	0.155**	Within PHC control /moderate effort	Process
					[0.0510]	[0.0546]		
Visible presence of hand washing supplies (soap and water)	rounds 1-6	(1)	438	0.83	0.080	0.014	Within PHC control /moderate effort	Process
					[0.0418]	[0.0613]		
Water available in the consulting room	rounds 1-6	(1)	453	0.29	0.281***	0.122	Within PHC control /moderate effort	Process
					[0.0808]	[0.0885]		
Water available in the bathrooms	rounds 1-6	(1)	393	0.35	0.029	0.066	Within PHC control /moderate effort	Process
					[0.0918]	[0.0805]		
Water available in the waiting room	rounds 1-6	(1)	448	0.32	0.132*	0.007	Within PHC control	Process
					[0.0647]	[0.0695]		

							/moderate effort	
Water available in the delivery room	rounds 1-6	(1)	460	0.84	0.036	-0.005	Within PHC control /moderate effort	Process
					[0.0501]	[0.0421]		
Is there a poster on display describing hand-washing behavior?	round 6	(1)	80	0.16	0.371***	0.006	Within PHC control /low effort	Process
					[0.100]	[0.0812]		

Notes: This table reports results from Linear Probability Models estimated with Ordinary Least Squares. The regressions include state fixed effects, survey round fixed effects, and SURE-P intervention status. Standard errors are clustered by healthcare facility in regressions including observations from multiple rounds (Model 1). Model 2 includes the baseline value of the outcome variable.

Table 3.7A: Cleanliness of Waiting Room, Toilets, and Bed Linens		Model	Obs.	Ctrl Mean	Treatment A	Treatment B	Within/Outside PHC control and effort level	Process vs. Intermediate outcome
Is the waiting room clean?	rounds 1-6							
binary 1=no such room at this facility		(1)	471	0.03	-0.026	-0.002	Within PHC control /moderate effort	Intermediate outcome
					[0.0223]	[0.0367]		
binary 1=very clean		(1)	471	0.26	0.136*	0.027		
					[0.0581]	[0.0535]		
binary 1=clean		(1)	471	0.54	-0.035	0.054		
					[0.0615]	[0.0691]		
binary 1=average		(1)	471	0.14	-0.063	-0.057		
					[0.0335]	[0.0334]		
binary 1=dirty		(1)	471	0.02	-0.008	-0.016		
					[0.0125]	[0.0117]		
binary 1=very dirty		(1)	471	0.01	-0.005	-0.006		
					[0.00516]	[0.00555]		
Are the patient toilet rooms clean?	rounds 1-6							
binary 1=no such room at this facility		(1)	467	0.01	0.082	0.006	Within PHC control /moderate effort	Intermediate outcome
					[0.0505]	[0.0197]		
binary 1=very clean		(1)	467	0.09	0.110*	0.060		
					[0.0453]	[0.0487]		
binary 1=clean		(1)	467	0.32	0.117	0.004		
					[0.0745]	[0.0673]		
binary 1=average		(1)	467	0.29	-0.144**	-0.024		
					[0.0466]	[0.0555]		
binary 1=dirty		(1)	467	0.22	-0.126*	-0.037		
					[0.0487]	[0.0516]		
binary 1=very dirty		(1)	467	0.07	-0.039	-0.008		
					[0.0254]	[0.0391]		
Are the stored bed linens clean?	round 6							
binary 1=not clean		(1)	80	0.19	-0.069	-0.047	Within PHC control /moderate effort	Intermediate outcome
					[0.0810]	[0.0878]		
binary 1=clean		(1)	80	0.53	0.082	0.096		
					[0.112]	[0.133]		

binary 1=no fresh linens available	(1)	80	0.28	-0.013	-0.049
				[0.0968]	[0.115]

Table 3.7B: Availability of Essential Drugs and Vaccines

	Sample	Model	Obs.	Ctrl Mean	Treatment A	Treatment B	Within/Outside PHC control and effort level	Process vs. Intermediate outcome
N. out of 9 essential drugs are available/in stock (*)	rounds 1-6	(1)	431	5.908	0.587*	0.283	Outside PHC control /high effort	Intermediate outcome
					[0.266]	[0.216]		
N. out of 6 essential vaccines are available/in stock (**)	rounds 1-6	(1)	117	4.909	-0.143	0.0418	Outside PHC control /high effort	Intermediate outcome
					[0.332]	[0.348]		
N. out of 9 essential drugs are unexpired/valid (*)	rounds 1-6	(1)	431	5.822	0.635*	0.31	Within PHC control /moderate effort	Intermediate outcome
					[0.260]	[0.211]		
N. out of 6 essential vaccines are unexpired/valid (**)	rounds 1-6	(1)	117	4.886	-0.14	0.0753	Within PHC control /moderate effort	Intermediate outcome
					[0.327]	[0.352]		
Is there a re-order level for drugs?	rounds 1-6	(1)	430	0.703	-0.0411	0.0384	Within PHC control /low effort	Process
					[0.0495]	[0.0493]		
Is there a re-order level for vaccines?	rounds 1-6	(1)	336	0.432	-0.0193	0.0117	Within PHC control /low effort	Process
					[0.0452]	[0.0466]		

Notes: This table reports results from Linear Probability Models estimated with Ordinary Least Squares. The regressions include state fixed effects, survey round fixed effects, and SURE-P intervention status. Standard errors are clustered by healthcare facility in regressions including observations from multiple rounds (Model 1). Model 2 includes the baseline value of the outcome variable. (*) Drugs defined as essential are Misoprostol, Oxytocin, Magnesium Sulfate (MG), Zinc, Chlorhexidine, Amoxycillin, ORS, ACT, Fansidar/IPT. (**) The essential vaccines are BCG, Penta, Polio, Measles, Yellow Fever, Hepatitis B.

Table 3.8: Patient Experience/Satisfaction

	N	Control Mean	Treatment A	Treatment B
Cleanliness of facility.	1,923	0.89	0.0417*	0.02
			[0.0206]	[0.0191]
Waiting time reasonable.	1,922	0.89	-0.00737	-0.03
			[0.0193]	[0.0288]
Staff courteous and respectful of patient.	1,916	0.98	0.0019	-0.01
			[0.00869]	[0.00881]
Staff explained the patient's condition clearly.	1,909	0.96	-0.0171	-0.01
			[0.0134]	[0.0117]
Patient had enough privacy during visit.	1,915	0.81	-0.0157	-0.03
			[0.0270]	[0.0228]
Staff spent sufficient time with patient.	1,924	0.89	0.0193	-0.02
			[0.0145]	[0.0186]
Hours facility open adequate to meet patient needs.	1,851	0.94	-0.014	0.00
			[0.0151]	[0.0114]
Patient trusts the staff's decision about medical treatment.	1,898	0.92	0.00317	0.01
			[0.0107]	[0.0102]

Notes: This table reports results from Linear Probability Models estimated with Ordinary Least Squares. The regressions include state fixed effects, survey round fixed effects, and SURE-P intervention status (see sections 3.4 and 4.4 for details). Standard errors are clustered by healthcare facility.

Table 3.9A – Multiple Hypothesis Testing Correction - Z-Scores (equally-weighted)

	(1)	(2)	(3)	(4)	(5)
				Regression Coefficients and Standard Errors	
	N	*N. (vars)*	*Control Mean*	*Treat-ment A*	*Treat-ment B*
Results by "Within/Outside PHC control and effort"					
"Within PHC control/Low effort" Index	80	15	0.000	1.663***	-0,301
			(1.000)	[0.270]	[0.185]
"Within PHC control/Moderate effort" Index	80	37	0.000	1.277***	0,172
			(1.000)	[0.200]	[0.207]
"Outside PHC control/High effort" Index	80	20	0.000	0.561	0,174
			(1.000)	[0.288]	[0.387]
Results by "Process vs Outcome"					
"Process" Index	80	58	0.000	1.455***	-0,0389
			(1.000)	[0.234]	[0.206]
"Outcome" Index	80	14	0.000	0.471*	0,308
			(1.000)	[0.205]	[0.204]

Note: This table shows estimated coefficients and standard errors from regressions of indexes of the outcome variables (Kling et al. 2007). The estimated coefficients are expressed in standard deviation units. We removed *"Displaying patient right charter/poster"* from the "Low effort" and "process" group here, as this had an extreme high coefficient (as almost all Treatment A clinics put up a poster) and thus highly influenced the indices. The 2 indicators *"Are there posters showing waste separation in the clinic?"* and *"Does this facility have any guidelines on health care waste management?"* show no variation in the control group and were therefore excluded here as well, bringing the total number of indicators used in Table 3.9A to 72 instead of 75.

Table 3.9B – Corrections to Control False Discovery Rate (FDR)

	Number of observations	Number of outcomes	***Count of rejected null – Treatment A vs. Control*** *Controlling FDR*	 *by unadjusted p-Values*
Results by "Within/Outside PHC control and effort"				
"Within PHC control/Low effort"	80	18	6	7
"Within PHC control/Moderate effort"	80	37	7	12
"Outside PHC control/High effort"	80	20	0	3
Total	80	75	13	22
Results by "Process vs Outcome"				
"Process" Index	80	61	10	18
"Outcome" Index	80	14	1	4
Total	80	75	11	22

Notes: This table shows the number of estimated coefficients from Tables 3.6A-F above that remain statistically significant after applying a p-value correction for the "False Discovery Rate" (Anderson 2012).

Table 3.9C – Corrections using List et al. (2016) FWER-Corrections

			Count of rejected null – Treatment A vs. Control	
	Number of observations	Number of outcomes	*by multiplicity adjusted*	*by unadjusted p-Values*
Results by "Within/Outside PHC control and effort"				
"Within PHC control/Low effort"	80	18	2	7
"Within PHC control/Moderate effort"	80	37	5	16
"Outside PHC control/High effort"	80	20	2	3
Total	80	75	9	26
Results by "Process vs Outcome"				
"Process" group	80	61		
"Outcome" group	80	14	2	5
Total	80	75		

Notes: This table shows the number of estimated coefficients from Table 6 above that remain statistically significant after applying the p-value correction proposed by List et al. (2016).

Table 3.10 – Staff Turnover (Doctors, Midwives, Nurses)

	Average number of staff in round 1	Average number of staff that stayed on through round 7 (one year after the end of the intervention)	Retention rate through round 7
Control	5.3	3.84	76%
Treatment A	5.8	3.92	69%
Treatment B	5.7	3.46	67%
Total	5.6	3.75	71%

3.10 Appendix

Appendix Table A3.1: Recommended standards/practices at treated PHCs

Standard	% of treated facilities with standard included in Safecare
Management & Leadership	
Carry out checks on expiry date of all pharmaceutical and laboratory supplies in all areas of the facility. Ensure proper documentation of these checks. Ensure the 'first expired first out' principle is adhered to.	83%
Document the organizational structure from governance and within the facility. Roles and responsibilities should be documented and education provided to all staff on work dynamics (clinical and administrative).	88%
Introduce a quality management system at the facility (form a quality team, appoint quality lead, organize weekly quality team meetings, take minutes, train staff).	83%
Institute effective mechanisms of communication and collaboration which include handover meetings, ward rounds, clinical meetings, quality team meetings, etc. Keep records. Document and implement action plans.	83%
General storage facilities should be secure, adequate, ventilated and well organised putting different groups of items in sections.	79%
Implement a stock management system with definitions of maximum & reorder levels. Records of stock received, distribution to different units and usage should be kept to prevent stock-outs. Ensure continuous monitoring of stock.	79%
Ensure all new supplies (medication, vaccines, kits, consumables, etc.) are checked for expiry, batch number, labels, signs of tampering, potency, completeness, colour, smell etc. Keep records of action taken if required.	63%
A list of all equipment, furniture and supplies at the facility should be available. This list should be dated, signed and updated periodically. A policy guiding this process should be available.	50%
Implement a system that ensures all equipment and supplies are available, properly stored and distributed to all relevant areas of the facility. A list of all equipment and supplies should be available.	21%
Obtain the national treatment guidelines and standing orders to guide all staff in their clinical practice.	21%
Establish an effective sterilization process (with regular testing) and provide the appropriate training for the personnel.	13%
Ensure completion of the bore-hole-overhead tank-facility system(re-install motor) and provide the means of supplying the water to the point of use. Provide Veronica buckets and other hand washing facilities.	8%
Ensure the provision of a minimum of 2 functional sanitary facilities (patient and staff) in the facility.	8%
Strengthen the community involvement process through establishing goals for the WDC and incorporating quality improvement indicators in the performance review for the	4%

Appendix Table A3.1 (continued)

Standard	% of treated facilities with standard included in Safecare
Human Resources Management	
Ensure the provision of the needed staff cadres according to the Minimum Standards for PHCs in Nigeria.	46%
Develop an orientation program for new staff at the facility. Keep appropriate records of program content and those in attendance including their signatures.	38%
Create a mechanism that ensures that at the facility levels, job descriptions are known and facility-level performance measurement is done to inform designation and delegation of duties.	4%
Patient Rights & Access to Care	
Obtain patients rights charter. Display strategically in the facility. Train all staff on the patient's right to privacy during examinations, counselling & provision of information (OPD, wards, pharmacy, laboratory, etc).	88%
Ensure the availability of ward screens in relevant areas of the facility (at least 1 ward screen to 2 beds). Ensure windows in patient interaction areas have drapes. Ensure doors are closed during examinations & counselling.	71%
Management of Information	
Ensure all patient records are standardized, dated, up to date, signed and contain the designation of personnel carrying out the assessment.	83%
Make available a secure cabinet/cupboard for the storage of patient files. Ensure files are neatly arranged according to colour, condition and unique identification number. Implement systems for easy retrieval of records.	67%
Ensure all national and local registers are completely filled with correct information. Designate an individual to oversee this process.	58%
Designate an individual to be responsible for the management of information. Establish policy-guided processes regarding data management and provide personnel education/training for the use of data at the facility level.	8%
Risk Management	
Obtain policy on waste management. Train personnel on waste segregation & appropriate containers for collection. Keep adequate records. Display posters on waste segregation at different areas of the facility.	100%
Designate an individual to be responsible for infection control and ensure the provision of continuous in-service training to all personnel. Retrain staff on disinfection techniques. Keep records of training.	96%

Appendix Table A3.1 (continued)

Standard	% of treated facilities with standard included in Safecare
Risk Management (continued)	
Develop & document a mechanism for summoning the assistance of external sources of security in an emergency (eg. Police, community guards, etc.). Make it known to all personnel. Have available contact details displayed in relevant areas.	92%
Ensure flammable materials (fuel, kerosene, meth spirit, etc.) are clearly labelled & have appropriate signage in its environs. Store these materials in well ventilated rooms or cupboards away from easily combustible materials.	92%
Ensure the availability of safety boxes and covered dustbins in all areas of the facility for waste collection. Dustbins should have colour coded bin liners or should be painted with the respective colour codes.	92%
A colour coded system should be employed for mops & brooms in cleaning different areas of the facility (a designated mop, broom and bucket for the labour room, laboratory, toilets, wards, etc.).	83%
Display posters addressing hand washing at different areas of the facility.	83%
Ensure availability of personal protective equipment (gloves, masks, aprons, boots, googles, e.t.c) for staff in all relevant areas. Ensure that personnel make proper use of personal protective equipment.	83%
Obtain the Government policy for the provision of Post Exposure Prophylaxis. Train staff on the policy and how to access these services.	79%
Develop a process that protects personnel & patients from assault. Ensure staff are aware. Control access to the facility & restricted areas. Display posters on no-tolerance for violence. Ensure no dark areas are within & around the facility.	63%
Ensure access control measures are in place at the pharmacy, laboratory, labour room and other restricted areas. Ensure doors are lockable and have appropriate signage eg. "authorized entry only", "restricted area", etc.	17%
Make provision for more waste bins in the facility.	13%
Guiding/supporting rails should be fitted for all staircases and along the high corridors.	13%
Provision for fire fighting equipment should be made. Staff should be trained on how to use these equipment and regular servicing of fire-fighting equipment should be done.	4%

Appendix Table A3.1 (continued)

Standard	% of treated facilities with standard included in Safecare plan
Primary Health Care Services	
Obtain national guidelines for the treatment of Malaria. Ensure that the management of malaria accords with national guidelines. Keep appropriate records of cases receiving ACT following a laboratory confirmation.	100%
Obtain patients rights charter. Display strategically in the facility. Train all staff on the patient's right to privacy during examinations, counselling & provision of information (OPD, wards, pharmacy, laboratory, etc).	88%
Ensure the use of partograph to monitor all deliveries and keep records of apgar score for newborns. Ensure all tests, observations and examinations are recorded for all antenatal and postnatal cases.	88%
Ensure the provision of soap, water and paper towels/single use towels at hand washing facilities. Water should be distributed to relevant areas of the facility with the use of buckets with tap heads (veronica buckets).	83%
Make provision for a delivery table with stirrups.	79%
Make arrangements within the community for a patient transport system. Document this system and make it known to all personnel. Contact telephone numbers should be available and functional.	71%
Create a check-list of parameters and patients that require early attention and document the system for identifying and fast-tracking these patients.	71%
Obtain a referral policy from the local/state government or SURE-P. Ensure policy includes the cases to be referred, services to be referred, a list of referral centers, and details of contact persons in the referral centers.	63%
Develop a health education plan for the facility's patient population. Have a standardized method of keeping records of health education provided to each patient.	54%
Make standing orders for CHOs/CHEWs and JCHEWs available. Make LSS and MLSS guidelines available at the facility.	50%
Make provision for an angle-poise lamp for adequate lighting in the delivery room.	38%
Supply the SURE-P ANC patients files to provide a template for proper records.	21%
Provision should be made for at least 2 security personnel who can run daily shifts, covering the facility round the clock.	21%
Put in place a system to identify newborns (eg. use of wristbands). Display posters reminding mothers not to leave their babies unattended to. Ensure only authorized access to the wards.	17%

Appendix Table A3.1 (continued)

Standard	% of treated facilities with standard included in Safecare
Primary Health Care Services (continued)	
Ensure regular supply of all essential drugs and family planning consumables to prevent stock outs.	17%
Make available facilities and equipment for the testing of malaria.	13%
Ensure care providers write a summary of care provided to each patient whilst on admission in the facility as well as follow-up instructions.	4%
Provide perimeter fencing and ensure a lockable gate is in place.	4%
Repairs of the dilapidated sanitary facilities (toilets and bathroom) for staff and patients.	4%
Rehabilitation of the staff quarters to solve the space constraints in the clinic area and renovation of all dilapidated structures in the facility.	4%
In-Patient Care	
Make provision for more ward screens in all relevant areas (ward, examination room etc.)	50%
Operating Theatre & Anaesthetics	
Ensure the availability and use of autoclaves for sterilization of all instruments. Calico and sterility tapes should be available for the sterilization process.	100%
Make available a secure and well ventilated storage area for sterilized instrument packs. These should be stored off the ground.	50%
Ensure a clear flow and dermacation of activities in the sterilization area (decontamination, washing, drying, packing, sterilizing and storage).	50%
Obtain a storage drum/ container for disinfected instruments.	50%
Laboratory Services	
Designate an individual (with documented job descriptions) to manage the laboratory. Ensure there are policy-guided processes that foster collaborative work between the other units and the laboratory.	13%
Medication Management	
Institute a system that tracks adverse drug reactions (immunization/medication) for patients. Records with details of preventive and remedial actions taken should be kept in registers and patient records as appropriate.	54%
Develop and implement a system for the disposal of expired stock. Records of all expired stock should be kept as well as method of disposal. Expired stock should be separated from all other stock and appropriately labeled.	21%
Designate an individual (with documented job descriptions) for medication management. Ensure there are policy-guided processes that foster collaborative work between the other units and the pharmacy.	17%

Appendix Table A3.1 (continued)

Standard	% of treated facilities with standard included in Safecare
Facility Management Services	
Ensure the provision of a regular source of power supply. Ensure that a back-up system for power supply is available and functional.	92%
Make provision for a reliable and safe source of water supply to this facility. Ensure that there is a back-up source of water in case of contamination or failure.	83%
Ensure all the identified structural defects in the facility (torn mosquito netting, damaged doors and windows, etc) are fixed. Establish a facility maintenance process.	54%
Ensure all sources of electricity are functional and provision is made for the regular supplies of needed fuel. For each shift, a designated individual should be available who oversees this function.	13%
Ensure all construction debri and broken furniture which are no longer useful are kept neatly in an area of the facility, cordoned off, and arrangements put in place to clear them out of the facility.	13%
Provide mosquito nets for all the windows and external doors in the facility.	4%
Support Services	
Ensure the availability of bed linen at this facility and secure storage facilities for these.	100%
Provision should be made for the secure storage of cleaning materials and equipment (mops, brooms, buckets, etc.). Chemicals for cleaning should be kept in a dedicated and secure cabinet clearly labelled for the purpose.	96%
Make available a schedule for emptying waste from the facility to the pit as well as a schedule for burning waste in the pit. Ensure implementation of these schedules.	83%
Construct waste disposal pit with a parapet and cover and train personnel in the use of	79%
Provide training on appropriate cleaning methods, frequency of cleaning & specialized cleaning of infectious areas to all housekeeping staff. Ensure all brooms & mops are properly cleaned & dried before storage.	67%
Provision should be made for at least 2 cleaners who will be responsible for the daily cleaning of the facility, and should be guided by written service-related policies and procedures.	21%

Appendix Table A3.2: Classification of standards by control and effort requirement

Within PHC control / Low effort (18)	
Organizational structure chart available	Currently working towards quality improvement targets
Staff providing suggestions for improvement	Patients’ rights charter visibly displayed
Have any posters been put up with clinical information last month	Flammable materials are clearly labeled
Posters put up showing waste separation	The waste bins for different types of waste are clearly marked
The facility has guidelines on health care waste management	There is a designated individual responsible for infection control
Contact phone numbers of external security sources are available	Written list of all clinical staff available
Facility submitted a request for additional staff	A partograph is available in the facility
The partograph is posted visibly (if available)	There is a poster on display describing hand-washing behavior.
There is a re-order level for drugs	There is a re-order level for vaccines
Within PHC control / Moderate effort (37)	
Any staff meetings held last month	Number of staff meetings held last month
Written summary available for the most recent meeting	Drugs and vaccines are labeled and organized by expiration date
Number (out of 7) of printed medical issue guidelines available	There is a waste bin in the clinic
Medical waste and regular waste are disposed of separately	Staff has received training in waste management
Different mops are available for high and low risk areas	There is a color system for the mops
Medical gloves are available	Staff were trained on disinfection techniques (last 6 months)
There is currently fuel for the generator available	Facility has system for measuring personnel performance
The PHC keeps individual ANC records.	Written records were used to answer questions about the number of deliveries and antenatal visits at the facility
Number of births where an “apgar” score was recorded (last 10 births)	Individual case records are kept at the PHC
Availability of individual case records were visibly confirmed by the enumerator (5)	Average completeness of the 5 patient records
Files are kept for all patients (not just selected ones)	Printed guidelines for malaria treatment available
Number of cases of malaria diagnosed via RDT/number of cases of malaria diagnosed	Number of cases of malaria diagnosed via lab/number of cases of malaria diagnosed

The facility keeps individual malaria records	There is a hand washing facility for patients
There is a hand washing facility for medical personnel	Hand washing supplies (soap and water) are visibly present
Water available in the consulting room	Water available in the waiting room
Water available in the bathrooms	Water available in the delivery room
Cleanliness of the waiting room	Cleanliness of the patient toilet rooms
Cleanliness of the stored bed linens	All essential drugs are unexpired/valid
All essential vaccines are unexpired/valid	
Outside PHC control / High effort (20)	
Number of ward screens available	Fire extinguishers are functional
There are materials for sterilization of equipment	A functional autoclave is available
A functioning electric dry heat sterilizer is available	The facility is connected to the national power grid
Average number of hours connected to the national power grid	Number of days without electricity interruptions in the past two weeks
The facility has a functional generator	Number of days without access to power last week
Clean water available all year	Number of days without access to clean water last week
Number of days without water supply interruptions in past two weeks	Facility has enough staff
Facility has system for rewarding personnel performance	Number of antenatal visits last month
Number of deliveries that took place at this facility last month	Number of deliveries without complications/number of deliveries in the PHC
All essential drugs are available	All essential vaccines are available

Appendix Table 3.3: Classification of standards by control and effort requirement

This table lists those Quality Improvement Plan (QIP) action items that line up with questions in the main survey for this impact evaluation ("IE survey question") and the SURE-P survey ("SURE-P survey question"), as well as the individual considered to be responsible for implementing the action ("Responsible"). The numbers listed after each QIP Action link back to SafeCare's full list of quality standards.

Category	Sub-category	QIP Action	Responsible	IE survey question	SURE-P survey question
Management-leadership	Organizational Chart	Design an organizational chart or document which describes the lines of authority and accountability from governance and within the service. (1.1.1.2)	Officer in charge (OIC)	Is an organizational structure chart available in the facility?	
	Quality Management – Communication	Introduce a quality management system in the facility (appoint quality manager, train staff, organize bi-weekly quality team meetings, keep minutes of these meetings). (1.3.1.2) Institute effective mechanisms of communication and collaboration which include handover meetings, ward rounds, clinical meetings, quality team meetings, etc. Keep records. Document and implement action plans. (1.3.1.2.)	OIC	Last month, were any staff meetings held at this facility?	
				How many meetings were held?	
				In minutes, what was the duration of the last meeting? (only for meetings held LAST MONTH)	
				Do you have a written summary for the most recent meeting last month?	
Drugs and vaccines stock management	Supply of drugs	Ensure regular supply of all essential drugs and family planning consumables to prevent stock outs. (6.8.4.3, 6.8.4.4, 6.6.1.3)	SURE-P	stockout of essential drugs/vaccines	
	Expiry checks	Carry out checks on expiry date of all pharmaceutical and laboratory supplies in all areas of the facility. Ensure proper documentation of these checks. Ensure the 'first expired first out' principle is adhered to. (1.2.6.8.; 9.3.1.9.; 11.5.1.7.)	OIC	drugs/vaccines expiration date	
				Is there an expiration date on the vial?	
				Expiration date: BCG	

	New supplies	Ensure all new supplies (medication, vaccines, kits, consumables, etc.) are checked for expiry, batch number, labels, signs of tampering, potency, completeness, colour, smell etc. Keep records of action taken if required. (1.2.6.4.)	OIC	Check the VVM (vaccine vial monitor) and record the stage	
	Stock management	Implement a stock management system with definitions of maximum and reorder levels. Records of stock received should be kept as well as records of distribution to different units of the facility.(1.2.6.4, 9.3.1.10, 11.5.1.2, 11.5.1.8)	OIC	Is there a re-order level for vaccines?	
				Is there a re-order level for drugs?	
	Expired stock disposal	Develop and implement a system for the disposal of expired stock. Records of all expired stock should be kept as well as method of disposal. Expired stock should be separated from all other stock and appropriately labeled. (11.5.1.9.; 1.2.6.9.)	OIC/Pharm Tech		
Drug storage	gen. Storage secure	General storage facilities should be secure, adequate, ventilated and well organised putting different groups of items in sections. (1.2.6.6.) - OIC Provide adequate storage facilities to improve the space constraints and enable better organization in the facility. (1.2.6.6)N - SURE-P	OIC / SURE-P	Is the drug storage neatly organized?	
	Med Kit/Storage	Medication/kit storage area should be well ventilated, secure and away from sunlight. Room and refrigerator temperature monitoring should be done daily and records kept. Records of corrective measures should also be kept. (11.2.1.4.)	OIC		12.1 Is there a separate pharmacy or drug storage area in the health facility?
					12.3 Enumerator: Record if the drug storage area is clean
					12.4 Are drugs protected from water and sunlight?
	Storage (instr) drum	Obtain a storage drum/ container for disinfected instruments. (8.2.5.3) Make available a secure and well ventilated storage area for sterilized instrument packs. These should be stored off the ground. (8.2.5.4.)	SURE-P/OIC	none	

HR management	Staff Orientation	Develop an orientation program for new staff at the facility. Keep appropriate records of program content and those in attendance including their signatures. (2.2.1.6.)	OIC/Midwife	none	
	Staff levels	Ensure the provision of the needed staff cadres according to the Minimum Standards for PHCs in Nigeria. (2.1.1.1,2.2.1.6) Using the Essential Staff Requirement gap analysis result, ensure the provision of the needed staff cadres (especially housekeeping and security). Provide the necessary personnel management with proper induction/orientation.(2.1.1.1,2.2.1.6)	SURE-P/LG	Given your normal patient load, do you feel this facility has enough staff?	
				What kind of staff do you need?	
				What action WOULD you take if you need additional staff? (Has this facility submitted a request for additional staff?)	
	Job descriptions	Create a mechanism that ensures that at the facility levels, job descriptions are known and facility-level performance measurement is done to inform designation and delegation of duties (2.2.2.1)	OIC/Matron	Does ${ros_name} have a written job description or performance agreement?	
				Do you have a system for MEASURING personnel performance?	
	Lab person	Designate an individual (with documented job descriptions) to manage the laboratory. Ensure there are policy-guided processes that foster collaborative work between the other units and the laboratory(9.1.1.1)	OIC	none	none
	Medication management person	Designate an individual (with documented job descriptions) for medication management. Ensure there are policy-guided processes that foster collaborative work between the other units and the pharmacy (11.6.1.1,11.8.1.1)	OIC	none	none
Patient rights	Patients rights	Obtain policy document on Patient's Right and Informed Consent and display in strategic areas of the facility. Train staff. Monitor implementation.(3.1.1.1-3.1.1.3,	OIC / SURE-P	Is there a patient rights charter posted in a public space?	

		3.6.1.1) btain patients rights charter. Display strategically in the facility. Train all staff on the patient's right to privacy during examinations, counselling & provision of information (OPD, wards, pharmacy, laboratory, etc). (3.1.1.2.;3.1.1.3)			
	Education plan	Develop a health education plan for the facility's patient population. Have a standardized method of keeping records of health education provided to each patient. (3.3.1.1.; 3.3.1.6.)	OIC/Midwife	none	
	Ward screens	Make provision for more ward screens in all relevant areas (ward, examination room etc.) (7.2.2.5) Ensure the availability of ward screens in relevant areas of the facility (at least 1 ward screen to 2 beds). Ensure windows in patient interaction areas have drapes & doors are kept closed during examinations & counselling. (3.2.1.1.-3.2.1.3.)	SURE-P	How many ward screens are available throughout the facility?	
Patient records	Transport	Make arrangements for a patient transport system within the community. Document this system and make it known to all personnel. Contact telephone numbers should be available and functional.(3.7.1.2)	OIC	none	1.2.20 Does this facility refer patients to other facilities?
					1.2.23 Does the facility have access to transportation for patients to take them to the referral health facility / hospital?
					1.2.24 What type of transportation for patients does the facility have access to?
	Referral Policy	Obtain a referral policy from the local/state government. Ensure policy includes the cases to be referred, services to be referred, a list of referral centers, and details of contact persons in the referral centers.(6.1.1.1)	OIC	none	
	Train staff privacy	Train all staff on the protection of the patient's right to privacy during all examinations, counselling and provision of information (OPD, in-patient ward, maternity ward, pharmacy, laboratory, etc).(3.2.1.1, 3.2.1.2, 3.2.1.3)	OIC	none	

	Patient Records	Ensure all patient records are standardised, dated, up to date, signed and contain the designation/name of personnel carrying out the assessment.(4.4.2.1)	OIC	Do you keep individual case records?	
				Case file 1: Is the following indicated...?	
				Name of patient	
				Date of visit	
				Initials or name of health worker	
				Condition (last visit)	
	Sec Patient files storage	Make available a secure cabinet/cupboard for the storage of patient files. Ensure files are neatly arranged according to colour, condition and unique identification number. Implement systems for easy retrieval of records. (4.1.1.6.)	OIC	What kind of files does the PHC keep?	
				In what form are files kept?	
				Does the facility keep records for….	
	Registers	Ensure all national and local registers are completely filled with correct information. Designate an individual to oversee this process.(4.3.1.1)	OIC	none	6.1.1 Does the facility have an MCH register?
	Information officer	Designate an individual to be responsible for the management of information. Establish policy-guided processes and provide personnel education/training for the use of data at the facility level (4.3.1.1, 4.4.2.1)	OIC	none	
	Early attention patients	Create a list of patients who require early attention and document the system for identifying and fast-tracking these patients. (6.3.1.4, 6.3.1.5)	OIC	none	
	Summary of care	Ensure care providers write a summary of care provided to each patient whilst on admission in the facility as well as follow-up instructions(6.1.1.1,6.1.1.4)	OIC	none	none
Waste management	Waste management policy	Obtain policy on waste management. Train personnel on waste segregation & containers for collection. Keep adequate records. Monitor implementation. Display posters on waste segregation at different areas of the facility. (5.6.2.1.) Establish a policy-guided waste management system at the facility. Provide relevant	OIC/SURE-P/LG	Is medical waste disposed together with regular waste or separately?	

		tools/resources (sharps boxes, PPEs, pedaled bins, etc). Train personnel, provide reminders and monitor the adherence to protocols(5.6.1.8, 5.6.2.1, 5.6.2.2,13.3.4.4).			
				How does this facility finally dispose of medical waste (other than sharps boxes)?	
				Are there posters showing waste separation in the clinic?	
				Does this facility have any guidelines on health care waste management?	
				Have you or any provider(s) received training in health care waste management practices in the past two years?	
				Do you have a schedule for burning waste?	
	Safety boxes	Ensure the availability of safety boxes and covered dustbins in all areas of the facility for waste collection. Dustbins should have colour coded bin liners or should be painted with the respective colour codes. (5.6.2.4.; 13.3.4.2.; 13.3.4.3.)	OIC/SURE-P	To enumerator: Are there waste bins in the clinic?	
				Are the waste bins covered?	
				Are the waste bin for different types of waste clearly marked? (for example color coded)	
	More waste bins	Make provision for more waste bins in the facility (5.6.2.3)	SURE-P	none	
Risk management	Colored mops	A colour coded system should be employed for mops & brooms in cleaning different areas of the facility (a designated mop/broom for the labour room, toilets, consulting area etc). (Std 5.6.1.)	SURE-P/Officer-in-Charge	Are there different mops available for high and low risk areas in the facilities?	
				Is there a color coded system for these mops	
	Availability of protective equipment	Ensure availability of personal protective equipment (gloves, masks, aprons e.t.c) for staff in all relevant areas.(5.6.1.8)	SURE-P/LG	Observe: are there medical gloves available?	
	Use of protective	Ensure that personnel make correct use of personal protective equipment(gloves, masks, aprons). (5.6.1.8)	OIC	none	

	Use of protective	Obtain the Government policy for the provision of Post Exposure Prophylaxis. Train staff on the policy and how to access these services. (5.2.1.7)	OIC	none	
	Infection ctrl	Designate an individual to be responsible for infection control and ensure the provision of continuous in-service training on infection control to all personnel. Keep records of all training(content of training and attendance list). (5.6.1.1, 5.6.1.4)	OIC	Is there a designated individual responsible for infection control at this facility?	
	Retrain on disinfection	Retrain staff on disinfection techniques.(5.6.1.4)	OIC	Were staff trained on disinfection technqiues? (last 6 months)	
				If yes, have you kept a record of the training?	
	Fire fighting	Provision for fire fighting equipment should be made. Staff should be trained on how to use these equipment and regular servicing of fire-fighting equipment should be done.(5.4.1.3, 5.4.1.7)	SURE-P/ LG	Are there fire extinguishers (functional)?	
	Flammable labled	Ensure all flammable materials (fuel, kerosene, methylated spirit, etc.) are clearly labelled and have appropriate signage in its environs.(5.4.1.4)	OIC	Are flammable materials clearly labelled? (fuel, kerosene, meth spirit, etc.)	
Hand-washing	Hand washing poster	Display posters addressing hand washing at different areas of the facility. (5.6.1.7)	OIC	Is there at least one poster on display describing hand-washing behavior?	
	Soap/Water	Ensure the provision of soap, water and paper towels/single use towels at hand washing facilities. Water can be distributed to relevant areas of the facility with the use of buckets with tap heads (veronica buckets).(5.6.1.6)	OIC/SURE-P/WDC	Visible presence of hand washing supplies (soap and water)	
Safety and security	Security (external)	Develop a mechanism for summoning the assistance of external sources of security in case of an emergency (eg. Police, community guards, etc.). Document this mechanism and make it known to all personnel.(5.3.1.5)	OIC	Do you have contact phone numbers of any external security sources e.g. police, civil defence and vigilantee?	
	Security personnel	Provision should be made for at least 2 security personnel who can run daily shifts, covering the facility round the clock.(5.3.1.3)	LG/WDC	none	
	Assault safety	Develop a process that protects personnel & patients from assault. Ensure staff are aware. Control access to the facility	OIC/CHC	none	

		& restricted areas. Display posters on no-tolerance for violence. Ensure no dark areas are within & around the facility. (5.3.1.5.)			
	Access control	Ensure access control measures are in place at the pharmacy, laboratory, labour room and other restricted areas. Ensure doors are lockable and have appropriate signage eg. "authorized entry only", "restricted area", etc. (5.3.1.2.)	OIC/CHC		12.2 Can the doors and windows be locked to keep the drug storage area secured?
	Rails	Guiding/supporting rails should be fitted for all staircases and along the high corridors. (Std. 5.1.1.)	CHC/SURE-P/LG	none	
	Repair Gate	Repair the gate at the entrance to the compound of the facility for security reasons. (5.3.1.1, 5.3.1.3) Provide perimeter fencing and ensure a lockable gate is in place. (5.3.1.1, 5.3.1.3)	SURE-P	none	
Deliveries	Partographs	Ensure the availability and use of partographs to monitor all deliveries at the facility. (6.6.5.4) - SURE-P Ensure the use of partograph to monitor all deliveries and keep records of apgar score for newborns. Ensure all tests, observations and examinations are recorded for all antenatal and postnatal cases. (6.6.2.4.; 6.6.3.6.; 6.6.5.4.; 6.6.6.3.)	LG/SURE-P/OIC	Is there a partograph available in the facility?	
				Is it posted visibly?	
	Newborn identification	Put in place a system to identify newborns (eg. use of wristbands). Display posters reminding mothers not to leave their babies unattended to. Ensure only authorized access to the wards. (6.6.5.6.)	OIC/Midwife	none	
	Apgar	Record the Apgar score for each newborn baby in the respective patient's card and delivery register.(6.6.5.4) Ensure the use of partograph to monitor all deliveries and keep records of apgar score for newborns. Ensure all tests, observations and examinations are recorded for all antenatal and postnatal cases. (6.6.2.4.; 6.6.3.6.; 6.6.5.4.; 6.6.6.3.)	OIC	Check the records for the 10 most recent births. How many have an "apgar" reported?	

	Lamp	Make provision for an angle-poise lamp for adequate lighting in the delivery room (6.6.4.1)	SURE-P	none	Delivery light: 11.7
	Delivery table	Make provision for a delivery table with stirrups (6.6.4.2)	SURE-P	none	Delivery table: 11.7
	Delivery room equipment	Provide the necessary tools and equipment required in the labor room (delivery table with stirrups, angle poise lamps, delivery kits). Provide documented training for the relevant personnel in the use of these (6.6.4.1)	SURE-P/WDC	none	generate an index from SURE-P data for all available and functional delivery equipments
					11.7 Is the following equipment Available and Functioning/Working (AF), Available but not Functioning/Working (ANF), or Not Available (NA)?
	ANC PNC records	Ensure all records of ANC, labour and post-natal care are kept for each patient in their respective patient cards. Provide individual patient records template for Labour, Postnatal & Inpatient care. (6.6.2.4, 6.6.3.6, 6.6.6.3)	OIC	Last month: how many antenatal visits did this facility receive?	
	SUREP records	Supply the SURE-P ANC patients files to provide a template for proper records. (6.6.2.4)	SURE-P	none	none
Equipments and guidelines	Equipment	Implement a system that ensures all equipment and supplies are available, properly stored and distributed to all relevant areas of the facility. A list of all equipment and supplies should be available.(1.2.6.5, 1.2.6.6, 1.2.6.7)	OIC	only sterilization equipment	generate an index from SURE-P data for all available and functional outpatient/lab equipments
					11.1 Where is the outpatient equipment located?
					11.4 Where is the lab equipment located?
					11.6 Where is the delivery and neonatal equipment located?
	Standing orders	Make standing orders for CHOs/CHEWs and JCHEWs available. Make LSS and MLSS guidelines available at the facility. (6.6.1.1.) Obtain the national treatment guidelines and standing orders to guide all staff in their clinical practice. (1.2.1.4)	OIC/LG/SURE-P	Do you have printed guidelines for the treatment of the following medical issues?	
Malaria	Malaria guidelines	Obtain national guidelines for the treatment of Malaria and ensure compliance these guidelines. Keep complete records for the malaria cases managed. (6.8.4.1)	OIC	Do you have printed guidelines for the treatment of the following medical issues?	

		Obtain national guidelines for the treatment of Malaria. Ensure that the management of malaria accords with national guidelines. Keep appropriate records of cases receiving ACT following a Laboratory confirmation. (6.8.4.1			
	Malaria testing records	Keep appropriate records of malaria cases treated on the basis of clinical diagnosis only. (6.8.4.1)	OIC	How many patients were diagnosed with malaria last month?	
				How many of those were diagnosed via rapid diagnostic test (RDT)?	
				How many were diagnosed with other lab testing methods? (for example microscope)	
				How were malaria patients treated?	
				"Silent question": Did the respondent use written records to answer any of the questions?	
	Malaria Testing Equipment	Make available facilities and equipment for the testing of malaria. (6.8.4.2.)	OIC/SURE-P/LG	none	none
Sterilization	Flow and Demarcation	Ensure a clear flow and demarcation of activities in the sterilization area (decontamination, washing, drying, packing, sterilizing and storage). (8.2.5.1.)	OIC/Midwife		
	Sterilization process	Establish an effective sterilization process (with regular testing) and provide the appropriate training for the personnel.	OIC		
	Autoclave	An autoclave should be provided & installed and used for sterilizing instruments. Staff should be trained on how to use the autoclave (8.2.5.6) Where autoclaves/pressure pots are present, these should be installed and used for sterilizing instruments. Provide training on the use.(8.2.5.6)	SURE-P/OIC	autoclave: Which of the following items are FUNCTIONAL?	
Facility characteristics - infrastructure	Toilets	Ensure the provision of a minimum of 2 functional sanitary facilities (patient and staff) in the facility.	SURE-P/WDC	Questions for in-charge (or main respondent): Which rooms do you have in this facility? Room11: toilet	
				Are the PATIENT toilet rooms clean? Please rate	

Other	Ward Development Committee (WDC)	Strengthen the community involvement process through establishing goals for the WDC and incorporating quality improvement indicators in the performance review for the unit(1.2.3.3,1.2.4.1)	OIC/SURE-P	none	

4. Bias in patient satisfaction surveys: a threat to measuring health care quality[82]

4.1 Abstract

Patient satisfaction surveys are an increasingly common element of efforts to evaluate the quality of health care. Many patient satisfaction surveys in developing countries frame statements positively and invite patients to agree or disagree, so that positive responses may reflect either true satisfaction or bias induced by the positive framing. In an experiment with more than 2,200 patients in Nigeria, we distinguish between actual satisfaction and survey biases. Patients randomly assigned to receive negatively framed statements expressed significantly lower levels of satisfaction (87 percent) than patients receiving the standard positively framed statements (95 percent – p-value<0.001). Depending on the question, the effect is as high as a 19 percentage point drop (p<0.001). Thus, high reported patient satisfaction likely overstates the quality of health services. Providers and policy makers wishing to gauge the quality of care will need to avoid framing that induces bias and to complement patient satisfaction measures with more objective measures of quality.

4.2 Introduction

As access to at least some level of health services increases in low- and middle-income countries, the focus of policymakers shifts to quality: How can we ensure that patients receive high-quality care? But even while measuring the provision of care is challenging in systems with limited data, measuring the quality of care invites a host of new complications. How can we regularly, systematically gauge the quality of medical attention and advice? The simplest, most direct approach seems to be to ask the patients themselves. To gauge the quality of care, many policymakers and researchers turn to the patient satisfaction survey.

In high-income countries, results from patient satisfaction surveys are used to identify gaps and to inform quality improvement plans in healthcare organizations and health systems (Browne et al., 2010), as well as in research (Aiken et al., 2012; Bjertnaes et al., 2012). Moreover, patient satisfaction is often used as a performance indicator that influences hospital reimbursements and, more and more frequently, physician compensation (Fenton et al., 2012; Medical Group Management Association, 2016). In low- and middle-income countries, these surveys are increasingly

[82] An amended version of this chapter was published in the British Medical Journal – Global Health (Dunsch et al., 2018). Co-authors are David Evans, Mario Macis, and Qiao Wang.

used. For example, in Africa alone, patient satisfaction instruments have been used in Kenya (Mwanga, 2013), South Africa (Chimbindi et al., 2014; Phaswana-Mafuyaet et al., 2011), Nigeria (World Bank, 2016), and Tanzania (Leonard, 2008; Khamis and Njau, 2014), among others.

These surveys often provide patients with a statement and then ask them to agree or disagree with that statement, such as “This health facility is clean. Do you agree or disagree?” If patients answer these questions favorably, does that actually reflect high levels of patient satisfaction, or rather does it reflect a bias? Patients in low-income environments with few options for health services may value any services, and indeed, other work supports high reported patient satisfactions even in the face of relatively low quality services (Evans and Welander-Tärneberg, 2018). Alternatively, patients may agree with the interviewer to be agreeable (“acquiescence bias”) or because “I agree” is the first option offered and requires the least effort (“satisficing”; Krosnick, 2000).

This is a substantive issue: of 26 recent patient satisfaction surveys in low- and middle-income countries, more than three-quarters phrased their statements positively. Specifically, of the 26 studies included in the World Bank's Central Microdata Catalog that used patient satisfaction questions, 20 (77%) were phrased positively and only 6 (23%) were phrased negatively or neutrally. This potential framing bias adds to other, previously identified challenges with patient satisfaction surveys, such as that patient satisfaction measured at clinics is rated much higher than patient satisfaction measured at home (Das and Pave, 2015).

4.3 How can we distinguish true patient satisfaction from bias induced by the survey?

We implemented an experiment in Nigeria to distinguish between actual satisfaction with health services and survey biases. The study was implemented in 80 primary healthcare centers in 6 Nigerian states: Anambra, Bauchi, Cross River, Ekiti, Kebbi, and Niger. Patient exit interviews were administered to all patients who visited the primary healthcare centers at the time of data collection. Surveys were administered face-to-face by trained enumerators with tablet computers in 8 monthly rounds between June 2014 and February 2015. Interviewers arrived unannounced as part of a larger randomized controlled trial that involved helping clinic staff to identify gaps in the quality of service delivery and to set goals to close those gaps (Dunsch et al., 2017). The patient exit interview did not mention the larger quality-improvement intervention. In total, 2,370 patients were interviewed, or roughly 30 patients per facility on average. In addition to patient satisfaction measures, data were collected on a set of demographic and socio-economic characteristics of the patients (including age, gender, education, employment and income).

Each patient was presented with 11 statements on the quality of care and asked to agree or disagree with each statement. Patients were randomly assigned to receive one of three treatments: the standard, positively framed statements (Table 4.1 Set A), a set of equivalent negatively framed statements (Table 4.1 Set B), or a random mix of the two.

Table 4.1: Positive and Negative Framed Patient Satisfaction Statements

Set A: Positively framed statement	*Set B: Negatively framed statement*
1. The lab fees today were reasonable.	The lab fees today were too expensive.
2. This health facility is clean.	This health facility is dirty.
3. The waiting time was appropriate.	The waiting time was too long.
4. The fees for medicines or drugs you received today were reasonable.	The fees for medicines or drugs received today were too expensive.
5. The staff at this facility is courteous and respectful.	The staff at this facility is rude and disrespectful.
6. The staff did a good a job of explaining your condition.	The staff did a poor job of explaining your condition.
7. You had enough privacy during your visit.	You had too little privacy during your visit.
8. The registration fees of this visit to the health facility were reasonable.	The registration fees of this visit to the health facility were too expensive.
9. The staff spent a sufficient amount of time with you.	The staff spent too little time with you.
10. The hours this facility is open are adequate to meet your needs.	The hours this facility is open are too short to meet your needs.
11. You completely trust the staff's decision about medical treatment in this facility.	You do not completely trust the staff's decision about medical treatment in this facility.

The random assignment of individual patients to treatments was generated by software ("SurveyCTO") on the tablets at the time of interview. As expected with randomization and a large sample of patients, patients were statistically indistinguishable across groups on age, gender, education, and employment. The enumerators did not know in advance which set of statements would be presented, the surveys were anonymous, and the interviews were conducted with spatial separation from the PHCs to ensure confidentiality.

For the negatively framed statements, we avoided statements with the word "not", as deciding whether you disagree with the statement "You did not have enough privacy during your visit" can be confusing to respondents due to the double negative (Lietz, 2008). As such, in that case, we framed the statement as "You had too little privacy during your visit" in the negatively framed statements. All questions were asked in two stages. In the first stage, the respondent had to decide whether to "agree", "neither agree nor disagree", or "disagree" with the presented statement. Then, in the second stage, the respondent decided – conditional on having chosen to agree or disagree – whether to agree or disagree strongly or not (see Figure 4.1). For the analysis, we reversed the sign on the negatively framed questions, so that we are comparing the people who agreed with positively framed statements to people who disagreed with negatively framed statements.

Figure 4.1: Experiment decision structure

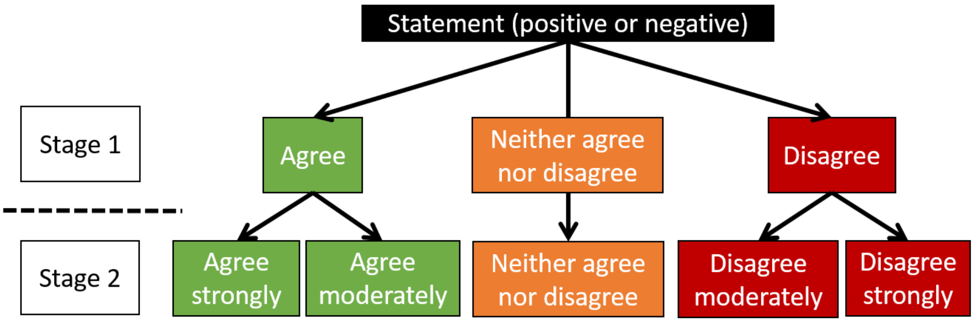


Table 4.2 shows the distribution of participants across treatment groups, by state and overall. In total, 42 percent of patients received the positively framed questions, 42 percent received the negatively framed questions, and 16 percent answered the random mix.[83]

Table 4.2: Distribution of participants across treatment groups, by state and overall

State	N.	Positive framing (%)	Negative framing (%)	Positive-Negative Mixed Framing (%)
Anambra	346	43%	44%	14%
Bauchi	456	40%	42%	18%
Cross River	265	43%	38%	19%
Ekiti	325	44%	43%	14%
Kebbi	444	45%	39%	16%
Niger	386	38%	47%	15%
Total	**2,222**	**42%**	**42%**	**16%**

In Table 4.3, we present average patients' characteristics, overall and by treatment condition. The average age of patients was 30.3 years. 72% of the patients interviewed were between 19 and 34 years old, 19% were between 35 and 54, 5% were 55 or older, and 3% were 18 or younger. Only 39% of the patients had at least some secondary school education, 83% report being self-employed, 10% were unemployed, and 90% were married. 72% of the patients had never been to a

[83] The third treatment condition, a mix of positively- and negatively-framed statements, was used only during the first three rounds of data collection (of eight total); this explains the fact that they account for a smaller share of the observations.

private health care facility. The random allocation of treatment conditions had the desired effect of achieving balance across all of these characteristics.

Table 4.3: Patient characteristics, overall and by treatment group

	Total		Positive framing		Negative framing		Positive-Negative Mixed Framing	
	N.	*mean*	*n.*	*mean*	*n.*	*mean*	*n.*	*mean*
Age	2,211	30.3	923	30.5	938	29.9	350	30.5
Age group:								
<=18 years	72	3%	27	3%	34	4%	11	3%
19-34 years	1600	72%	668	72%	685	73%	247	71%
35-54 years	424	19%	173	19%	177	19%	74	21%
>=55 years	115	5%	55	6%	42	4%	18	5%
Gender								
% female	1,859	84%	772	83%	802	85%	285	81%
Employment								
Employed	150	7%	72	8%	56	6%	22	6%
Self-employed	1,840	83%	749	81%	791	84%	300	85%
Unemployed	230	10%	108	12%	92	10%	30	9%
Education Level								
Low	1,365	61%	577	62%	569	61%	219	62%
High	855	39%	352	38%	370	39%	133	38%
Marital Status								
Married	1,991	90%	831	89%	842	90%	318	90%
Single	182	8%	80	9%	78	8%	24	7%
Widowed	42	2%	18	2%	16	2%	8	2%
Divorced	5	0%	2	0%	2	0%	1	0%
Ever been to a private health care facility	611	28%	259	28%	242	26%	110	31%

Notes: Low education = primary school or less (no completed education, adult literacy education, arabic, vocational, other); High education refers to secondary school and higher, including college and higher (university, master's degree, MSc/MA, Ordinary National Diploma, Higher National Diploma, Nigeria Certificate in Education.

4.4 How patient satisfaction questions are framed makes a big difference

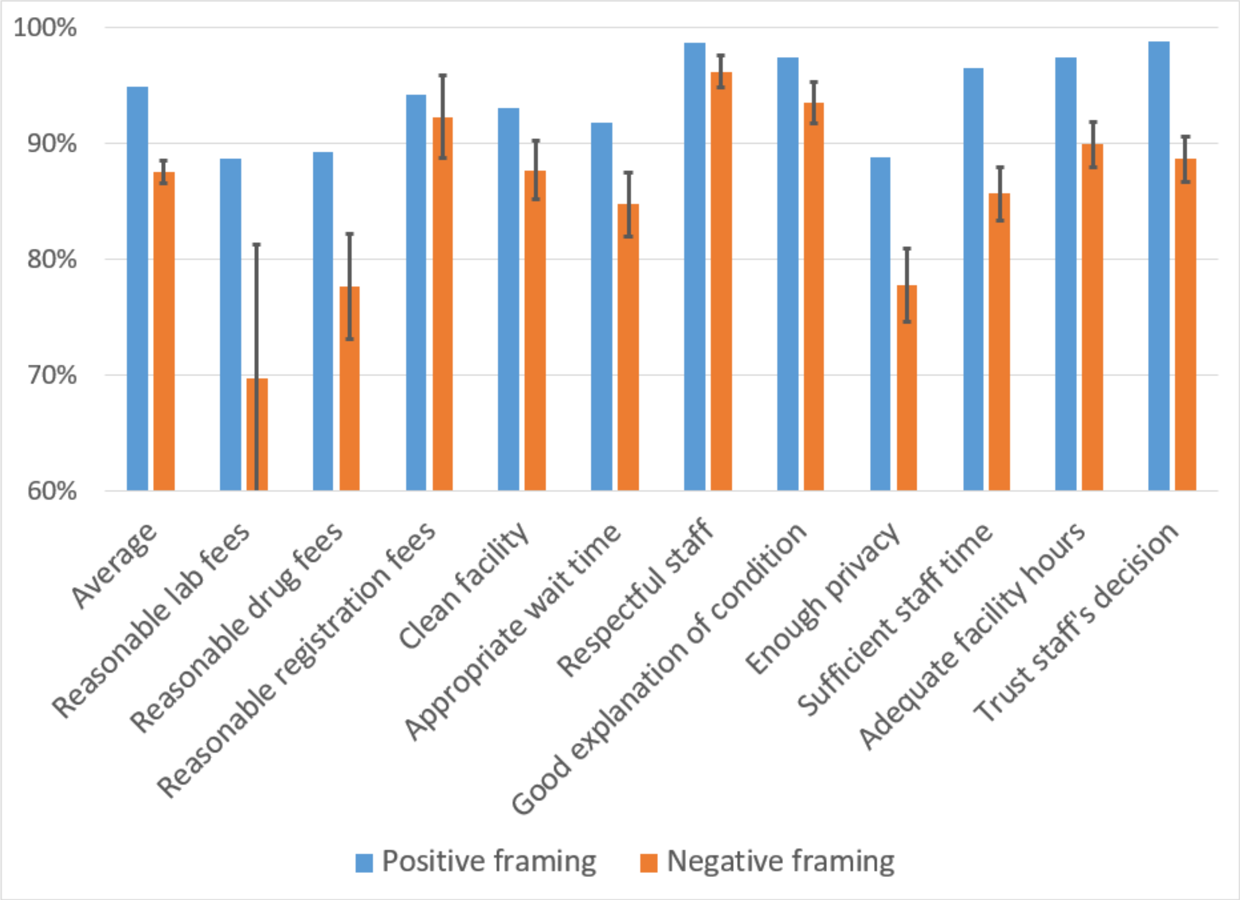
Analysis

We estimate three linear probability models. We have estimated ordinal logit models with similar results. Here, we use linear probability models both because it is one of the most common methods of estimation with patient satisfaction survey analysis (Evans and Welander-Tärneberg, 2018) and for ease of interpretation (Angrist and Pischke, 2008).

$$(1)\ favorable_{ik} = \beta_0 + \beta_1 neg_k + \varepsilon_{ik}$$

where *favorable*$_{ik}$ takes the value 1 if patient i gave a favorable response to statement k, and 0 otherwise, and *neg* denotes negatively framed statements. Because we have balance across observed characteristics (gender, education, age, and income), we do not control for them in our main specification, although we do so as a robustness check. The results of this specification are reported in Table 4.4. Figure 4.2 shows the results visually, and provides confidence intervals around the estimates. With positively framed statements, patients report extremely high levels of satisfaction. At least 88% of patients agree with all 11 statements; for more than half of the statements, agreement exceeds 94%. However, when patients are presented with negatively framed questions, satisfaction drops significantly on 10 out of 11 questions, with an average drop of 7.5 percentage points across all questions and including drops as large as 18.9 and 11.6 percentage points. These results are consistent across patient genders, ages, education levels, and income levels.

Figure 4.2: The Impact of Positive and Negative Framing on Patient Satisfaction



(2) $favorable_{ik} = \beta_0 + \beta_1 neg\_w\_neg_k + \beta_2 pos\_w\_mix_k + \beta_3 neg\_w\_mix_k + X_i + \varepsilon_{ik}$

In this second specification, we examine whether including a negatively framed statement within a mix of positively and negatively framed statements affects reporting. *neg_w_neg* denotes negatively framed statements in sets of all negative statements, and *pos_w_mix* and *neg_w_mix* denotes positively and negatively framed statements, respectively, in sets of mixed positive and negative statements (the omitted (or reference) category thus consists of positively framed statements in sets of all positive statements). The results of this specification are reported in Table 4.5. For 10 out of the 11 statements (and on average), the impact of negative statements with all negative statements is negative and statistically significant. Patients who received a negative statement in the mixed battery of statements were also less likely to respond favorably. Here, 8 of the 11 statements show significant effects.

(3) $favorable_{ik} = \beta_0 + \beta_1 neg_k + \beta_2 X_i + \beta_3 neg_k \times X_i + \varepsilon_{ik}$

To probe the robustness of the results, in the third specification, we examine whether the impact of negative framing differs by patient characteristic, where X represents a patient characteristics such as gender, education, assets, or previous experience with private facilities. The results of this specification are reported in Appendix table A4.1. In all cases we obtain very similar results to our main specification. We see no statistically significant differences of framing by these characteristics, as demonstrated in the coefficients of the interaction terms. That is, the pattern of acquiescence bias that we uncovered seems to affect patients irrespective of their gender, income, or education.

We find the same result – that the positive or negative framing is crucial to patient responses – if we focus on the more detailed "stage 2" patient responses, when they are asked – conditional on agreement with each statement – if they *strongly* agree or disagree (Appendix Table A4.2). Of the 11 items, 8 are significant for the *neg_w_neg* group and 7 out of 11 in the *neg_w_mix* group. The effects are slightly smaller for the *neg_w_neg* group when compared to the stage 1 results and about the same for the *neg_w_mix* group. In the *neg_w_mix* group, statement 4 (drug fees) is insignificant for stage 2. For the *neg_w_neg* group, statements 2 (cleanliness) and 5 (respect) become insignificant. The largest effect in this group can be observed for the "lab fees" item.

Table 4.4: Impact of framing on patient satisfaction – The simple specification

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Lab fees	Drugs fees	Registration fees	Clean	Wait time	Respect	Explain	Privacy	Staff time	Open hours	Trust	Overall
Negative	-0.189	-0.116	-0.0195	-0.0534	-0.0709	-0.0246	-0.0395	-0.111	-0.109	-0.0755	-0.102	-0.0746
	[0.002]	[<0.001]	[0.275]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Positive (Control Mean)	0.886	0.892	0.942	0.930	0.918	0.986	0.974	0.888	0.965	0.974	0.988	0.949
Obs. (N)	178	1004	784	2219	2219	2213	2204	2209	2219	2144	2193	19586
Missing values	2	7	37	3	3	9	18	13	3	78	29	202
Obs with perfect response rate	180	1011	821	2222	2222	2222	2222	2222	2222	2222	2222	19788

Notes: Dependent variable = 1 if the patient responded favorably in stage 1 (i.e., "agree" on positively framed questions or "disagree" on negatively framed questions), 0 otherwise. “Negative” refers to a negatively framed item. “Positive” refers to a positively framed item. The numbers reported below the coefficients are p-values. The total patients asked each question differs because certain questions only applied to a subset of patients.

Table 4.5: Impact of framing on patient satisfaction – The detailed specification

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Lab fees	Drugs fees	Registration fees	Clean	Wait time	Respect	Explain	Privacy	Staff time	Open hours	Trust	Overall
	-0.204	-0.105	-0.0282	-0.0561	-0.0643	-0.0213	-0.0293	-0.0940	-0.106	-0.0659	-0.0807	-0.0665
Negative with negative	[0.002]	[<0.001]	[0.092]	[<0.001]	[<0.001]	[0.004]	[0.002]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
	-0.165	-0.148	-0.120	-0.0358	-0.110	-0.0323	-0.0835	-0.212	-0.132	-0.143	-0.218	-0.124
Negative with mixed	[0.113]	[0.004]	[0.022]	[0.163]	[<0.001]	[0.055]	[0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
	-0.0352	0.0283	-0.227	0.0014	-0.00449	0.00955	0.0114	-0.00597	-0.00633	-0.0177	0.000658	-0.005
Positive with mixed	[0.731]	[0.433]	[0.003]	[0.945]	[0.846]	[0.161]	[0.303]	[0.820]	[0.697]	[0.271]	[0.940]	[0.454]
Pos with Pos												
(Control Mean)	0.892	0.888	0.963	0.930	0.918	0.985	0.972	0.889	0.966	0.977	0.988	0.949
Obs. (N)	178	1004	784	2219	2219	2213	2204	2209	2219	2144	2193	19586
Missing values	2	7	37	3	3	9	18	13	3	78	29	202
Obs with perfect response rate	180	1011	821	2222	2222	2222	2222	2222	2222	2222	2222	19788

Notes: Dependent variable = 1 if the patient responded favorably in stage 1 (i.e., "agree" on positively framed questions or "disagree" on negatively framed questions), 0 otherwise. "Negative with negative" refers to a negatively framed item with a battery of negatively framed items. "Negative with mixed" refers to a negatively framed item with a random mix of negative and positive items. The numbers reported below the coefficients are p-values. The total patients asked each question differs because certain questions only applied to a subset of patients.

4.5 Conclusion

There is broad consensus that improving patients' experience as they obtain health care is an intrinsically desirable goal. Some elements of that improved experience are likely to be universal: patients value short waiting times and clean facilities, and they appreciate providers that respond to their needs and treat them with respect. Other elements may vary across contexts, such as the extent to which patients value being involved in the medical decision making process (Browne et al., 2010). Accordingly, routine measurement of patient experience and satisfaction are becoming commonplace in health care organizations in both high-income and developing countries.

Our results demonstrate that patient satisfaction measurements are deeply sensitive to the framing of the questions. Specifically, we find strong evidence of acquiescence bias, or the tendency of individuals to agree to the statement they are presented, irrespective of their content. As such, the standard ("positive") formulation results in consistently inflated measures of patient satisfaction. These results suggest that high reported patient satisfaction likely overstates the quality of health service provision in resource-constrained environments, adding to evidence that patient satisfaction is imperfectly related to health outcomes (Fenton et al., 2012). Inflated patient satisfaction reports can potentially distort decisions about effort and resource allocation. This highlights the need to supplement patient satisfaction with other measures to provide an overall indication of service quality. These may include the measurement of actual health outcomes, as well as the use of vignettes to gauge provider knowledge and standardized patients to gauge provider effort. Furthermore, there may be significant ceiling effects with positively framed questions, since the average tends to be so high that it is difficult to distinguish across performance levels (Voutilainen et al., 2016; Andrew et al., 2011).

The main implication of our study is that designers of patient satisfaction surveys should avoid using all positively phrased statements. Providing a mix of positively and negatively framed statements would attenuate the overall bias, although bias would still be present in the responses to each individual statement. Avoiding agree/disagree, yes/no response formats would also reduce acquiescence bias. Several major patient satisfaction surveys in use already incorporate these recommendations. For example, the Picker Patient Experience Questionnaire (PPE-15) avoids agree/disagree statements altogether (Jenkinson et al., 2002), and the Patient Experience Questionnaire (PEQ) has agree/disagree statements but includes both positive and negative framing (Steine et al., 2001).

Reduced bias would make patient satisfaction measures more meaningful, allowing better distinguishing across facilities, and would be beneficial for programs wishing to use patient satisfaction to identify gaps and areas where changes are needed.

Appendix Table A4.1: Impact of framing – Interaction with patient characteristics

	(1)	(2)	(3)	(4)	(5)
	Baseline (no control)	*Baseline + Gender interaction terms*	*Baseline + Age group interaction terms*	*Baseline + Education interaction terms*	*Baseline + Wealth Quintile interaction terms*
Depedent Var.	overall effect	overall effect	overall effect	overall effect	overall effect
Independent Var.					
Neg with Neg	-0.0665	-0.0603	-0.0644	-0.052	-0.0548
	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Neg with Mix	-0.124	-0.147	-0.16	-0.151	-0.119
	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Pos with Mix	-0.00528	0.0143	-0.0189	-0.00477	-0.000419
	[0.454]	[0.237]	[0.204]	[0.650]	[0.974]
Female		-0.00628			
		[0.382]			
Female * Neg with Neg		-0.00714			
		[0.624]			
Female * Neg with Mix		0.0275			
		[0.335]			
Female * Pos with Mix		-0.0251			
		[0.087]			
Age group (24-44)			0.00384		
			[0.568]		
Age group (>=45)			0.0282		
			[0.007]		
Age group (24-44) * Neg with Neg			-0.00236		
			[0.851]		
Age group (24-44) * Neg with Mix			0.0437		
			[0.152]		
Age group (24-44) * Pos with Mix			0.0196		
			[0.256]		
Age group (>=45) * Neg with Neg			0.000149		
			[0.995]		
Age group (>=45) * Neg with Mix			0.0885		
			[0.019]		
Age group (>=45) * Pos with Mix			0.0229		
			[0.302]		
Education (Low)				-0.00788	
				[0.218]	
Education (Low) * Neg with Neg				-0.0237	
				[0.040]	
Education (Low) * Neg with Mix				0.0431	
				[0.082]	
Education (Low) * Pos with Mix				-0.000812	
				[0.954]	

Quintile (Poorest)					-0.0326
					[0.001]
Quintile (Less poor)					-0.00997
					[0.257]
Quintile (Average)					-0.0164
					[0.077]
Quintile (Less poor)					-0.0032
					[0.731]
Quintile (Poorest) * Neg with Neg					-0.0206
					[0.232]
Quintile (Poorest) * Neg with Mix					0.0364
					[0.273]
Quintile (Poorest) * Pos with Mix					0.0107
					[0.613]
Quintile (Less poor) * Neg with Neg					-0.0186
					[0.263]
Quintile (Less poor) * Neg with Mix					0.0168
					[0.603]
Quintile (Less poor) * Pos with Mix					-0.0241
					[0.258]
Quintile (Average) * Neg with Neg					-0.00425
					[0.810]
Quintile (Average) * Neg with Mix					-0.0235
					[0.523]
Quintile (Average) * Pos with Mix					-0.0092
					[0.655]
Quintile (Less poor) * Neg with Neg					-0.0224
					[0.212]
Quintile (Less poor) * Neg with Mix					-0.0587
					[0.126]
Quintile (Less poor) * Pos with Mix					-0.00668
					[0.739]
Pos with Pos (Control Mean)	0.949	0.949	0.949	0.949	0.949
Obs. (N)	19586	19586	19222	19568	19361

Appendix Table A4.2: Impact of framing on patient satisfaction – Second stage (“strongly agree”)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Lab fees	Drugs fees	Registration fees	Clean	Wait time	Respect	Explain	Privacy	Staff time	Open hours	Trust	Overall
Neg with Neg	-0.178	-0.0859	-0.0306	-0.0264	-0.0572	-0.00178	-0.0263	-0.0845	-0.114	-0.0452	-0.0867	-0.0571
	[0.028]	[0.005]	[0.175]	[0.175]	[0.002]	[0.880]	[0.046]	[<0.001]	[<0.001]	[0.002]	[<0.001]	[<0.001]
Neg with Mix	-0.0867	-0.103	-0.191	0.0347	-0.124	-0.0209	-0.0896	-0.196	-0.130	-0.117	-0.223	-0.112
	[0.463]	[0.068]	[0.003]	[0.296]	[0.001]	[0.371]	[0.003]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Pos with Mix	-0.0802	-0.0335	-0.181	0.0622	-0.0255	0.00152	0.00767	0.0414	0.0155	-0.0223	0.0188	0.00479
	[0.571]	[0.550]	[0.019]	[0.034]	[0.433]	[0.940]	[0.712]	[0.170]	[0.526]	[0.392]	[0.149]	[0.695]
Pos with Pos (Control Mean)	0.723	0.770	0.917	0.785	0.830	0.932	0.925	0.802	0.892	0.911	0.958	0.874
Obs. (N)	178	1004	784	2219	2219	2213	2204	2209	2219	2144	2193	19586
N. of missing response	2	7	37	3	3	9	18	13	3	78	29	202
Obs with perfect response rate	180	1011	821	2222	2222	2222	2222	2222	2222	2222	2222	19788

Dependent variable = 1 if the patient responded strongly favorably in stage 2 (i.e., "strongly agree" on positively framed questions or "strongly disagree" on negatively framed questions), 0 otherwise.

4.6 References

Aiken, L. H., Sermeus, W., Van den Heede, K., Sloane, D. M., Busse, R., McKee, M., ... & Tishelman, C. (2012). Patient safety, satisfaction, and quality of hospital care: cross sectional surveys of nurses and patients in 12 countries in Europe and the United States. *Bmj*, *344*, e1717.

Andrew, S., Salamonson, Y., Everett, B., Halcomb, E. J., & Davidson, P. M. (2011). Beyond the ceiling effect: using a mixed methods approach to measure patient satisfaction. *International Journal of Multiple Research Approaches*, *5*(1), 52-63.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bjertnaes, O. A., Sjetne, I. S., & Iversen, H. H. (2012). Overall patient satisfaction with hospitals: effects of patient-reported experiences and fulfilment of expectations. *BMJ Qual Saf*, *21*(1), 39-46.

Browne, K., Roseman, D., Shaller, D., & Edgman-Levitan, S. (2010). Analysis & commentary measuring patient experience as a strategy for improving primary care. *Health Affairs*, *29*(5), 921-925.

Chimbindi, N., Bärnighausen, T., & Newell, M. L. (2014). Patient satisfaction with HIV and TB treatment in a public programme in rural KwaZulu-Natal: evidence from patient-exit interviews. *BMC health services research*, *14*(1), 32.

Das, J., & Sohnesen, T. P. (2006). *Patient satisfaction, doctor effort, and interview location: Evidence from Paraguay*. The World Bank.

Dunsch, F. A., Evans, D. K., Eze-Ajoku, E., & Macis, M. (2017). *Management, Supervision, and Health Care: A Field Experiment*(No. w23749). National Bureau of Economic Research.

Dunsch, F., Evans, D. K., Macis, M., & Wang, Q. (2018). Bias in patient satisfaction surveys: a threat to measuring healthcare quality. *BMJ global health*, *3*(2), e000694.

Evans, D. K., & Welander Tärneberg, A. (2018). Health-care quality and information failure: Evidence from Nigeria. *Health economics*, *27*(3), e90-e93.

Fenton, J. J., Jerant, A. F., Bertakis, K. D., & Franks, P. (2012). The cost of satisfaction: a national study of patient satisfaction, health care utilization, expenditures, and mortality. *Archives of internal medicine*, *172*(5), 405-411.

Jenkinson, C., Coulter, A., & Bruster, S. (2002). The Picker Patient Experience Questionnaire: development and validation using data from in-patient surveys in five countries. *International Journal for Quality in Health Care*, *14*(5), 353-358.

Khamis, K., & Njau, B. (2014). Patients' level of satisfaction on quality of health care at Mwananyamala hospital in Dar es Salaam, Tanzania. *BMC health services research*, *14*(1), 400.

Krosnick, J. (2000). The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Newsletter*, *20*(1), 4-8.

Leonard, K. L. (2008). Is patient satisfaction sensitive to changes in the quality of care? An exploitation of the Hawthorne effect. *Journal of health economics*, *27*(2), 444-459.

Lietz, P. (2010). Research into questionnaire design. *International Journal of Market Research*, *52*(2), 249-272.

Medical Group Management Association. (2016). Medical Practice Management Compensation Survey Report.

Mwanga, D. M. (2013). Factors Affecting Patient Satisfaction at Kenyatta National Hospital, Kenya: A Case Of Cancer Outpatient Clinic. *Nairobi: University of Nairobi*.

Phaswana-Mafuya, N., Davids, A. S., Senekal, I., & Munyaka, S. (2011). Patient satisfaction with primary health care services in a selected district municipality of the Eastern Cape of South Africa. *Modern Approaches to Quality Control [homepage on the Internet]. InTech Publishing*, 85-98.

SafeCare. (2019). SafeCare Standards. Retrieved September 20, 2019, from http://www.safecare.org/index.php?page=safecare-standards.

Steine, S., Finset, A., & Laerum, E. (2001). A new, brief questionnaire (PEQ) developed in primary health care for measuring patients' experience of interaction, emotion and consultation outcome. *Family practice*, *18*(4), 410-418.

Voutilainen, A., Pitkäaho, T., Kvist, T., & Vehviläinen-Julkunen, K. (2016). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of advanced nursing*, *72*(4), 946-957.

World Bank. (2016). *Service delivery indicators*. Washington, DC: World Bank. http://datatopics.worldbank.org/sdi/.

5. Job Preferences of Frontline Health Workers in Ghana: A Discrete Choice Experiment[84]

5.1 Abstract

The lack of supply of adequately skilled and motivated health workers especially in rural areas poses a major obstacle to health service delivery in Ghana, as well as in many other developing countries. In this paper, we present the results of a discrete choice experiment (DCE) conducted with community health officers (CHO) and community health volunteers (CHV), which are extension workers that provide health services and consultations to mostly rural populations.

CHOs and CHVs completed the cadre-specific discrete choice experiment that elicited preferences for attributes of potential job postings. Data was collected from 404 CHOs and 206 CHVs in 8 Ghanaian districts in 4 regions. For CHOs, next to increases in salary, the choice of job posting was most strongly influenced by facility quality, followed by career development opportunities and transport subsidies. Additional supervision showed no effects. For CHVs, next to receiving a monthly stipend, facility quality was also most important, followed by training opportunities.

We are corroborating the notion that other non-financial incentives can have strong effects on job preferences, for example the equipment of the facilities, which includes housing, as well as training and career development and training opportunities. Given that Ghana's health wage bill is already very high, this may open new policy avenues for health workers recruitment and retention to converge towards the aspired universal health coverage.

[84] Co-author of this study is Edit Velenyi (World Bank Senior Economist).

5.2 Introduction & Background

Ghana's economy has grown very rapidly since 2000, with the GDP per capita outpacing population growth at a peak rate of 14.5 percent in 2011 (Wang et al., 2017). However, while improving, the performance of the health sector is lagging the surging economy. There are a variety of reasons for these lagging health system outcomes, both on the demand side (service access & utilization)[85], as well as on the supply side (for example lack of human resources, facilities, pharmaceutical supply chains, and health financing). There are important gaps in service provisions, especially for the extremely poor, which is an expression of the "inverse care law" formulated by Hart (1971), which stipulates that the people that need care services the most receive the least attention. While 96.7% in the highest income quintile benefitted from skilled birth attendance, this was only the case for 46.9% in the lowest quintile (Wang et al., 2017).

Appiah-Denkyira et al. (2013, p. 4) see poor human resources as the main problem: "Perhaps most important, however, it is the lack of skilled service providers, or HRH – particularly in rural areas – that prevents critical health services from being accessed and adequately delivered to those who need them most." Ghana has been a source of (outward) migrating health workers due to the high-quality training and relatively low wages (Stilwell et al., 2004; Willis-Shattuck et al., 2008; Antwi and Phillips, 2013).

This article focuses on ways to improve retention and motivation of community extension health workers, which are a key cadre in the Ghanaian health system. As in many other developing countries, the lack of supply of adequately skilled health workers, especially in rural areas, poses a major service delivery problem in Ghana.

In addition, the current policy trends place increasing emphasis on the role of Community-Based Health Planning & Services (CHPS) zones in expanding service coverage of essential care to all communities. The goal is to "[reach] every community with a basic package of essential health services towards attaining Universal Health Coverage and bridging the access inequity gap by 2020" (GHS, 2016). In general, effective coverage requires that the following conditions are satisfied: (i) there is an adequate operational environment, in terms of infrastructure, physical inputs, and human resources; (ii) providers are technically endowed/skilled; and (iii) providers are motivated to do their best.

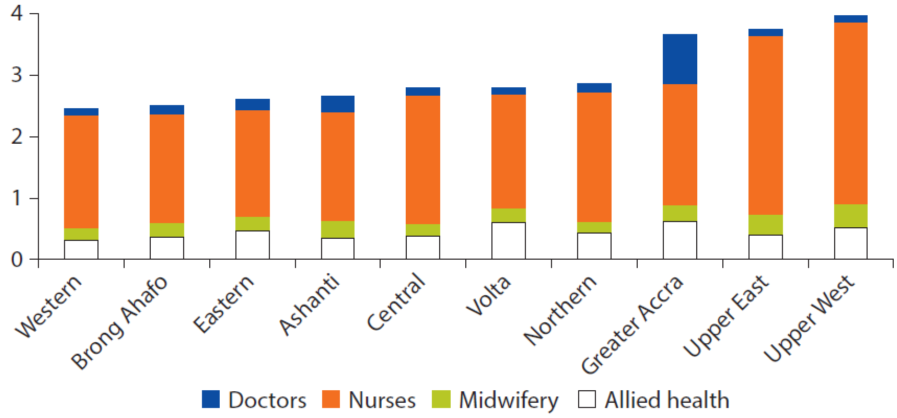
However, recent data shows clear gaps in CHPS-based delivery of maternal and child health and (MCHN) services (e.g. ANC, PNC, referral to skilled delivery, child growth monitoring and immunization) and related inputs, as compared to the standards set in the National CHPS Policy (GHS

[85] See for example Wang et al. (2017) for insights on the important National Health Insurance Scheme (NHIS).

2016). A recent study finds that 46% of CHPS do not provide ANC services; 33% of CHPS do not provide PNC for newborns; 37% do not provide PNC for the mother. About half (47%) of CHPS do not have compounds (World Bank, 2016). This undermines effective coverage as compound possession was highly correlated with the capacity to store essential medicines and rapid diagnostic tests on sight, as well as positively correlated with the frequency with which CHOs performed home visits. Critical equipment and logistics for maternal, child health service delivery is generally low. 78% of CHPS lack a hemoglobin scale, 40% of CHPS do not have an outreach bag or blood pressure meter and 30% do not have a thermometer. The retention of health workers, especially in rural areas, is critical. Currently 69% of CHPS have fewer than 3 Staff, the standard set forth in the National CHPS Policy (World Bank, 2016). As summarized by Heerdegen et al. (2019), health workers in rural areas face higher workloads, professional isolation, unsustainable work environments, lack of opportunities for professional and educational advancement, lack of clear contract terms, poor housing, dearth of opportunities, and good schools for their families (see also Lori et al., 2012 for a study on midwives). These factors result in health workers emigrating or being drawn to work in the large urban centers. 65 percent of doctors and 40 percent of nurses work in Greater Accra or Kumasi, although these areas are home of only 33 percent of the Ghanaian population (Adzei and Atinga, 2012).

The World Health Organisation established that a health worker density of 4.45 per 1,000 people would be needed to be able to achieve UHC (WHO, 2015). Figure 5.1 below shows that in no region in Ghana this threshold is currently achieved (Wang et al., 2017, p. 12).

Figure 5.1: Number of Health Workers per 1,000 People, per Region (2016)



Source: Payroll data, Ministry of Health.

The Government of Ghana acknowledges this problem and has committed itself to improve Ghana's stock, distribution, and performance of health workers (Appiah-Denkyira et al., 2013)

In addition to infrastructure problems, shortcomings in human resources are an important factor explaining utilization rates that – despite some progress – lack behind expectations. While skilled birth attendance overall rose sharply from 47.1 % in 2003 to 73.7 % in 2014, immunization rates declined between 2008 and 2014. The portion of children with fever taking antimalarial medication was lower in 2014 than in 2003 (Wang et al.; 2017). As one cornerstone to achieve UHC, Ghana needs to expand the strategic deployment of health personnel and incentivize the retention of existing staff, especially in rural areas. There is also some evidence that better performance of the health sector can boost economic growth further (Velenyi, 2016).

Antwi and Phillips (2013) show that higher wages can reduce attrition (for workers under 35), thus the results could feed into the calibration of monetary performance incentives for the various cadres but also provide insights into the importance of non-monetary incentives, which a) would often be more affordable, and b) can result in even higher retention rates and performance of health workers but do not have high budget implications (or are budget neutral), especially as Ghana has been struggling with health worker pay and compensations already constituting a large portion of the country's overall health spending (McCoy et al., 2008).[86] A systematic review on potential interventions for increasing the proportion of health workers in rural areas concludes in light of lacking existing research on the issue: "Rigorous evaluation of the effectiveness of various strategies is required to determine the true impact of these interventions and to better inform future policy" (Grobler, 2015, p. 13).

Given these circumstances and to better understand the motivation and related program alternatives, we decided to conduct a discrete choice experiment (DCE) which is highly relevant for HRH policymaking. DCEs have become popular in the field of health economics to elicit preferences (in our case of health workers) as opposed to direct survey responses, which can be severely biased (Arrow et al., 1993). Whereas most studies focus on individual factors that incentivize health workers to take and keep jobs in rural areas, DCEs are a way to see how attributes factor into employment decisions jointly as well as to rank them (Araújo and Maeda, 2013).

[86] In 2006, health worker pay and emoluments absorbed 76% of government spending on health (McCoy et al., 2008).

What are DCEs?

"DCEs are an attribute-based approach to collect SP (stated preference) data. They involve presenting respondents with a sequence of hypothetical scenarios (choice sets) composed by two or more competing alternatives that vary along several attributes, one of which may be the price of the alternative or some approximation for it. In a Lancasterian framework (Lancaster, 1966), it is assumed these attribute levels determine the value (utility) of each alternative. For each choice set, respondents are asked to choose their preferred scenario. It is assumed that individuals will consider all information provided and then select the alternative with the highest utility. Responses enable the analyst to model the probability of an alternative being chosen as a function of the attributes and the socio-economic characteristics of the respondents. This allows an estimation of the relative support that respondents show for the various competing alternatives. Other policy outputs include marginal rates of substitution across nonmonetary attributes as well as WTP or WTAC for an improvement or deterioration of one of those attribute welfare measures for a proposed change in levels of the attributes and predicted uptake or demand." (Ryan et al., 2007, p. 4)

The remainder of the paper is structured as follows.[87] Section 5.3 provides a brief overview of the literature on the preferences of health personnel in developing countries. Section 5.4 presents the data and DCE method, while Section 5.5 specifies the econometric model and the estimations. Sections 5.6 and 5.7 discusses the results for CHOs and CHVs, and Section 5.8 concludes.

5.3 Literature

Human Resources for Health (HRH) issues, such as subpar health worker motivation and high turnover rates are a major obstacle for the delivery of high quality of health care in developing countries, especially in remote areas (WHO, 2010; Dussault and Franceschini, 2006; Araújo and Maeda, 2013). Poor health worker motivation results in worse health outcomes for the population due to absenteeism, low team morale, and higher worker turnover. Chaudhury et al. (2006) find that a shocking 35 percent of public health workers were absent during random unannounced visits in a series of low- and middle-income countries.

In contrast, better motivated workers have stronger relationships with the communities they serve, build their competencies, strengthen teamwork, and are less likely to leave their posts (Buykx et al, 2010; Bonenberger et al., 2014). Alhassan et al. (2013) find that staff satisfaction is positively correlated with working conditions and higher-level quality of care.

[87] The paper follows a similar structure of Kolstad (2011).

As a consequence of poor conditions, Ghanaian health workers also migrate to other countries that benefit from the well-trained health workers (Willis-Shattuck et al., 2008; Antwi and Phillips, 2013), which puts further stress on the Ghanaian health care system.

There are very few studies that investigate the impacts of higher salaries in the public sector on job performance, mainly because it is hard to vary salaries among public sector employees (De Ree et al., 2017). In one such study, Dal Bo et al. (2013) show that higher salaries attract better qualified candidates in Mexico during the application process and the government more easily is able to fill vacancies. In Zambia, Ashraf et al (2015) experimented with two ways to advertise a position for health workers. One advertisement stressed the social impact the workers would have while the other stressed career advancement. Candidates were more qualified in the group that respondent to the latter poster, stressing promotions and career advancement (as measured by high school test scores and past educational performance). Prosocial motivation did not vary between the groups. Finan et al. (2015) provide a good overview over recent research, especially in the field of public sector recruitment.

Most evidence focuses on financial rewards based on performance rather than unconditional pay raises. In the field of behavioral economics there is vast evidence that financial rewards can induce higher effort (e.g. Camerer et al., 1999; Ashraf et al, 2014; Garbers et al, 2014; DellaVigna et al., 2017). Financial incentives seem especially effective for judgement tasks, and lower complexity tasks where increased effort also augments performance (Camerer et al., 1999). However, the impacts seem to be highly specific to the nature of the tasks performed, the level of effort and skills required, and prior intrinsic motivation levels.

There are also concerns around negative effects of financial incentives through the potential to crowd-out intrinsic motivation (Deci, 1971; Bénabou et al., 2006; Kamenica, 2012), especially for social tasks, although this is not well substantiated yet in the realm of government work. Gneezy et al. (2000) find evidence of the phenomenon in prosocial tasks like volunteering work and Mellstrom et al. (2008) for blood donations.[88] Deserrano (2015) finds that financial incentives in a job description lead to less pro-socially motivated individuals applying for a health promoter position in Uganda. Skill levels of applicants however were the same in the groups expecting a lower salary when compared to the group expecting a higher salary.

There are a few examples of impact evaluations that specifically measure the impact of financial incentives on the effort of health care providers in low- or medium-income countries. An experimental impact evaluation in Rwanda found that financial rewards can increase performance of health workers, especially for lower effort tasks that draw higher payouts (Basinga et al., 2011).

[88] However, there is some evidence that shows the opposite (Lacetera et al., 2013).

A study on doctors in the Philippines found increased clinical knowledge of medical procedures, as these were rewarded (Peabody et al., 2011). Olken et al. (2014) demonstrate that incentives to villages in Indonesia increased school enrollment rates and health effects in the short-run. De Walque et al. (2015) find that performance incentives for health workers increases HIV testing in Rwanda. In a non-randomized study, Sun et al. (2016) show that prescription quality augments after the roll-out of a pay-for-performance program. In a systematic review, Willis-Shattuck et al. (2008) find that both monetary and non-financial incentives impact motivation and retention in low- and middle-income countries.

Mathauer et al. (2006) find that health workers in Africa overall are "strongly guided by their professional conscience", and that many health workers are "unable to satisfy their professional conscience and impeded in pursuing their vocation due to lack of means and supplies and due to inadequate or inappropriately applied human resources management (HRM) tools." They conclude that recognition, career development opportunities, and further qualification can further strengthen the professional ethos of health workers. Indeed, there are a few well identified experimental studies that buttress this claim, Ashraf et al. (2014) show that a non-financial recognition program was more effective to incentivize health extension workers to sell condoms than financial incentives. Dunsch et al. (2017) find that in Nigeria, periodic supportive supervision can increase quality of care in health clinics without additional financial expenditure in infrastructure or equipment. Schlechter et al. (2015) show that non-financial rewards have impacts on prospective employees' perceived attractiveness of a job offering.[89] Overall, it is hard to ascertain whether non-financial reward systems work or not, as 1) there are not too many experimental studies available, and 2) non-financial rewards are hardly comparable (unlike financial rewards) which makes a synthesis very difficult.

Discrete Choice Experiments have become a popular instrument to elicit preferences of health workers in recent years, especially the trade-offs between financial and non-financial rewards, as well as to elicit the willingness to pay for tradeoffs. "It forces respondents to choose between two scenarios of employment packages, thereby making trade-offs and identifying hierarchical preferences" (Lori et al., 2012, p. 3). De Bekker-Grob et al. (2012) and Clark et al. (2014) conducted a literature review summarizing the state of the discipline. Most empirical studies focus on retention and preferences of health workers in hospitals or health centers.

In a study on Ghanaian midwives, Lori et al. (2012) found that educational opportunities were most important to entice workers to be posted in rural areas, followed by the quality of facilities, and improved quality education for children.

[89] This study is, however, based on a convenience sampling technique.

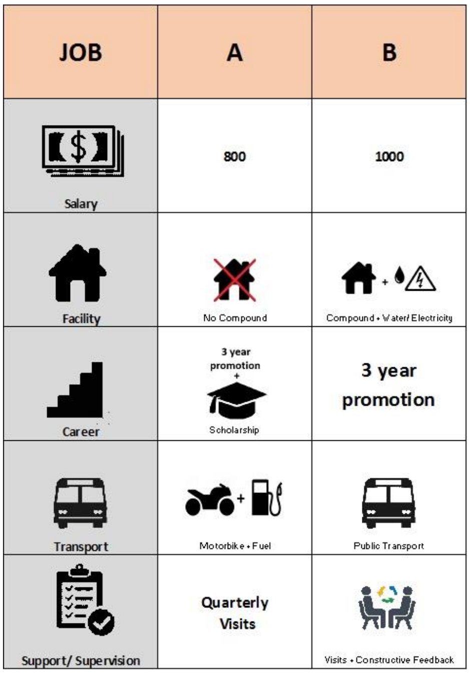
Takemura et al. (2016) conducted a DCE with clinical officers in Kenya and find that a 1-year study leave after 3 years of service would have the highest impact on retention, followed by good quality health infrastructure and a 30% pay raise. In a study in China (Song et al., 2015), doctors and nurses revealed highest preferences for sufficient welfare benefits, sufficient essential equipment, and respect from the community. For 4 provinces in Mozambique, a DCE elucidated the preferences of non-physician health workers (Honda and Vio, 2015). Here, housing was the most relevant attribute for choosing a job, followed by formal education opportunities, and the availability of equipment and medicine at the facility. Smitz et al (2016) investigate what contributes to retention of health worker retention in rural areas of Timor-Leste. Doctors showed the highest satisfaction related to professional development opportunities and good working conditions. For nurses and midwifes, skill upgrading emerged as the most important attribute. A DCE with midwives and nurses in Peru (Huicho et al., 2012) shows that the health workers preferred urban jobs over rural postings, although a set of financial and non-financial incentives proved to double uptake of rural jobs in policy simulations. Kolstad (2011) finds that offering continuing education to newly hired clinical officers in Tanzania was the most powerful incentive to make rural postings more attractive, followed by salary increases and hardship allowances.

This is, to our knowledge, the first DCE study that focuses on the job preferences on (rural) health extension workers and volunteers (as opposed to health workers in a clinical setting).

5.4. The Set-Up of the DCE – Tools and Methods

We conducted the DCE with 404 Community Health Officers (CHOs) and 206 Community Health Volunteers (CHVs) in 8 districts of Ghana.[90] The CHOs and CHVs are anchored in the Community Health Planning Services (CHPS) zone, which is the main platform for service delivery in remote areas. CHPS zones are administrative units, encompassing roughly 3000 to 4500 people. Each CHPS zone has one Community Health Team (CHT) that provides health services to the population, including basic preventative care, curative care, and promotional health services in homes or compounds. Severe cases are referred to hospitals (Shiratori et al., 2016). CHTs are made up of 1-2 CHOs and ca. 4 CHVs. CHOs are paid GHS staff while CHVs are recruited from the communities and receive no formal pay.

Figure 5.2 – A simplified example of a choice between two hypothetical jobs made by the respondents on the Android Tablet Device



JOB	A	B
Salary	800	1000
Facility	No Compound	Compound + Water/ Electricity
Career	3 year promotion + Scholarship	3 year promotion
Transport	Motorbike + Fuel	Public Transport
Support/ Supervision	Quarterly Visits	Visits + Constructive Feedback

During the survey, each health worker received 8 choice sets. The DCE module was part of a baseline questionnaire for an impact evaluation aimed at measuring the impacts of a World Bank community-based performance-based financing (CPBF) project pilot. We conducted the survey using electronic tablet computers. Because many of the respondents were illiterate, we used graphical illustrations of the choice alternatives (icons) along with verbal explanations. Figure 5.2 shows an example of such as choice set with the alternatives "Job A" and "Job B". The participation in this study was voluntary and the respondents were not compensated for their collaboration.

In preparation for the DCE design (prior to the quantitative data collection that is the basis of the DCE), we also conducted focus group discussions (FGDs). These FGDs were conducted with CHOs and CHVs in a sample of 7 districts in 4 regions (Northern, Upper East, Upper West and Volta), between October and November 2015. The main goal of the Focus Group Discussions was to find out what the main factors are that motivate CHOs and CHVs in their respective functions. In each of the 7 districts, two groups of CHOs and one group of CHVs were interviewed. Each group

[90] Agortime Ziope, Kadjebi (Volta Region), West Gonja, North Gonja (Northern Region), Bawku West, Talensi (Upper East), Lawra and Nandom (Upper West).

consisted of 8 individuals. The total sample size was therefore 112 CHOs and 56 CHVs.[91] The FGDs were facilitated by two Ghanaian consultants who had previous experience conducting FGDs in the target regions and who speak the local languages. The FGDs were conducted separately for the CHVs and the CHOs, and supervisors of the CHOs were not permitted to assist so that each group would speak with ease.

The FGDs were semi-structured, meaning that the facilitators had a script that they read from, but they also allowed for the conversations to flow in a fluid way. Nonetheless, the Facilitators were instructed to revert to the original structure of the script and guide the conversation in a way that that all topics would be addressed. Following the discussions, the facilitators were asked to fill out a form in which they rated the top 10 attributes that impact their job motivation. The facilitator guided this process but ensured that there was consensus between the participants; by going through each attribute that was addressed during the discussions one by one to validate the ranking.

For the selection of attributes that were eventually used in the DCE, the FGD ranking was used, as well as a triangulation with the Government, as it was important that the attributes that were included were also policy relevant. In order to restrict the complexity of the choice sets, we capped the number of attributes at 5 (with a maximum of 4 levels per attributes), which is in line with standards in health economics (Scott, 2002). Table 5.2 shows the attributes and their levels. These are in line with the WHO's (2010) global policy recommendations for interventions focused on rural retention: education, regulation, financial incentives, and personal and professional mechanisms (see also Shiratori et al., 2016).

Similar to, for example, Kolstad (2011), a D-optimal (unlabeled binary choice) design was developed to construct the hypothetical choice situations. The design was developed "to achieve an efficient combination of orthogonality, level balance and minimum overlap" (Kolstad, 2011, p. 200).

The final choice sets consisted of 40 choices between pairs of jobs. The 40 choice sets were divided into 5 blocks of 8 choices each. Each respondent was then randomly assigned with each of these blocks of 8 decisions during the survey. The options are simply named job A and job B, akin to most studies in the health economics literature (Blaauw et al., 2010).

[91] In the Volta region, the FGDs were conducted in one district, whilst in the other regions, the FGDs were conducted in two districts for each region.

Descriptive Statistics for CHOs and CHVs

The demographic characteristics for the CHOs and CHVs, based on data from our surveys, are shown in table 5.1. The mean age of the CHOs is 29.4 years, 53% are female, and the mean monthly salary is 921.53 Ghanaian Cedi (ca. 175 USD in 2019). 83% are of Christian faith. Interestingly only 4% were born in the CHPS zone they are currently working in. 45% of the CHOs are married. 76% have received tertiary (medical) education and 47% have children. 47% report to be the head of their household.

The mean age for CHVs is higher than for CHOs with 42.5 years (SD = 13.7). 35% of CHVs are female and 68% state that their religion is Christianity. In contrast to the CHOs, 65% were born in the CHPS zone they are currently volunteering for. 80% of the CHVs are currently married. Only 4% of CHVs have received tertiary education, and 90% have children. 65% of the CHVs report to be the head of their household.

Table 5.1: Descriptive Statistics

	CHOs	CHVs
Age (mean)	29.4	42.5
% Female	53 %	35 %
Monthly average salary	921.5	-
Religion (Christian)	83 %	68 %
Born in CHPS Zone of work	4 %	65 %
Married	45 %	80 %
Received tertiary education	76 %	4 %
Has children	47 %	90 %
Head of household (yes)	47 %	65 %

Table 5.2 – Attributes and Levels

CHOs		**CHVs**	
Attribute 1	Salary	Attribute 1	Monthly stipend
Level 1	800 Cedis	Level 1	No monthly stipend
Level 2	900 Cedis	Level 2	40 cedis/month
Level 3	1000 Cedis	Level 3	80 cedis/month
Level 4	1100 Cedis	Level 4	120 cedis/month
Attribute 2	Facility	Attribute 2	Facility
Level 1	No compound available in the CHPS zone	Level 1	No compound / limited equipment
Level 2	Compound available with electricity and potable water	Level 2	Compound
Level 3	Compound available with electricity and potable water, AND complete list of essential equipment	Level 3	Compound with equipment
Level 4	Compound available, including electricity and water, AND complete list of essential equipment AND free accommodation for staff		
Attribute 3	Career	Attribute 3	Trainings
Level 1	First promotion after 5 years	Level 1	No trainings
Level 2	First promotion after 3 years	Level 2	Two free trainings a year (including travel to training site)
Level 3	First promotion after 3 years + scholarship for further education based on exceptional performance	Level 3	Four free trainings
Attribute 4	Transport	Attribute 4	Transport
Level 1	No transport available/ no transport subsidy	Level 1	No subsidized transport
Level 2	Fuel subsidy for public transport, sufficient for all monthly outreaches and home visits	Level 2	Monthly stipend for outreaches with public transport
Level 3	Motorbike (A100) + fuel subsidy for motorbike	Level 3	Bicycles for CHVs for home visits
Attribute 5	Supervision	Attribute 5	Awards
Level 1	Quarterly visits by sub-district/ district supervisors (Critical Feedback)	Level 1	None
Level 2	Quarterly visits by sub-district/ district supervisors with Constructive feedback to CHOs to improve performance	Level 2	CHV of the Month awards
Level 3	Quarterly visits by sub-district/ district supervisors with constructive evaluation of CHT performance AND accompanied outreaches	Level 3	Gift bag based on performance (clothes, rice, etc.)
		Level 4	Gift bag + free NHIA enrollment

5.5 Specification of the DCE Model

DCE data utilizes the random utility theory model that stipulates that the utility function of an individual cannot be directly observed by researchers (McFadden 1973; McFadden 1986).

The utility function is modelled to have an explainable component and an unexplainable (random) component. The random component is being analyzed using a probabilistic framework. The utility of the *n*th respondent is stated by

$$U_n = \lambda \times V_n + \varepsilon_n$$

with V_j as the systematic component depicting observed influences of attributes and levels (also referred to as the representative utility), ε_j as the stochastic component reflecting unobserved influences which is treated as being random. It is assumed that ε_j is a distributed IID extreme value. λ is the scale parameter reflecting the variance of the unobserved influences (Hensher et al., 2005; as stated in Hoefman et al., 2014). It is assumed that everyone, when having the choice between 2 (or more) alternatives, chooses the one that maximizes their utility (Hauber et al., 2016). The underlying assumption of the model is that health workers always choose the alternative that maximizes their utility.

The specified model was the following, where U_{nj} describes the representative utility of person *n* from job alternative *j*, and *x* is a vector of the attributes of the job *j.*

$$U_{nj} = \beta' \mathrm{x}_{nj} + \varepsilon_{nj}$$

For analysis, all attributes were coded as dummy variables, with the lowest level of each attribute left out of the regression. The "basic job" for CHOs in our model is a job that pays 800 Cedi, there is no CHPS Compound, promotions occur after 5 years, there is no subsidy for transportation, and basic supervision takes place 4 times a year. While this appears to be bleak, it reflects often the reality for rural health workers in Ghana. In fact, health workers I spoke with and observed at their work complained to me that their "office" was "under the mango tree."

The basic position for CHVs is one without a monthly stipend, no compound and limited equipment, no trainings, no transport subsidies, and no performance awards.

We estimated main effects using only the attribute variables in a conditional logit model (using Stata's clogit command) to consider the grouping of each set of job choices for an individual. In addition, we estimated separate (split sample) models using gender, age (over/under the median), and higher and lower income groups (over/under 900 Cedis per month). Due to the panel nature and the likely correlation across choice occasions, standard errors were clustered at the individual respondent level (Cameron and Trivedi, 2010; Lancsar et al., 2017).

In addition to the main regressions, we calculate the willingness-to-pay ("WTP") for attribute levels. It is measured as the amount of salary or stipend a CHO or CHV is willing to give up in order to receive a higher level of another job attribute.

We also run "policy simulations" that estimate the proportion of CHOs or CHVs that would prefer a hypothetical job vs. the most "basic" job. In order to simulate policy impact, we change only one attribute at a time while holding all others constant and observe how the probabilities change.

The logit probability of choosing alternative *i* rather than alternative *j* is given by

$$P_i = \frac{e^{\beta' x_i}}{\sum e^{\beta' x_j}}$$

where x is a vector of attribute coefficients (Ryan et al., 2012). As stated in Kolstad (2011), "this effect can be understood as the change in probability of taking the baseline job because of a change in the level of one of the job attributes."

5.6 Results – CHOs

This section describes the results of the DCE for CHOs. With the logistic regression results (Table 5.3), we can calculate the willingness to pay (WTP; Table 5.4) for certain attributes, as well as the change of probability of favoring a job, when one of the attributes changes ("policy simulation", Table 5.5). Both can be very informative for policymakers that aim to improve the work conditions for frontline health workers in a resource constrained environment such as rural Ghana.

Other than for "supervision", the estimated employment attribute coefficients were all significant and of the anticipated (positive) sign for the CHOs (Table 5.3). The equipment levels of the facility, career opportunities and transport subsidies and support were major predictors for job choice. Increased salaries are also significant.[92] Generally, the results indicate that the CHOs received a higher level of utility from the better attribute level and made rational choices.

[92] Surprisingly, when running the analysis with salary as a dummy dependent variable, only the highest salary level (Level 4) was significant, while the smaller increases (step 2 and 3) were not.

Table 5.3: Preferences for Job Attributes – CHOs

Variable	Coefficient	Robust SE	P > |z|
Salary	0.003	0.000	0.000
Facility Level 4	2.058	0.167	0.000
Facility Level 3	1.536	0.162	0.000
Facility Level 2	1.168	0.139	0.000
Career Level 3	1.040	0.165	0.000
Career Level 2	0.600	0.107	0.000
Transport Level 3	0.915	0.138	0.000
Transport Level 2	0.686	0.107	0.000
Support Level 3	0.050	0.114	0.661
Support Level 2	0.139	0.085	0.102
Constant	0.017	0.065	0.789
Log likelihood	-958.19106	Prob > Chi2	0.0000
Number of obs.	3296	Pseudo R^2	0.1612
Wald chi^2 (11)	204.76		

CHO – Willingness to Pay

The willingness-to-pay (WTP) calculations (Figure 5.3 and Table 5.4) illustrate which the most impactful attributes are. WTP measures should be interpreted as the willingness to give up additional salary (a potential raise) in exchange for an attribute using the actual average salary of 921 Cedi as a benchmark.

When calculating willingness-to-pay measures, we see that a fully equipped facility with sleeping quarters ("Facility Level 4") are of utmost importance for the CHOs. They would be willing to forgo a raise of 85.6% of their actual average salary to be able to work in a fully equipped facility which includes electricity, water, essential equipment and overnight accommodation for staff. This is in line with the results from the preparatory focus group discussions (see annex table A5.1) where the issue of inadequacy of facilities was often raised as well as with other results in the literature (see e.g. Takemura et al., 2016). Even if upgraded compounds did not include free accommodation CHOs were still willing to forgo a lot of salary to work in an environment with full availability of essential equipment (63.9%). Even for "Facility Level 2" (*"Compound available with electricity and water"*) the WTO is still high with 447.7 Cedi (48.6% of salary), which is higher than the highest "Career" or "Transport" attributes. These results show the importance to upgrade of compounds, not just for the sake of having better health infrastructure, but also its motivational effect on rural health workers.

Compared to the "base job" where the first promotion occurs after 5 years of service, "Career Level 3" *("First promotion after 3 years + scholarship for further education based on exceptional performance")* has a WTP of 398.6 Cedis, equivalent to 43.3% of salary, while "Transport Level 3" *("Motorbike A100 + fuel subsidy for motorbike"* – compared to no transport subsidy for the base job) has a WTP of 350.7, which corresponds to 38.1% of the average salary. Levels 2 for "Career" (first promotion after 3 years) and "Transportation" (subsidy for public transportation) still show high willingness to pay with 25% and 28.5% respectively. The importance of improving career opportunities is in line with previous research by Kwansah et al. (2012) and Snow et al. (2011) who also found that low chances for career development were main reasons for doctors to prefer urban areas.

The results show that the health workers would not trade salary for increased support or supervision (quarterly visits by supervisors with constructive feedback on Level 2 + accompanied outreaches at level 3), which might be surprising as other studies report that "feeling forgotten" at rural posting was a source for preferring urban posting (Kwansah et al., 2012) and this was also mentioned during the preparatory focus group discussions.

In sum the results show that upgraded facilities are most important for CHOs, followed by better career incentives and transport subsidies. Increased supportive supervision was not important to CHOs.

Figure 5.3: Willingness to Pay - CHOs

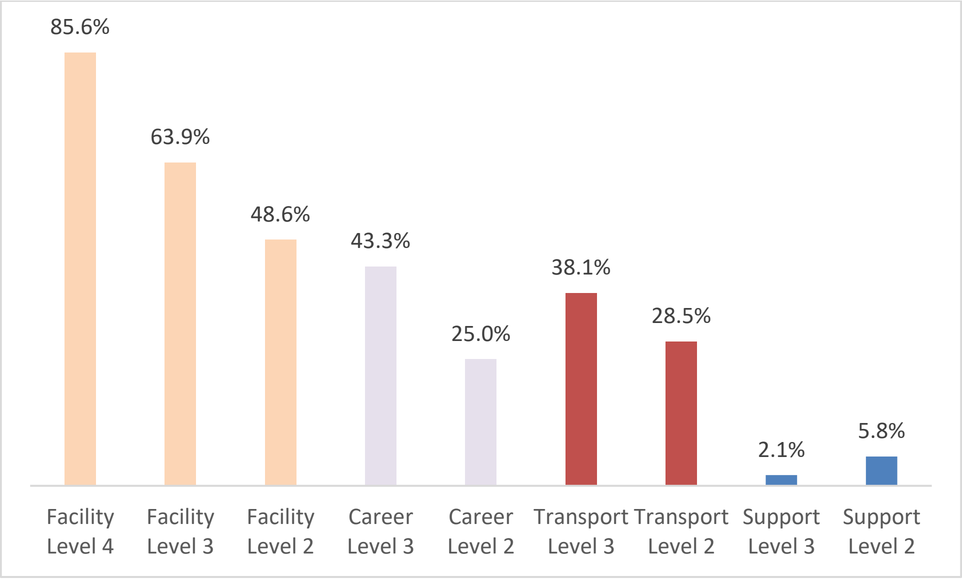


Table 5.4: Willingness-to-pay (WTP) – CHOs

	Willingness to pay	Upper limit	Lower limit	% of actual average salary (921.53)
Facility Level 4	789	997	581	85.6%
Facility Level 3	589	753	424	63.9%
Facility Level 2	448	578	317	48.6%
Career Level 3	399	521	276	43.3%
Career Level 2	230	314	146	25.0%
Transport Level 3	351	471	230	38.1%
Transport Level 2	263	356	170	28.5%
Support Level 3	19	103	-64	2.1%
Support Level 2	53	115	-8	5.8%
Number of observations	3296			
Wald Chi^2 (11)	204.76			
Prob > Chi2	0.000			
Pseudo R^2	0.1612			

Marginal Probabilities – Policy simulations

Table 5.5 shows the changes in probabilities to accept a job, when only one attribute is changed from the most basic job profile (see section 5.5 for the basic job profile description). This “policy simulation” shows that – in line with the willingness to pay results – that upgrading the facility, not increasing the salary, seems to be by far the most powerful policy instrument. Providing a CHPS compound with water and electricity already increases the probability to accept the hypothetical job by 53% (Facility Level 2). Providing complete and functional equipment increases the probability by another 12 percentage points (to 65%; Level 3). Adding full accommodation to the compound and the equipment adds another 12 percentage points (to 77%; Level 4). Given the fact that the health posts are often severely under-equipped, this is not a surprising finding.

Career opportunities are also important for CHOs. Receiving the first promotion after 3 years instead of 5 years increases the probability to prefer a job by 29% (Level 2). Adding a scholarship conditional on exceptional performance after 3 years adds another 19 percentage points to the probability (Level 3). Providing a monthly fuel subsidy for outreaches increases the probability of choosing a job by 33% (Level 2). Providing motorbikes (AG100 model; Level 3) adds another 10 percentage points (to a 43% total). We see a linear increase of the probability of accepting a job with the rise in salary. Offering a salary of 900 Cedi (compared to a base of 800 Cedi) increases the probability of job choice by 13 percentage points. Each additional 100 Cedi increment (to 1000 and 1100) increases the probability by another ca. 13 percentage points each. This shows that, while salary increases are effective, they need to be high in order to match the effects of the other non-financial incentives. Better facilities and equipment, a faster promotion and the potential for a scholarship, as well as providing a motorbike in additional to fuel subsidies for outreaches beat out the highest salary increase. The results of this study run counter to a similar study done with CHOs by Shiratori et al. (2016) who finds that a 50% salary increase from the basic job only leads to a 6% uptick in probability to take up a rural posting. (These results are however sensitive to the entire set of chosen attributes for the DCE and also the definition of the base job.)

Subgroup analysis (Table 5.6) shows that that male health workers are more incentivized by a higher salary than females. Men and women are more balanced on the "facility" variable, but women are more inclined to prefer a job than men when offered better career advancement prospects and transport subsidies.

Figure 5.4: Changes in probabilities of choice compared to base job - CHOs

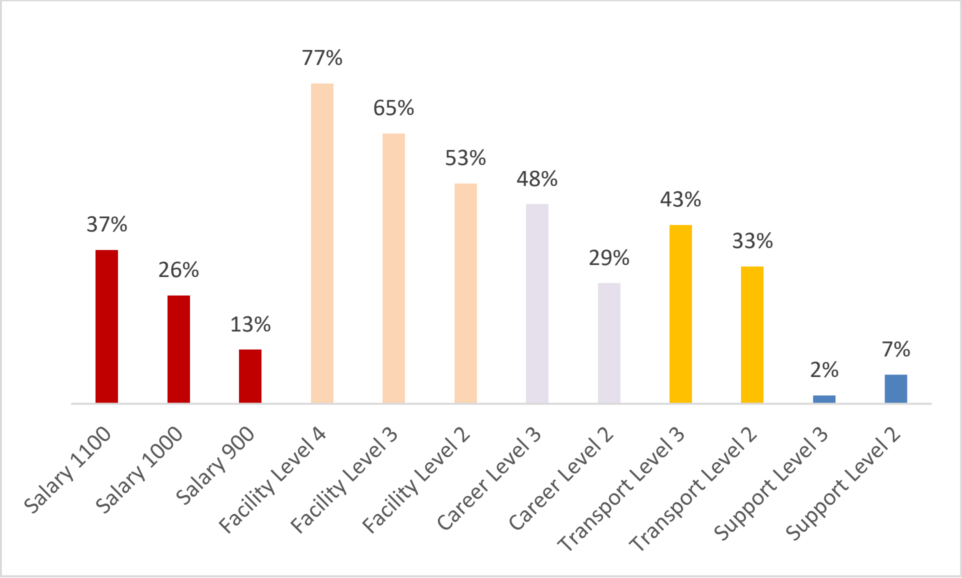


Table 5.5: Changes in probabilities of choice compared to base job - CHOs

Change from baseline scenario	Probability (in %)	Standard Error	P > |z|
Salary 1100	37%	0.050	0.000
Salary 1000	26%	0.036	0.000
Salary 900	13%	0.019	0.000
Facility Level 4	77%	0.033	0.000
Facility Level 3	65%	0.047	0.000
Facility Level 2	53%	0.050	0.000
Career Level 3	48%	0.064	0.000
Career Level 2	29%	0.049	0.000
Transport Level 3	43%	0.056	0.000
Transport Level 2	33%	0.048	0.000
Support Level 3	2%	0.057	0.660
Support Level 2	7%	0.042	0.101
Number of observations	3296		
Wald Chi^2 (11)	204.76		
Prob > Chi2	0.000		
Pseudo R^2	0.1612		

Table 5.6:
Subgroups – CHO – Changes in Probabilities

Variable	Male	Female	Age, older	Age, younger	Income, high	Income, low
Salary 1100	0.408***	0.338***	0.361***	0.408***	0.401***	0.343***
	(0.0695)	(0.0716)	(0.0692)	(0.0691)	(0.0648)	(0.0741)
Salary 1000	0.281***	0.231***	0.247***	0.281***	0.276***	0.234***
	(0.0512)	(0.0510)	(0.0498)	(0.0509)	(0.0476)	(0.0529)
Salary 900	0.143***	0.117***	0.125***	0.143***	0.141***	0.119***
	(0.0272)	(0.0266)	(0.0261)	(0.0270)	(0.0252)	(0.0276)
Facility Level 4	0.771***	0.780***	0.782***	0.771***	0.772***	0.778***
	(0.0526)	(0.0427)	(0.0455)	(0.0484)	(0.0500)	(0.0449)
Facility Level 3	0.649***	0.647***	0.670***	0.630***	0.653***	0.645***
	(0.0775)	(0.0577)	(0.0623)	(0.0711)	(0.0709)	(0.0621)
Facility Level 2	0.504***	0.547***	0.559***	0.479***	0.470***	0.586***
	(0.0802)	(0.0658)	(0.0652)	(0.0793)	(0.0755)	(0.0627)
Career Level 3	0.564***	0.385***	0.515***	0.439***	0.559***	0.386***
	(0.0880)	(0.0915)	(0.0865)	(0.0932)	(0.0755)	(0.104)
Career Level 2	0.297***	0.285***	0.275***	0.325***	0.359***	0.219**
	(0.0752)	(0.0653)	(0.0704)	(0.0644)	(0.0625)	(0.0764)
Transport Level 3	0.491***	0.365***	0.401***	0.481***	0.427***	0.438***
	(0.0819)	(0.0781)	(0.0781)	(0.0799)	(0.0838)	(0.0756)
Transport Level 2	0.345***	0.316***	0.333***	0.335***	0.354***	0.321***
	(0.0649)	(0.0705)	(0.0678)	(0.0706)	(0.0670)	(0.0670)
Support Level 3	0.115	-0.0582	0.0263	0.00476	0.103	-0.0801
	(0.0780)	(0.0821)	(0.0723)	(0.0928)	(0.0755)	(0.0815)
Support Level 2	0.173**	-0.0271	0.0586	0.0734	0.0899	0.0330
	(0.0558)	(0.0607)	(0.0573)	(0.0666)	(0.0563)	(0.0618)
Number of observations	1552	1744	1840	1456	1568	1728
Wald Chi^2 (11)	89.68	137.28	97.31	128.94	99.80	144.73
Prob > Chi2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Pseudo R^2	0.1546	0.1770	0.1594	0.1703	0.1534	0.1774

5.7 Results – CHVs

For the CHVs, all attribute coefficients are significant at the 99% level and positive (Table 5.7).

In line with the results for CHOs, CHVs express the highest WTP for facility improvements (Compound with equipment; Level 3), with 142.8 Cedis, followed by the "Additional Training" attributes (Level 3: 124.5; Level 2: 121.3; Table 5.8). For the training attributes, we see that there is little additional WTP for Training Level 3 over Level 2 (4 free trainings vs. 2 free trainings). This shows that while there is a large preference for receiving training at all, the marginal utility is decreasing. For simply being able to work out of a compound, CHVs would forgo a stipend of 105.1 Cedi.

Transport subsidies are a bit less attractive with a WTP of 93.1 to have bicycles for home visits (Transport Level 3) and 64.8 to receive a stipend that covers public transportation (Level 2).

Non-financial incentives also had an effect. The WTP to receive a CHV of the month award was 60.6 (Level 2). This may be of particular interest for policy makers, since this virtually free recognition, while small, can have measurable effects on the willingness to engage as a volunteer. To receive a gift bag based on performance that may include some clothes or food like rice showed a WTO of 72.3 Cedi (Level 3) and the gift bag in addition to free enrollment with the National Health Insurance Agency (NHIA; Level 4) exhibited a WTP of 94.2 Cedi.

Table 5.7: Preferences for Job Attributes – CHVs

Variable	Coefficient	Robust SE	P > |z|
Stipend	0.011	0.001	0.000
Facility Level 3	1.499	0.114	0.000
Facility Level 2	1.104	0.071	0.000
Transport Level 3	0.977	0.090	0.000
Transport Level 2	0.681	0.075	0.000
Training Level 3	1.307	0.096	0.000
Training Level 2	1.274	0.081	0.000
Award Level 4	0.989	0.101	0.000
Award Level 3	0.759	0.100	0.000
Award Level 2	0.637	0.085	0.000
Constant	-0.115	0.052	0.028
Log likelihood	-1823.9529	Prob > Chi2	0.0000
Number of obs.	6464	Pseudo R^2	0.1858
Wald chi^2 (11)	463.59		

Figure 5.5: Willingness to Pay - CHVs

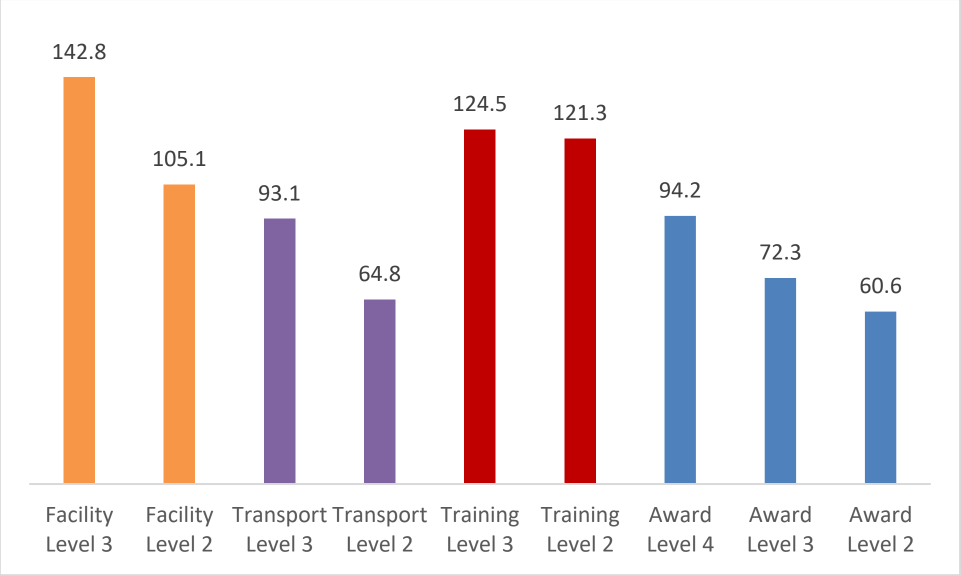


Table 5.8: Willingness-to-pay (WTP) – CHVs

	Willingness to pay	Upper limit	Lower limit
Facility Level 3	142.8	166.4	119.1
Facility Level 2	105.1	126.0	84.2
Transport Level 3	93.1	109.5	76.7
Transport Level 2	64.8	80.5	49.2
Training Level 3	124.5	145.8	103.1
Training Level 2	121.3	145.5	97.1
Award Level 4	94.2	113.2	75.1
Award Level 3	72.3	90.6	54.0
Award Level 2	60.6	77.5	43.8
Number of observations	6464		
Wald Chi^2 (11)	463.59		
Prob > Chi2	0.0000		
Pseudo R^2	0.1858		

Table 5.9 shows "policy simulations" that show how the probabilities of accepting a posting change if only one attribute is changed from the basic scenario (holding all others constant). Similar to the WTP results, as for the CHOs, upgraded facilities exhibit the highest change in probability to prefer a position (63.5% for Facility Level 3) followed by the Training Level 3 (4 free trainings; 57.4%) and 2 (2 free trainings) and the highest stipend level of 120 Cedi (55.8%). The high results for the Level 2 training variable shows that even a few offered trainings can go a long way to motivate the volunteers.

Surprisingly, a small stipend of 40 Cedi had the smallest effect (20.7%) showing that the volunteers do not care much about being paid when compared to the other potential benefits. This also explains why the WTP results are so high.

Also, of lesser interest are Transport Level 2 (a monthly public transport stipend for outreach) and installing a "CHV of the month" award as a non-financial incentive (Award Level 2).

Figure 5.6: Changes in probabilities of choice compared to base job - CHVs

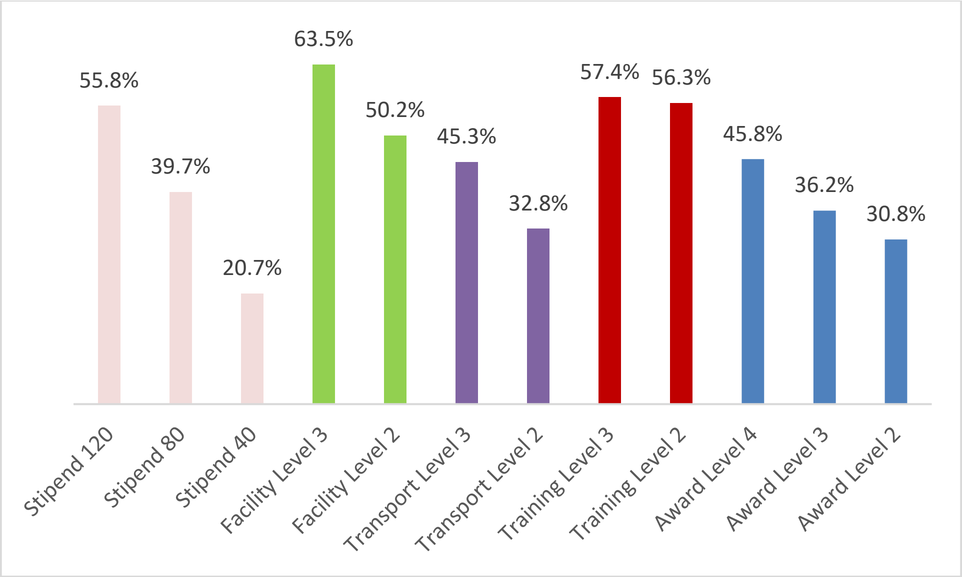


Table 5.9: Changes in probabilities of choice compared to base job - CHVs

Change from base line	Probability	Standard Error	P > |z|
Stipend 120	0.558	0.047	0.000
Stipend 80	0.397	0.038	0.000
Stipend 40	0.207	0.022	0.000
Facility Level 3	0.635	0.034	0.000
Facility Level 2	0.502	0.027	0.000
Transport Level 3	0.453	0.036	0.000
Transport Level 2	0.328	0.034	0.000
Training Level 3	0.574	0.032	0.000
Training Level 2	0.563	0.028	0.000
Award Level 4	0.458	0.040	0.000
Award Level 3	0.362	0.044	0.000
Award Level 2	0.308	0.039	0.000
Number of observations	6464		
Wald Chi^2 (11)	463.59		
Prob > Chi2	0.0000		
Pseudo R^2	0.1858		

* The basic job has no stipend, no compound and limited equipment, no trainings, no subsidized transportation and no awards based on performance.

Table 5.10 shows subgroup analysis for the marginal probabilities, comparing men and women, and older and younger CHVs (over/under 41 years, which is the median). As for the CHOs, men seem to be slightly more incentivized by higher stipends. They are also more inclined to choose a job with an award system than females. Women exhibit higher probabilities to prefer a job for the facility variable.

Younger CHVs prefer jobs with higher stipends, the 120 Cedi stipend increases the probability of younger workers to 64.5% to favor the job. Younger workers also prefer the award scheme and the training attributes.

Table 5.10: Subgroups – CHV

Variable	Main	Male	Female	Age, older	Age, younger
Stipend 120	0.558	0.583*** (0.0560)	0.532*** (0.0844)	0.463*** (0.0800)	0.645*** (0.0527)
Stipend 80	0.397	0.418*** (0.0467)	0.376*** (0.0674)	0.322*** (0.0608)	0.471*** (0.0469)
Stipend 40	0.207	0.219*** (0.0269)	0.195*** (0.0377)	0.165*** (0.0330)	0.250*** (0.0282)
Facility Level 3	0.635	0.610*** (0.0416)	0.695*** (0.0567)	0.617*** (0.0527)	0.657*** (0.0442)
Facility Level 2	0.502	0.466*** (0.0330)	0.577*** (0.0446)	0.462*** (0.0371)	0.542*** (0.0381)
Transport Level 3	0.453	0.464*** (0.0442)	0.449*** (0.0640)	0.447*** (0.0514)	0.464*** (0.0513)
Transport Level 2	0.328	0.374*** (0.0412)	0.244*** (0.0579)	0.290*** (0.0500)	0.368*** (0.0455)
Training Level 3	0.574	0.575*** (0.0388)	0.589*** (0.0578)	0.528*** (0.0479)	0.621*** (0.0429)
Training Level 2	0.563	0.561*** (0.0329)	0.581*** (0.0487)	0.559*** (0.0400)	0.570*** (0.0376)
Award Level 4	0.458	0.486*** (0.0474)	0.414*** (0.0751)	0.424*** (0.0601)	0.495*** (0.0529)
Award Level 3	0.362	0.385*** (0.0519)	0.327*** (0.0794)	0.309*** (0.0630)	0.418*** (0.0594)
Award Level 2	0.308	0.298*** (0.0478)	0.337*** (0.0663)	0.282*** (0.0544)	0.337*** (0.0548)
Number of observations		4224	2240	3168	3296
Wald Chi^2 (11)		349.73	144.25	230.69	259.95
Prob > Chi2		0.0000	0.0000	0.0000	0.0000
Pseudo R^2		0.1887	0.1938	0.1781	0.1998

5.8 Discussion & Conclusions

This DCE is valuable as it, unlike most other studies, quantifies the relative importance of job attributes for rural health workers. This has the potential to inform upcoming policy decisions of the Government of Ghana for the pressing policy matter of ensuring equity in the distribution of health workers across the country. It is one of the few studies in Ghana that investigates the job preferences of community health extension workers, which are becoming an integral part of the Ghanaian health system through the expansion of the so-called CHPS zones to reach rural communities.

The results for the CHOs indicate that providing a functional and well-equipped CHPS compound where the CHOs can also sleep had the biggest impact on job choice. Higher salaries, better career opportunities and providing transport subsidies were also important to motivate rural health workers. However, increasing supportive supervision did not increase the CHOs' likelihood to prefer job.

For CHVs, it was shown that the most sought-after attributes of a job were also upgraded facilities, followed by offering training opportunities and the provision of a stipend. However, offering the smallest stipend amount (40 Cedis) had the smallest impact on the marginal probability to accept a posting.

This study corroborates previous findings that show that adequate clinical infrastructure as well as accommodation are of utmost importance to health workers (see also Zaidi, 1986; Vujicic, 2004; Kruk et al., 2010). It also buttresses findings by Snow et al. (2011) who found (in a qualitative study) that career development incentives and salary top-ups were important factors to accept postings in rural areas. A study with Ghanaian nurse-assistants, community health nurses, and other laborers in the health system (Alhassan et al., 2013) found that staff satisfaction was generally low, but non-financial incentives such as transportation to work, career development prospect, and clinical resource availability are also important sources for staff motivation.

The fact that both CHOs and CHVs were highly incentivized by better facilities and equipment – even over salary – is surprising but also encouraging. The results of this DCE study can be utilized to construct incentive packages that can help to motivate health professionals to work in remote areas and has the potential to increase retention rates. The results can also function as a justification to campaign for more investments in facilities, housing of health workers in remote areas, and improved equipment.[93] Our study also shows that even small and less costly tokens of gratitude can make substantial impacts, including offering small subsidies for transport, or non-

[93] The results of this study were presented to the Government of Ghana in a seminar in 2017.

financial incentives for CHVs, like installing a “CHV of the Month” award of offering a gift bag with some clothes and food, or conducting a few trainings. Finding effective non-financial incentives for health workers is crucial for the Government, as the health sector wage bill of the country is already relatively high when compared to neighboring countries (McCoy et al., 2008).

In the future, it would be interesting to cost out the different attributes to add more value to the willingness to pay estimates. If we know what alternatives to providing additional salary would cost the government, policy advice could be more precise and pertinent (similar to recommendations by Kolstad, 2011).

This research is limited by some common concerns with these types of studies, such as whether the subjects of the research have adequately understood the questions, and also whether social desirability bias played a role when answering the choice sets (placing higher value on non-financial attributes). The study is also limited by the fact that it only observes stated preferences, and not revealed preferences, as we do not observe actual behavior of health extension workers. While DCEs are considered to be a better tool to elicit preferences when compared to simple survey responses, they are not as good as observing actual choices. However, these can often not be randomized on a large scale. If possible, it would be an interesting avenue for future research to conduct a policy experiment where some of the attributes subject to this DCE are actually implemented and their impact on real choices tested.

This study only looked at one aspect of improving health care, specifically the retention and motivation of extension health workers. Obviously achieving universal health care consists of many more components, such as the quality of training, the quality of equipment, the demand for services, pricing of medication, the overall health wage bill, etc. This study should therefore not be analyzed in isolation, but rather as an additional piece to a much larger and more challenging puzzle.

5.9 References

Adzei, F. A., & Atinga, R. A. (2012). Motivation and retention of health workers in Ghana's district hospitals: addressing the critical issues. *Journal of health organization and management*, *26*(4), 467-485.

Alhassan, R. K., Spieker, N., van Ostenberg, P., Ogink, A., Nketiah-Amponsah, E., & de Wit, T. F. R. (2013). Association between health worker motivation and healthcare quality efforts in Ghana. *Human Resources for Health*, *11*(1), 37.

Antwi, J., & Phillips, D. C. (2013). Wages and health worker retention: Evidence from public sector wage reforms in Ghana. *Journal of Development Economics*, *102*, 101-115.

Appiah-Denkyira, E., Herbst, C. H., Soucat, A., Lemiere, C., & Saleh, K. (2013). Towards Interventions in Human Resources for Health in Ghana: Evidence for Health Workforce Planning and Results. World Bank Publications.

Araujo, E., & Maeda, A. (2013). How to recruit and retain health workers in rural and remote areas in developing countries: a guidance note.

Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., & Schuman, H. (1993). Report of the NOAA panel on contingent valuation. *Federal register*, *58*(10), 4601-4614.

Ashraf, N., Bandiera, O., & Jack, B. K. (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, *120*, 1-17.

Ashraf, N., Bandiera, O., & Lee, S. S. (2015). Do-gooders and go-getters: career incentives, selection, and performance in public service delivery. *Harvard Business School*.

Basinga, P., Gertler, P. J., Binagwaho, A., Soucat, A. L., Sturdy, J., & Vermeersch, C. M. (2011). Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *The Lancet*, *377*(9775), 1421-1428.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, *96*(5), 1652-1678.

Blaauw, D., Erasmus, E., Pagaiya, N., Tangcharoensathein, V., Mullei, K., Mudhune, S., Goodman, C., English, M., & Lagarde, M. (2010). Policy interventions that attract nurses to rural areas: a multicountry discrete choice experiment. Bulletin of the World Health Organization, 88(5), 350-356.

Buykx, P., Humphreys, J., Wakerman, J., & Pashen, D. (2010). Systematic review of effective retention incentives for health workers in rural and remote areas: Towards evidence-based policy. *Australian Journal of Rural Health*, *18*(3), 102-109.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, *19*(1-3), 7-42.

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics using stata* (Vol. 2). College Station, TX: Stata press.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006). Missing in action: teacher and health worker absence in developing countries. *Journal of Economic perspectives*, *20*(1), 91-116.

Clark, M. D., Determann, D., Petrou, S., Moro, D., & de Bekker-Grob, E. W. (2014). Discrete choice experiments in health economics: a review of the literature. *Pharmacoeconomics*, *32*(9), 883-902.

Dal Bó, E., Finan, F., & Rossi, M. A. (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. The Quarterly Journal of Economics, 128(3), 1169-1218.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, *18*(1), 105.

DellaVigna, S., & Pope, D. (2017). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies*, *85*(2), 1029-1069.

de Bekker-Grob, E. W., Ryan, M., & Gerard, K. (2012). Discrete choice experiments in health economics: a review of the literature. *Health economics*, *21*(2), 145-172.

Deserranno, E. (2017). Financial incentives as signals: Experimental evidence from the recruitment of village promoters in uganda.

De Walque, D., Gertler, P. J., Bautista-Arredondo, S., Kwan, A., Vermeersch, C., de Dieu Bizimana, J., ... & Condo, J. (2015). Using provider performance incentives to increase HIV testing and counseling services in Rwanda. *Journal of health economics*, *40*, 1-9.

Finan, F., Olken, B. A., & Pande, R. (2015). *The personnel economics of the state* (No. w21825). National Bureau of Economic Research.

Dunsch, F. A., Evans, D. K., Eze-Ajoku, E., & Macis, M. (2017). *Management, Supervision, and Health Care: A Field Experiment* (No. w23749). National Bureau of Economic Research.

Dussault, G., & Franceschini, M. C. (2006). Not enough there, too many here: understanding geographical imbalances in the distribution of the health workforce. *Human resources for health*, *4*(1), 12.

Garbers, Y., & Konradt, U. (2014). The effect of financial incentives on performance:: A quantitative review of individual and team‐based financial incentives. *Journal of occupational and organizational psychology*, *87*(1), 102-137.

GHS – Ghana Health Service (2016). National Community-Based Health Planning and Services (CHPS) Policy. Accelerating Attainment of Universal Health Coverage and Bridging the Access Inequity Gap.

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies*, *29*(1), 1-17.

Grobler, L., Marais, B. J., & Mabunda, S. (2015). Interventions for increasing the proportion of health professionals practising in rural and other underserved areas. Cochrane database of systematic reviews, (6).

Hart, J. T. (1971). The inverse care law. *The Lancet*, *297*(7696), 405-412.

Heerdegen, A. C. S., Bonenberger, M., Aikins, M., Schandorf, P., Akweongo, P., & Wyss, K. (2019). Health worker transfer processes within the public health sector in Ghana: a study of three districts in the Eastern Region. *Human Resources for Health*, *17*(1), 45.

Hensher, D. A., Rose, J. M., & Greene, W. H. (2005). Applied choice analysis: a primer. Cambridge University Press.

Hoefman, R. J., van Exel, J., Rose, J. M., Van De Wetering, E. J., & Brouwer, W. B. (2014). A discrete choice experiment to obtain a tariff for valuing informal care situations measured with the CarerQol instrument. Medical Decision Making, 34(1), 84-96.

Hole, A. R. (2018). Discrete Choice Methods in Health Economics. In *Health Econometrics* (pp. 85-99). Emerald Publishing Limited.

Honda, A., & Vio, F. (2015). Incentives for non-physician health professionals to work in the rural and remote areas of Mozambique—a discrete choice experiment for eliciting job preferences. Human resources for health, 13(1), 23.

Huicho, L., Miranda, J. J., Diez-Canseco, F., Lema, C., Lescano, A. G., Lagarde, M., & Blaauw, D. (2012). Job preferences of nurses and midwives for taking up a rural job in Peru: a discrete choice experiment. *PloS one*, *7*(12), e50315.

Kamenica, E. (2012). Behavioral economics and psychology of incentives. *Annu. Rev. Econ.*, *4*(1), 427-452.

Kolstad, J. R. (2011). How to make rural jobs more attractive to health workers. Findings from a discrete choice experiment in Tanzania. Health economics, 20(2), 196-211.

Kruk, M. E., Johnson, J. C., Gyakobo, M., Agyei-Baffour, P., Asabir, K., Kotha, S. R., ... & Dzodzomenyo, M. (2010). Rural practice preferences among medical students in Ghana: a discrete choice experiment. *Bulletin of the World Health Organization*, *88*, 333-341.

Kwansah, J., Dzodzomenyo, M., Mutumba, M., Asabir, K., Koomson, E., Gyakobo, M., ... & Snow, R. C. (2012). Policy talk: incentives for rural service among nurses in Ghana. *Health policy and planning*, *27*(8), 669-676.

Lacetera, N., Macis, M., & Slonim, R. (2013). Economic rewards to motivate blood donations. Science, 340(6135), 927-928.

Lancaster, K. J. (1966). A new approach to consumer theory. Journal of political economy, 74(2), 132-157.

Lancsar, E., Fiebig, D. G., & Hole, A. R. (2017). Discrete choice experiments: a guide to model specification, estimation and software. *PharmacoEconomics*, *35*(7), 697-716.

Lori, J. R., Rominski, S., Richardson, J., Agyei-Baffour, P., Kweku, N. E., & Gyakobo, M. (2012). Factors influencing Ghanaian midwifery students' willingness to work in rural areas: a computerized survey. *International journal of nursing studies*, *49*(7), 834-841.

Mathauer, I., & Imhoff, I. (2006). Health worker motivation in Africa: the role of non-financial incentives and human resource management tools. *Human Resources for Health*, *4*(1), 24.

McCoy, D., Bennett, S., Witter, S., Pond, B., Baker, B., Gow, J., ... & McPake, B. (2008). Salaries and incomes of health workers in sub-Saharan Africa. *The Lancet*, *371*(9613), 675-681.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. Frontiers in Econometrics. New York: Academic Press; 1974. p. 105–42.

McFadden, D. (1986). The choice theory approach to market research. Marketing science, 5(4), 275-297.

Mellström, C., & Johannesson, M. (2008). Crowding out in blood donation: was Titmuss right?. *Journal of the European Economic Association*, *6*(4), 845-863.

Olken, B. A., Onishi, J., & Wong, S. (2014). Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia. *American Economic Journal: Applied Economics*, *6*(4), 1-34.

Ryan, M., Gerard, K., & Amaya-Amaya, M. (Eds.). (2007). Using discrete choice experiments to value health and health care (Vol. 11). Springer Science & Business Media.

Ryan, M., Kolstad, J., Rockers, P., & Dolea, C. (2012). How to conduct a discrete choice experiment for health workforce recruitment and retention in remote and rural areas: a user guide with case studies. *World Health Organization & CapacityPlus: World Bank*.

Schlechter, A., Thompson, N. C., & Bussin, M. (2015). Attractiveness of non-financial rewards for prospective knowledge workers: An experimental investigation. Employee Relations, 37(3), 274-295.

Scott, A. (2002). Identifying and analysing dominant preferences in discrete choice experiments: an application in health care. Journal of economic Psychology, 23(3), 383-398.

Shiratori, S., Agyekum, E. O., Shibanuma, A., Oduro, A., Okawa, S., Enuameh, Y., ... & Ansah, E. (2016). Motivation and incentive preferences of community health officers in Ghana: an economic behavioral experiment approach. *Human resources for health*, *14*(1), 53.

Smitz, M. F., Witter, S., Lemiere, C., Eozenou, P. H. V., Lievens, T., Zaman, R. U., ... & Hou, X. (2016). Understanding health workers' job preferences to improve rural retention in Timor-Leste: findings from a discrete choice experiment. PloS one, 11(11), e0165940.

Snow, R. C., Asabir, K., Mutumba, M., Koomson, E., Gyan, K., Dzodzomenyo, M., ... & Kwansah, J. (2011). Key factors leading to reduced recruitment and retention of health professionals in remote areas of Ghana: a qualitative study and proposed policy solutions. *Human resources for health*, *9*(1), 13.

Song, K., Scott, A., Sivey, P., & Meng, Q. (2015). Improving Chinese primary care providers' recruitment and retention: a discrete choice experiment. Health policy and planning, 30(1), 68-77.

Stilwell, B., Diallo, K., Zurn, P., Vujicic, M., Adams, O., & Dal Poz, M. (2004). Migration of health-care workers from developing countries: strategic approaches to its management. *Bulletin of the World health Organization*, *82*(8), 595-600.

Sun, X., Liu, X., Sun, Q., Yip, W., Wagstaff, A., & Meng, Q. (2016). The Impact of a Pay-for-Performance Scheme on Prescription Quality in Rural China. *Health economics*, *25*(6), 706-722.

Takemura, T., Kielmann, K., & Blaauw, D. (2016). Job preferences among clinical officers in public sector facilities in rural Kenya: a discrete choice experiment. Human resources for health, 14(1), 1.

Velenyi, E. V. (2016). Health care spending and economic growth. In World Scientific Handbook of Global Health Economics and Public Policy: Volume 1: The Economics of Health and Health Systems (pp. 1-154).

Velenyi, E., Dunsch, F., Miller, A., Tjaden, J. (2016). Leveling the Playing Field. A Comparative Analysis of the Population Coverage, Equipment, Human Resources, and Services of Primary Health Facilities and CHPS Zones across 8 Districts in Ghana. World Bank, Washington D.C. Unpublished.

Vujicic, M., Zurn, P., Diallo, K., Adams, O., & Dal Poz, M. R. (2004). The role of wages in the migration of health care professionals from developing countries. *Human resources for Health*, *2*(1), 3.

Wang, H., Otoo, N., & Dsane-Selby, L. (2017). *Ghana National Health Insurance Scheme: Improving Financial Sustainability Based on Expenditure Review*. World Bank Publications.

Willis-Shattuck, M., Bidwell, P., Thomas, S., Wyness, L., Blaauw, D., & Ditlopo, P. (2008). Motivation and retention of health workers in developing countries: a systematic review. *BMC health services research*, *8*(1), 247.

World Health Organization. (2010). *Increasing access to health workers in remote and rural areas through improved retention: global policy recommendations*. World Health Organization.

World Health Organization. (2015). Health workforce and services: Draft global strategy on human resources for health: workforce 2030. Rep. Secr. Exec. Board EB13836 [Internet].

Zaidi, S. A. (1986). Why medical students will not practice in rural areas: evidence from a survey. *Social science & medicine*, *22*(5), 527-533.

5.10 Annex Tables

Table A5.1: Top attributes for CHOs mentioned in preparatory focus group discussions

Attribute	Description - Details
1. Career Progress and Performance Assessment	- Opportunity for promotion after 3 years if placed in remote CHPS zones - Opportunity to move from "Certificate" to "Diploma" status more easily based on experience as a CHO
2. Salary of CHOs	- Salary of Certificate-holding CHOs (CHNs) at the basic level should be increased to 1,000 GHC *(200 GHC less than Diploma holding Nurses)*
3. CHPS Compounds	- CHPS Compounds should ideally be built prior to posting CHOs in a CHPS zone - Compounds should have the basic equipment to provide community-level services, and social amenities to make living comfortable for posted CHOs - CHPS Compounds should be at most 50 meters from the community to enhance security - If CHOs do not have a Compound and are posted, they should be given a stipend of 50 GHC for accommodation
4. Equipment/Logistics	- Essential equipment which was lacking included: Fridge, BP Apparatus, Infant and Adult Weighing scales, Outreach bags and Delivery beds
5. Performance-based Incentive / Bonus Payment	- 100 GHC per CHO per month would be an interesting top-up for CHOs
6. Motor Bike and Fuel	- AG 100 motorbike - 2 Gallons of fuel per week (equivalent to about 30 GHC per week/ 120 GHC/month)
7. Sufficient Staffing	- CHOs feel that CHPS zones should have 2 CHNs, 1 Enrolled Nurse and a midwife
8. On-the-job Trainings	- CHOs should receive on the job trainings every quarter to stay updated on newest health policies - If a training is held at regional level, districts should bring sub-districts and CHPS zones up-to-speed by holding debrief training at lower level
9. Volunteer Motivation	- Active volunteers should receive 50 GHC/ month to keep them motivated as Volunteer turnover is high - Volunteers should receive free registration in the NHIS Note: *The above points on Volunteer Motivation represent the perspective of CHOs and CHNs regarding the CHV financial motivation. Within the main attributes for CHVs listed below, the attribute of financial motivation is also identified from the perspective of Volunteers themselves*.
10. Non-Financial Incentives	- Certificate of "Best CHO of the Month / Year" would instill pride in their work

Table A5.2: Top attributes for CHVs mentioned in preparatory focus group discussions

Attribute	Description - Details
1. Financial Motivation/Stipend	- If active Volunteers received a monthly stipend of 100 GHC/ month, this would be an interesting financial motivation
2. Equipment and Supplies	- Key equipment mentioned for CHVs included outreach box (with essential medications such as antimalarials and a first aid kit if CHVs are appropriately trained), weighing scale, rain coat, rain boots, torch light
3. Transportation	- Bicycles for home visits - A transportation allowance of 20 GHC/ month for Volunteers without a Bicycle.
4. Work Opportunities in Health Facilities	- CHVs should be hired in health facilities as cleaners, messengers or security guards
5. Opportunities for Training	- CHVs wanted training in preventive care and first aid
6. Performance based Financial Incentives	- 30 GHC/ month per CHV would be an interesting top-up motivation for CHVs
7. Non-financial Awards/“Tokens” Of Appreciation	- Certificates - T-shirts - Bicycle - Free NHIS Registration

6. Experimentation in the Social Sciences & the Problem of External Validity

6.1 Abstract

The quantity of impact evaluations (IEs) in the field of development economics utilizing experimental or quasi-experimental methods has picked up considerably over the past 2 decades. This is partly because impact evaluations are being welcomed by governments and researchers as a tool to evaluate the effectiveness of their policies and to provide policy guidance for future decisions – by showing “what works”. However, IEs inherently struggle with "external validity", i.e. the generalizability of results to other contexts. This chapter argues that the problem of context variance (as an extension of the problem of induction) is an overlooked aspect for lack of external validity of IE results as this age-old principle has been neglected by economists and policymakers alike. No matter how many instances of a causal effect are observed, although implicitly claimed otherwise by many IE supporters, it is not possible to claim with certainty that an identified cause-and-effect relationship holds over time, or across contexts. However, IEs are still a valuable tool for research and policy. Throughout history, trial-and-error has been the premier method of social and economic development, almost exclusively in non-scientific fashion. While IEs can never reliably verify whether a policy is worth implementing, falsifications of existing theories can be informative and foster the creation of new ones, so that the trial-and-error process is accelerated, and better working policies can be identified faster.

"I know that a statement is wrong, but not necessarily what statement is correct. If I see a black swan, I can certify that *all swans are not white! (…)* We can get closer to the truth by negative instances, not by verification" (Taleb, 2007, p. 56)

6.2 RCTs – Proliferation and the "What Works" Agenda

This essay discusses the problem of external validity for RCTs in the field of development (economics).[94] It provides an overview of the rise of the use of RCTs and the "what works" agenda (section 6.2), then discusses how the crucial problem of context variance virtually renders external validity impossible by contrasting opposing viewpoints from leading academics on the issue (section 6.3), and ends with recommendations on how to improve the utility of RCTs in the absence of external validity (section 6.4).

Cook et al. (2002, p. 34) define validity as follows:

> We use the term validity to refer to the approximate truth of an inference. When we say something is valid, we make a judgment about the extent to which relevant evidence supports that inference as being true or correct. (…) Assessing validity always entails fallible human judgments. We can never be certain that all of the many inferences drawn from a single experiment are true or even that other inferences have been conclusively falsified. That is why validity judgments are not absolute; various degrees of validity can be invoked.

Campbell (1957; cited in Cook et al., 2002, p. 37) first defined internal validity as: "did in fact the experimental stimulus make some significant difference in this specific instance?" Cook et al. (2002, p. 83) then state that "external validity concerns inferences about the extent to which a causal relationship holds over variation in persons, settings, treatments, and outcomes." An experiment is *not* externally valid, if the results of the study cannot be generalized to a larger population, a different context, or in the same context over time. This article argues that RCTs and other forms of quasi-experimental impact evaluations (IE) cannot fully fulfill this premise of external validity partly due to the often-overlooked phenomenon of "context variance" related to the problem of induction, which is well known in philosophy, and increasingly so in social psychology, but has received relatively little attention in the development economics debate around the usefulness of RCTs for policymaking.

Ronald Fisher (1925) is often credited with pioneering randomized controlled trials (see e.g. Jamison, 2017 for a history of RCTs; Banerjee et al., 2016). However, the use experimentation really "took off" in the field of development economics around the year 2000. The quantity of experiments has spiked considerably over the last 10 to 15 years (See e.g. Figure 1.2 in the introductory chapter; Banerjee et al., 2016). The basic premise of RCTs is simple and powerful: Random assignment is meant to balance out potential confounding effects, so that the average

[94] Other issues on the viability of RCTs were discussed in the introductory chapter 1.

treatment effect of a project can be identified (see chapter 1). Unlike for other methods, no complicated assumptions about covariates and no models are required (Deaton and Cartwright, 2018).

The most prominent RCT proponents Esther Duflo, Abhihit Banerjee, and Michael Kremer just won the Nobel Prize in Economics: "This year's Laureates have introduced a new approach to obtaining reliable answers about the best ways to fight global poverty" (The Prize in Economic Sciences, 2019). Duflo and Kremer (2005, p. 205) are convinced that "credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and non-governmental organizations (NGOs) beyond national borders." Cook et al. (2002, p. 276) state that "randomized experiments provide a *precise* answer about whether a treatment worked". Referring to the use of impact evaluation methods, a Lancet editorial titled in 2004: "The World Bank is finally embracing science", after being accused by Banerjee (2007; cited in Deaton, 2010, p. 438) of "lazy thinking", and a "resistance to knowledge" by not adopting RCTs earlier. Banerjee (2007) classifies experimental results as "hard" evidence which is based on "evidence" rather than "trends" (Duflo, 2005). The implicit argument here is that other – non-experimental – studies are less "internally valid" enough and RCTs, through this revolution, would produce improved results with higher precision, which could then be used to propose better policies. Angrist and Pischke (2010, p. 4) are convinced that "[e]mpirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact", and Duflo and Kremer think that RCTs can "revolutionize social policy" in the 21st century just as it did with medicine in the 20th (Duflo and Kremer, 2005, p. 32). Pulitzer-Prize winning New York Times columnist Nicholas Kristof even called RCTs "the hottest thing in the fight against poverty" (2011). Considering this development, David McKenzie asks whether "RCTs have taken over development economics?" (McKenzie, 2016) and Banerjee and Duflo spoke of an "explosion" of the method (2009). When interviewed for The New Yorker magazine, Esther Duflo, one of the leading figures in the field, states that randomization "takes the guesswork, the wizardry (...) out of whether something works or not". Without rigorous impact evaluations we are no better than "medieval doctors and their leeches", she is quoted by Parker (2010). Banerjee (2007) affirms that the best argument for experiments is that it can spur innovation by making it easy to see "what works". To summarize, proponents claim that RCTs are the "gold standard" to establish causality, as they are deemed to be better, more credible, or more rigorous than other empirical methods that do not rely on randomization for causal attribution of impacts (Deaton and Cartwright, 2018) and these results provide solid guidance on which policies "work".[95] What is meant is that they produce the most *internally valid* estimates of causal effects or unbiasedness of point estimates[96] and that these results can be transferred to the future or other context (guiding policy).

The witty RCT critic Lant Pritchett sees three main reasons for the recent rise in RCTs in development economics (in Ogden, 2017):

1. There were debates around "identification" in economics for a longer time, which reached a boiling point. More and more studies were criticized for not being well-identified and so the trend towards RCTs was accelerated to take the issue of the table.[97]
2. Major bilateral and multilateral donors were increasingly pressured to show more and better impacts and impact evaluations/RCTs seemed like a viable way to provide "proof" of impacts and subsequently improve policies.
3. Pritchett sees a rise in philanthropic enterprise[98], which is intrinsically more interested in the smaller, more tangible issues and development problems, contrary to behemoths like the World Bank that attempt to change government policies and institutions.[99]

Indeed, more and more organizations and governments use randomized controlled trials (RCTs) as a tool for evaluation to find out and prove which policies "work", partly as a response to the rather disappointing results of decades of development aid and cooperation. With public pressure mounting to deliver results (see also chapter 1), for many organizations it is no longer acceptable to just assess the subjective satisfaction of officials or beneficiaries to claim that a project was a success (Woolcock, 2013). There are now a series of institutions entirely focused on the use of experimental and quasi-experimental methods as a means to provide actionable policy advice. The World Bank has multiple units charged with impact evaluations (e.g. the Development Impact Evaluation Unit (DIME), SIEF, and the Gender Lab), "The Abdul Latif Jameel Poverty Action Lab" (J-PAL) and "Innovations for Poverty Action" (IPA) conduct RCTs around the world. On the J-PAL website it reads that "randomized evaluations are generally considered the most rigorous and, all else equal, produce the most accurate (i.e. unbiased) results" (quoted in Deaton and Cartwright,

[95] As correctly pointed out by Woolcock (2013), the actual "gold standard" in clinical trials is that the allocation is "triple blind", i.e. neither the subjects, nor the researchers and supervisors are aware of who is part of the treatment and who is part of the control until the study has been concluded. This is obviously almost always impossible for policy-relevant trials.

[96] This claim is also often criticized, but it's not the focus of this article see chapter 1 of this dissertation or Barrett and Carter, 2014, for a summary.

[97] Angus Deaton makes a similar point in his interview in Ogden (2017) in the same edited volume.

[98] De Souza Leão and Eyal (2019, p. 388) call this group "philanthro-capitalists".

[99] Pritchett states: "From the charity perspective, there's a nice confluence between the methodological demand for statistical power and of being able to tweak at the individual level" (Ogden, 2017, p. 142).

2016, p. 6). 3ie is an organization entirely dedicated to the promotion and funding of RCTs (funded by DFID), universities have built units that centered their work around experimental policy evaluations (e.g. Harvard's EPod, or Berkeley's CEGA). There are conferences illustrating "what works" like the "What Works Global Summit" (WWGS) held in London in 2017 and also "centers" such as the US Department of Education's What Works Clearing House, The Campbell Collaboration (parallel to the Cochrane Collaboration in health), the Scottish Intercollegiate Guidelines Network (SIGN), the US Department of Health and Human Services Child Welfare Information Gateway, the US Social and Behavioral Sciences Team, What Works Centers established by the UK Government (Deaton and Cartwright, 2018).

Separate from "internal validity", a more formal way to discuss "what works", is the concept of *external validity*. Bracht and Glass (1968, p. 438) define external validity in the following way: "To the extent and manner in which the results of an experiment can be generalized to different subjects, settings, experimenters, and, possibly, tests, the experiment possesses external validity." Proponents of RCTs stating that these can show us "what works" going forward are essentially arguing for the claim that RCTs, when done right and establishing solid internal validity first, are or can also be externally valid. However, others disagree on this point and criticize RCTs for a potential *lack* of "external validity", i.e. being inapt to generalize results from one experiment across contexts or even the same context over time. There have been lively discussions around the question how external validity is deficient, and how it can be improved, which is also the focus of this chapter.[100]

Nuanced proponents of RCTs do agree that there are challenges surrounding external validity, but Banerjee et al. (2017a, p. 74), for example, are convinced that "it is far from unattainable."[101] Lant Pritchett is one of the most prominent critics of RCTs and specifically their use as a tool for policy advice. He sums up his general critique in an interview with the economist Timothy Ogden in 3 main points, of which external validity is one (Ogden, 2017):

1. Small development problems: RCTs can only tackle small problems (treatment units are often individuals, schools, health centers) because by their nature they cannot tackle large, systemic questions (because countries cannot be randomized into different interest rate regimes, for example). He states that "nothing super important about development happens at the individual level." RCTs therefore have by design only limited value when

[100] The concept of "validity" was first established by Campbell and Stanley in 1966 (Campbell and Stanley, 2015).

[101] There are many other issues or problems that arise when implementing RCTs that are not mentioned in this paper. See for example: Barrett and Carter, 2014.

it comes to answering the very large questions about growth, poverty, and other proxies for development (Ogden, 2017, calls this the "Trivial Significance Critique" of RCTs).

2. Political economy: Secondly, Pritchett believes that the "randomistas" misunderstand that results of RCTs do not automatically translate to policy change. "They have this unbelievably Cro-Magnon simple model of policy adoption that essentially asserts that once there is knowledge in the world about policies, that will lead to better policies being adopted and implemented" (Ogden, p. 136). If countries have strategic interests, effectiveness might not be most important factor for decision-making (Prowse, 2007).
3. External validity: Since there is huge heterogeneity in program impacts for "non-rigorous" studies and RCTs, external validity cannot be achieved. Pritchett and Sandefur (2014; 2015) for example show that non-experimental studies from the same context (country) can be more precise and have more external validity than RCTs from a different context.

Points 1 and 2 were briefly touched on in the introductory chapter 1. This chapter will mainly address this third point, RCTs' inherent lack of external validity.

6.3 The "Invariance Law" and the Problem of Induction

"I think economists, especially development economists, are sort of like economists in the 1950s with regressions. They have a magic tool, but they don't yet have much of an idea of the problems with that magic tool." (Deaton; in Ogden, 2017, p. 38)

Banerjee and Duflo (2017) themselves define the problem related to "external validity" in their "Handbook of Economic Field Experiments" as follows:

> The formal way of thinking about this problem is to recognize that though the random assignment ensures that unobservables are distributed identically across treatment and control groups and that the treatment is not correlated with these unobservables, the estimated program effects are for not for the treatment alone, but rather for the treatment interacted with the unobservable characteristics in the study sample. If these unobservable characteristics vary between the study sample and the universe to which we seek to extrapolate the findings to, then the estimated treatment effects may not be valid because the interactions may change. (Banerjee and Duflo, 2017, p. 353)

The crucial, and often overlooked, aspect is just how much weight we give these context *unobservables*. This is a conundrum, as we do not know what and how important these unobservable characteristics are (as they are unobservable). In other words: Results are only generalizable, if the context stays stable.[102] The more "technical" problems with extrapolations from RCTs that are laid out in section 1.3 are valid but miss the more profound point that the instability or variance (and immeasurability) of social contexts over space and time makes it hardly possible to exert any claims on external validity from one or a series of experiments. This point is often not in the focus, even in articles that summarize problems of external validity, as they often look at more technical aspects (e.g. in Peters et al., 2018). It is usually overlooked that even if all the precautions are taken ex-ante and the hazards appropriately addressed or discussed, this does not mean that one comes closer to external validity, as we have no test to check whether the new context (across space and/or time) to which a previously tested program or policy should be applied is anything like the context in which the original RCT, from which results are being extrapolated, was conducted.

The "what works" approach is rooted and can only work in line with a positivist worldview, which assumes that there are universal laws that can be uncovered through repeated related experiments, evoking LaPlace's classic all-intelligent future-predicting "demon".[103]

> Positivism seeks to understand the social world by uncovering universal 'laws' through the measurement of the 'constant conjunction of events' between two or more phenomena. These 'laws' are empirical generalisations which are seen to be (mainly) independent of time/space and are neutral and value-free" (Steinmetz, 1998; cited in Prowse, 2007, p. 2).[104]

To illustrate this problem in regard to external validity, Pritchett mentions (in his interview with Ogden, 2017) for example that even very simple games, such as the ultimatum game, the results differ widely across contexts. If even these simple games show no external validity, how can

[102] Deaton and Cartwright (2018) also mention the important fact that *one* randomization likely does not balance out the treatment and control group. While the estimator in one trial would be unbiased (in the ideal case), it might not be the truth: The result might not be the effect of the treatment, but rather stemming from an observable or unobservable covariate which randomization did not balance out. They even argue for re-randomization in case some key (observable) covariates are unbalanced. For this, some theories around the impact mechanisms is required, which will be discussed further down in this paper.

[103] "We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it – an intelligence sufficiently vast to submit these data to analysis-it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes" (LaPlace, 1902, chapter II).

[104] The fact that humans do not always act rationally has entered the economic mainstream (Elster, 2000). Daniel Kahneman (2002) won the Nobel Prize based on experiments proving this. For example, despite low costs, parents in developing countries often do not vaccinate their children against deadly diseases, they use polluted water, even if clean sources are available, and they often do not save enough money. At their core, RCTs try to uncover behavioral patterns that hold over time. However, if there is no contingency of rational behavioral patterns, then the search even for basic for mechanisms maybe be futile.

program impacts?[105] In order to make his point, Pritchett refers to the "invariance law", a concept stemming from physics, that states that the laws of physics apply in the same way across space and time (for example gravitation), which is why the field of physics suffers less from context variance, experiments can easily be repeated and results extrapolated. Some argue that this is the reason why advances were faster in chemical and mechanical processes as there is less context heterogeneity in the physical world than there are in social sciences (Deaton, 2010).

Proponents of RCTs as a tool to directly inform policies implicitly believe that human behavior and their relationships *is* indeed also invariant over time, Moreover, in order to claim generalizability, in addition to the people's make-up being invariant over the type, in order to make claims of "what works", we would also need to assume that the contaxt is 1) also invariant, or 2) that context doesn't matter, which is unlikely to be the case:

> What we do have says that people and systems are not invariant. So, for instance, no one ever expected that the impact on a person of having the offer of microcredit at a given interest rate would be constant in some way. That's just absurd. The fundamentals of how to do science somehow got lost in the enthusiasm for RCTs. (Pritchett in Ogden, 2017, p. 139)[106]

Context variance can take many forms. Alcott and Mullainathan (2012) for example conclude that randomized experiments can suffer from "partner selection bias" (in the framework of an energy conservation project in partnership with microfinance institutions), that unobservables at the population level or in the economic environment are often overlooked, and that it would be valuable to describe the context when thinking about external validity and how it can differ from the sample in an RCT. Weighing in on the importance of contexts, Deaton and Cartwright (2018) state:

> For example, a trial that relies on providing incentives for personal promotion is of no use in a state in which a political system locks people into their social and economic positions. Conditional cash transfers cannot improve child health in the absence of functioning clinics. Policies targeted at men may not work for women. We use a lever to toast our bread, but levers only operate to toast bread in a toaster; we cannot brown toast by pressing an

[105] Woolcock (2013) also illustrates this point by referencing an experiment by Chong et al. (2014): The researchers sent 10 letters to deliberately non-existent addresses in 159 countries to see how many letters would come back (as mandated by ratified international conventions) and how long it would take. In countries of the bottom half of the world's education distribution, the average return rate was only 21 percent of the letters. This is to illustrate that even seemingly simple processes vary widely depending on context.

[106] Kaushik Basu echoes this problem of variance of contexts across time: "If we are fussy about proper randomization for our study and take the view that we should not accept the wisdom of samples drawn in a biased manner or from the wrong population, we should also take the view that we cannot say anything about the future" (Basu, 2005, p. 4337).

> accelerator, even if the principle of the lever is the same in both a toaster and a car. If we misunderstand the setting, if we do not understand why the treatment in our RCT works, we run the same risks as Russell's chicken.[107] (Deaton and Cartwright, 2018, p. 12)

Social programs are embedded in complex contexts with a myriad of factors that are unknown, which also interact with each other in ways we cannot begin to understand (see e.g. Cesario, 2014). In the field of social psychology context variance is described with term "auxiliary assumptions" (Earp and Trafimow, 2015).

The issue of context variance is linked to the problem of induction which stipulates that rules cannot be drawn from an incomplete collection of observations, as summarized by the famous quote by John Stuart Mill: "No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion" (cited in Taleb, 2005, p. 117). As phrased by John Stuart Mill (1884): "The uniform experience, therefore, of the inhabitants of the known world, agreeing in a common result, without one known instance of deviation from that result, is not always sufficient to establish a general conclusion." Apparent reoccurring causal phenomena observed in the past do not provide one with certainty to predict the future as no two circumstances are identical. Especially in the social sciences the human mind ultimately is the judge as to what context qualifies as identical (or similar enough). The most extreme form of this thinking is that science in fact cannot generate proven knowledge at all (see e.g. Lakatos and Musgrave, 1970).

While the problem of context variance and by extension the problem of induction indeed looms large for experimenters in social science, it is often neglected or overlooked, and in its current application in development economics hardly ever mentioned. Proponents of RCTs routinely gloss over the importance of the context in which their study showed impacts. Surprisingly, RCT proponent Banerjee does partly accept the argument that RCT results are not generalizable over space and time in a relatively early publication (2005), following Hume, admitting that experiments lack the rational basis for induction, which however did not alter the trajectory of the field's claims.

The use of experiments started in the natural sciences and it subsequently became the paradigm of advancing knowledge in medicine.[108] Experiments in pharmacology are rightly credited as a vehicle of one of the greatest achievements of human history, the arrival of modern medicine. So

[107] Meant here is Bertrand Russell's "chicken": "The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken" (Russell, 2001, p. 29).

[108] See Jamison (2017) for an overview of how experimentation spread from the medical field to social science.

why is it appropriate to use induction in the medical field but not as much in the social sciences?[109] The main difference here is the invariance law mentioned earlier. The methods are the same, but the research "objects" are different. Pharmacology studies the reactions of the human body, which is a relatively fixed entity. Human bodies react similarly to drugs all over the world, it is essentially the observation of a chemical reaction in a closed environment. In fact, human bodies are 99.5% to 99.9% genetically identical, so that we can be sure that the “context” (which is in case of the medical field, the human body) is the same across time and space. (But even in the medical field, there are animated discussions around external validity of results, see e.g. Rothwell, 2005.)

The difference between the *body* and the *behavior* of a human being lies in the sheer complexity – or, as some would argue, randomness – of potential reactions to a "treatment" which can be in the form of social programs, for example, conditional cash transfers, pay-for-performance schemes, or non-financial incentives to elicit certain behaviors.

> RCT studies focus on generating consistent and unbiased estimates of treatment effects of development interventions. In the biophysical sciences from which the RCT tradition arises, this often works because basic physio-chemical laws ensure a certain degree of homogeneity of response to an experiment. (…) [T]here is such heterogeneity of microenvironments that one has to be very careful about model mis-specification. (Barrett and Carter, 2014, p. 75)

Even under the assumption that the human mind would work according to the strict laws of physics (in yet unknown ways), the reaction crucially depends on the behavior of all the other people. However, group tastes change. What was perceived as good in 1992 is not as good anymore, if not bad, in 2012. Thus, even when we perceive the "matter" that is subject to our experiments as "fixed" and “map-able” we can still not be sure that the interactions of matter are similar in all circumstances.[110] Thus, this problem is aggravated in the society, where neither the entities *nor* the relationships between them can be perceived as stable in their make-up. Basu (2005) enforces this point:

[109] There is of course also an exhaustive discussion in the medical field about the pros and cons of RCTs. Even in physics there are issues around replicability (Tsang and Kwan, 1999).

[110] Rosalind Eyben is one of the few scholars pushing for a more relational understanding of development for quite some time (Eyben, 2012; 2010). It would, for example, be possible that a project improves the political culture or the trust of citizen among each other, but these changes cannot be measured in numbers. This would signify a success, even if growth, education or health statistics do not change. If, for any reason, quantitative methods dominate the evaluation field, this might lead to a situation where functioning projects are negatively reviewed or not even implemented because they cannot be evaluated with metrics.

> One may try to counter this by arguing that between yesterday and tomorrow there is no fundamental difference and so no reason to expect a relation that was true yesterday to be not true tomorrow. But the difference between yesterday and tomorrow is not just a matter of time. Between yesterday and tomorrow there can be war and pestilence; between yesterday and tomorrow can be 9/11, altering the way world politics is conducted; between yesterday and tomorrow we can have a warmer globe.[111] (Basu 2005, p. 4337)

Thus, the key question is whether social systems (like societies) work more than a clock (highly predictable – we can find out "what works") or like a cloud (highly unpredictable). Tiny changes of movements (or tiny butterfly-effect errors in data collection for example) of particles can lead to vastly different outcomes in the clouds that are formed (Tetlock and Gardner, 2015). Nate Silver (2012) reminds us that while chaos theory doesn't imply randomness, it does mean that complex systems are notoriously hard (or more likely: impossible) to predict. He tells the story of a weather forecasting computer that seemingly erratically changed the forecast although the data that it was fed was the same from one round to another. The team later found that the barometric pressure in one cell was entered as 29.517 instead of 29.5168. It has been truncated by a technician. This rounding of the decimal resulted in vastly different predictions.

Despite the problem of induction, using information from the past to make inferences about the future is not entirely wrong. We make predictions – foreseeing the future using data from the past – from inductive reasoning all the time, both in physics, the social sciences, and everyday life. "Without the influence of custom, we would be entirely ignorant of every matter of fact beyond what is immediately present to the memory and senses" (Hume, 2000, p. 21-2). The question then is not whether or not to use information from the past to make the decisions in the future. The question is rather the degree of certainty that can be claimed and to favor *probabilities* instead of proclaiming *certainty* ("what works"), to avoid, as formulated by John Maynard Keynes (1939; cited in Mookherjee, 2005, p. 4328) the "slippery problem of passing from statistical description to inductive generalisation."[112]

As mentioned, the goal of the proponents of the positivist "what works" agenda is to find the underlying structure or "laws" of what makes the world work. Once uncovered, the right policies

[111] This is akin to Hume: "For all inferences from experience suppose, as their foundation, that the future will resemble the past, and that similar powers will be conjoined with similar sensible qualities. If there be any suspicion that the course of nature may change, and that the past may be no rule for the future, all experience becomes useless, and can give rise to no inference or conclusion. It is impossible, therefore, that any arguments from experience can prove this resemblance of the past to the future, since all these arguments are founded on the supposition of that resemblance. Let the course of things be allowed hitherto ever so regular; that alone, without some new argument or inference, proves not that, for the future, it will continue so" (Hume, 2000).

[112] "If we be, therefore, engaged by arguments to put trust in past experience, and make it the standard of our future judgment, these arguments must be probably only (…)" (Hume, 2000).

can be applied. Many admit that one experiment alone does not suffice to uncover these patterns. Thus, *replication* of studies in different contexts is often offered as a solution to increase external validity (see e.g. Athey and Imbens, 2017, Duflo and Kremer, 2008; Cohen and Easterly, 2009; Angrist and Pischke, 2010).

As a result, donors like "3ie" have started to fund replications of already completed RCTs. If one experiment alone cannot reveal a pattern, the argument goes, multiple studies of the same kind should be able to do so. Banerjee (2005) sees as the only solution to build trust in experimental results as tools for extrapolation to replicate them in multiple different locations. There have indeed been some efforts to coordinate RCTs aiming to produce evidence on the same topic, probably most prominently displayed by the "Graduation from Ultra Poverty" experiments in 6 countries (Banerjee et al., 2015) or Meager's (2015) effort to analyze data from seven RCTs on microfinance. And: "Context dependence can be assessed by replications, either of the same experiments or of related experiments (that is, by experiments that test programs inspired by the same general idea)" (Banerjee et al., 2017a, p. 96). What the authors here mean by "the same general idea" – which would be central to his argument as it relates to the problem of context variance – remains unclear.[113] Unknowingly, proponents of the "what works" agenda navigate themselves into a paradox, as Deaton (2010) points out:

> [R]unning RCTs to find out whether a project works is often defended on the grounds that the experimental project is like the policy that it might support. But the 'like' is typically argued by an appeal to similar circumstances or a similar environment, arguments that depend entirely on observable variables. Yet controlling for observables is the key to the matching estimators that are one of the main competitors of RCTs and that are typically rejected by the advocates of RCTs on the grounds that RCTs control, not only for things we can observe but for things we cannot. (Deaton, 2010, p. 450)

This point is key and worth repeating: RCTs make claims about which projects work based on the assumptions that the context in which the experiment is conducted (often a pilot) is similar to the context in which the policy is implemented. However, in order to be more confident about whether or not the context is similar, studies would need to include more covariates. Then again, one of the main advantages of RCTs is that they do *not* depend on co-variates to make causal claims. There is essentially a trade-off between internal and external validity (Cartwright, 2007). Pursuing maximum internal validity "puts severe constraints on the assumptions a target popula-

[113] One additional problem that makes it difficult to repeat similar experiments in different regions is the fact that exact repetitions of already devised experiment designs are rarely an attractive option for scientists (Cohen and Easterly 2009; Angrist and Pischke 2010).

tion must meet to justify exporting a conclusion from the test population to the target" (Cartwright, 2007, p. 12). Barrett and Carter (2014) argue the same point:

> If we cannot count, model, and potentially measure factors that might be spuriously or otherwise correlated with key variables in observational data, then we can similarly never know if a universe of unknowable factors mediates the effects of even randomly distributed treatments. Well-identified local average treatment effects become data-weighted averages of multiple response regimes with unknowable dimensionality. Generalization to other populations, where the relative preponderance of the regimes may be different, becomes indefensible. Radical skepticism thereby destroys in equal measure the internal validity of observational studies and the external validity of RCTs. One cannot invoke one without unleashing the other. (Barrett and Carter, 2014, p. 59)

The belief in the power of replication is what is the underlying thought for many proponents of meta studies or systematic reviews of evidence. They try to advance a field by clustering studies of one topic together to identify common trends and create a "warehouses of verified instruments" with little regard to the context in which these individual studies were conducted (Woolcock, 2013). In the field of development, there are many systematic reviews on different topics, for example in the fields of education (Evans and Yuan, 2017; Evans and Popova, 2015), cash transfers on schooling outcomes (Baird et al., 2014), business training programs (McKenzie and Woodruff, 2013), or water quality interventions (Clasen et al., 2015). In order to aggregate studies for a meta-study or a systematic review, authors have to make assumptions on which studies "belong together", and should therefore be grouped (while assuming contexts are invariant). Unfortunately, there are no statistical procedures that can help with this and the grouping purely relies on assumptions made by researchers (humans!). Even if the observable context variables are quantified, generalizability might depend highly on unobservable context characteristics. Whenever we group studies, we implicitly either state that the contexts are similar, or that they do not matter, which is in stark contrast to, for example, Pritchett and Sandefur (2014, p. 193) who call the "design space" of any intervention as "hyper-dimensional" and even Hume (2000) who warned that new context "(…) may be only in appearance similar" (Hume, 2000). Deaton and Cartwright criticize the Cochrane Review, an institution dedicated to systematic reviews, which "(…) seems to suppose that there is a single effect to be uncovered that, once established, is implied by internal validity" (Deaton and Cartwright, 2016, p. 52). Pritchett and Sandefur (2014) also express concern with meta-studies for the exact reason of context variance:

> We are wary of the trend toward meta-analyses or 'systematic reviews' in development, as currently sponsored by organizations like DFID and 3ie. In many cases, the

> transplantation of meta-analysis techniques from medicine and the natural sciences presupposes the existence of a single set of universal underlying parameters, subject to the same type of conditional invariance laws (…).[114] (Pritchett and Sandefur, 2014, p. 163)

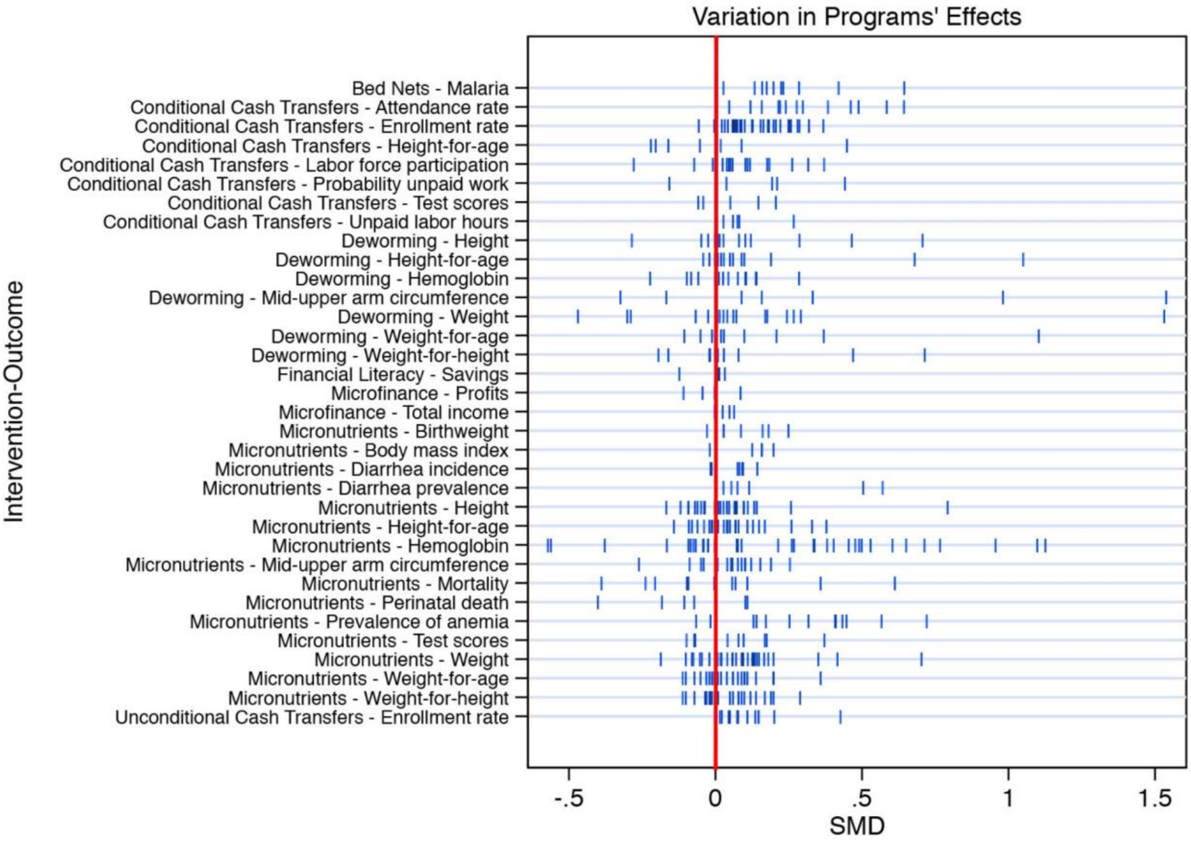
Maybe as a result of the overlooked problem of context variance, replications sometimes produce interesting if not outright funny results. Bold et al. (2013) find that the strong effects shown for a program of hiring contract teachers by an NGO in Kenya loses its effects when scaled up by the government country wide. This is also sometimes called the "piloting bias". The first RCT (the "pilot") is often run with a large amount of supervision by the research team which at a minimum ensures that there is no breakdown in the research protocol (see e.g. Banerjee et al., 2017a, which also includes more examples). Similarly, Vivalt (2015) compares RCTs run by the government with others run by NGOs or researchers and finds that the government RCTs usually find impacts that are on average smaller.

Another prominent recent example of where the "what works" agenda ostensibly failed is a prominent study by Miguel and Kremer (2004) in Kenya that showed that providing deworming medication to children in Kenya had a positive effect on test scores and school attendance in the treatment group and they also show positive spillovers to the control group, consequently recommending deworming as a cost-effective solution to improve educational outcomes. The policy was scaled up and by 2015, 200 million children were affected. After the data was made available online 2014, other researchers tried to replicate the findings, but could not validate the main outcome (de Souza Leão and Eyal, 2019), which resulted in uproar (the so-called "worm wars") and many following articles and blog posts in the academic community (see e.g. Majid et al., 2019).

In one of the largest meta studies on results of RCTs, Vivalt (2019) indeed finds that the reported results for seemingly similar programs are more heterogeneous than for example in the medical field (see Figure 6.1). Wilblin and Harris (2018) state in their podcast with Vivalt that the typical study result differs from the average effect found in similar studies so far by almost 100%.

[114] Pritchett and Sandefur (2015, p. 474): "Our results suggest that as policymakers draw lessons from experimental impact evaluations, they would do well to focus attention on heterogeneity in program design, context, and impacts, and may learn little from meta-analyses or 'systematic reviews' that focus exclusively on rigorous estimates of average effects for broad classes of interventions."

Figure 6.1: Variation in Programs' Effects (Vivalt, 2019)



Thus, so far, the "explosion" of academic articles brought about by the prominent method (Banerjee and Duflo, 2009) does not yet paint a coherent picture. Earlier in her career, Esther Duflo herself states the following (2006):

> In the absence of a well-funded alternative frame of analysis, these rejections appear now as a collection of random results that do not fit very well within any existing theory, and that we don't necessarily fully understand. This makes it difficult to generalize results and give them meaning, as some of the critics of randomized evaluation have pointed out. However, criticizing the experiments on this ground, like many have done, is a little bit like shooting the messenger. One may instead want to accept the message that they deliver: that we need to work on a new theoretical framework that can accommodate these results and predict new ones. (Duflo, 2006, p. 25)

And a bit later:

> We need a framework for interpreting what we find. For example, can we go beyond the observation that different inputs into the educational production function have different productivities? Is there any way to group the different inputs into broader input categories on a priori grounds, with the presumption that there should be less variation within

> the category? Or, on the outcome side, can we predict which outcomes of the educational system should commove more closely than the rest? Or is every experimental result sui generis?" (Banerjee and Duflo, 2009, p. 174)

In light of these results, one could assume that RCTs in the development sphere might very well suffer from the same "replication crisis" (Loken and Gelman, 2017; Maxwell et al., 2015) more commonly known in the field of psychology. Many experiments in the same field on the same issue fail to reproduce the same effects because they just might be "sui generis" as feared by Banerjee and Duflo. One reason why we maybe do not speak about a replication crisis in development economics (yet) is that there are just fewer replications of field experiments as they are harder to do in the field than in the lab (and replications are also less prestigious when compared to original research). Repeating an experiment in a different context requires as much work as a "new" experiment but generates significantly less scientific prestige (Earp and Trafimow, 2015).

The "context variance" problem and the empirical findings from meta-studies show that replication in a highly complex environment is probably not the solution, and seemingly buttressing Francis Bacon's claim that "(…) induction that proceeds by simple enumerations is childish" (Bacon, 1859; cited in Deaton and Cartwright 2018, p. 11). In fact, replication would only be useful if we can be sure that the contexts are identical or nearly identical in which the trials take place. Kaushik Basu's (2014, p. 464) therefore calls the claim that external validity (or claims to "what works") can be achieved scientifically "for the most part, a delusion".[115] Interestingly, this is fundamentally at odds with the rationale that RCTs can obtain "reliable answers about the best ways to fight global poverty", which is the rationale for awarding Duflo, Banerjee, and Kremer the Nobel Prize in Economics in 2019 (The Prize in Economic Sciences, 2019).

6.4 Ways Forward

Now that I have tried to show that RCT proponents often overstate the potential for extrapolation of results, an open question is what this means for the way forward. In this section I am arguing for the value of 3 related concepts: 1) understanding RCTs as a research method among others, 2) the use of more explicit theories in experiments, and focusing on "mechanism" experiments instead of project or program evaluations, and 3) re-emphasizing the value of falsification to advance knowledge.

[115] However, Basu (2014) is also inconsistent. One the one hand he says there is no such thing as external validity, but he also argues in the same article he states that repeated trials can help to understand why certain things happen.

6.4.1 RCTs as a Research Method

As stated earlier, this chapter does not argue that RCTs are a not a valuable research method (as exemplified by its wide application in my doctoral thesis), quite the opposite. However, they should only be used when the academic problem at hand can be meaningfully advanced by its use, and researchers must be more careful than they have been with making predictions from past experiments.

One viewpoint is that RCTs should be "re-relegated" to what they are supposed to be: a great method to establish causal relationships when it is possible to implement them.[116] However, it remains just one tool in the toolbox of a social scientist and does not per se trump other methods, as each method is designed to answer a different set of questions (Woolcock, 2013; Deaton and Cartwright, 2018). There is no inherent hierarchy of research methods (Ogden, 2017, p. xxiv, calls this the "Nothing Magic Critique" of RCTs). Methods are designed to answer different types of questions. Thus, RCTs are by no means the "gold standard" of research methods (de Souza Leão and Eyal, 2019), they do not automatically trump other forms of research (Deaton, 2010). The fact that not all questions can and should be answered by means of RCT is a point that even Banerjee and Duflo (2017) concede (even if only in a footnote). Ideally, IEs should be used in concert with other research methods. As stated by Deaton and Cartwright (2018, p. 3): "You cannot know how to use trial results without first understanding how the results from RCTs relate to knowledge that you already possess about the world, and much of that knowledge is obtained by other methods." RCTs are therefore one useful tool in the portfolio of scientific methods of evaluation, and it is important and correct that they now also find application in the context of development. Research findings, in the aggregate, can guide policy in areas where policymakers have often needed anecdotal evidence. Nevertheless, they are by no means the royal road to combating poverty and other development problems, as some have claimed. These critiques are not aimed at the method per se, but rather it's application (Basu, 2005).

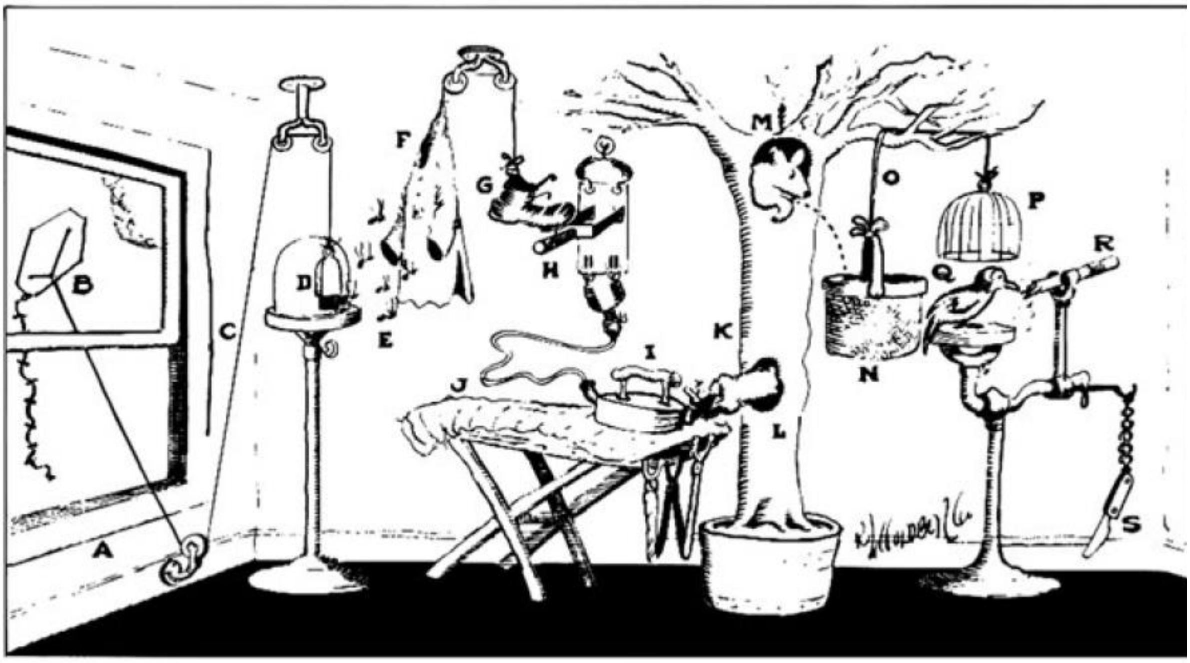
6.4.2 The Value of Theory & Mechanism Experiments

The great value of impact evaluations is that they can often credibly show that something *worked* (in the past) by demonstrating high levels of internal validity. However, as demonstrated, they cannot tell us whether it will work in the future, or *why* it worked. We can observe a cause and an effect, but traditional RCTs do not allow us to look into the "black box" of *why*. This phenomenon

[116] "The price for this success is a focus that is too narrow to tell us 'what works' in development, to design policy, or to advance scientific knowledge about development processes. Project evaluation using randomized controlled trials is unlikely to discover the elusive keys to development, nor to be the basis for a cumulative research program that might progressively lead to a better understanding of development. (…) I argue that evidence from randomized controlled trials has no special priority (...). Randomized controlled trials cannot automatically trump other evidence, they do not occupy any special place in some hierarchy of evidence" (Deaton, 2010, p. 426).

can be amusingly illustrated by the Rube Goldberg machine (Figure 6.2; Cartwright and Hardie, 2012, p. 77), in which flying a kite sharpens a pencil through a maze of different mechanisms.

Figure 6.2 – Rube Goldberg Machine (Cartwright and Hardie, 2012, p. 77)



The Professor gets his think-tank working and evolves the simplified pencil sharpener.

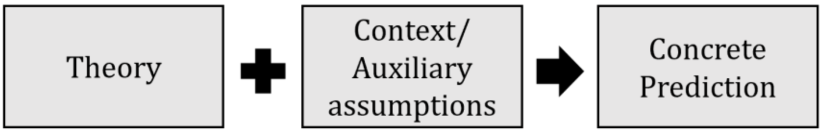
Open window (**A**) and fly kite (**B**). String (**C**) lifts small door (**D**), allowing moths (**E**) to escape and eat red flannel shirt (**F**). As weight of shirt becomes less, shoe (**G**) steps on switch (**H**) which heats electric iron (**I**) and burns hole in pants (**J**). Smoke (**K**) enters hole in tree (**L**), smoking ou opossum (**M**) which jumps into basket (**N**), pulling rope (**O**) and lifting cage (**P**), allowing woodpecker (**Q**) to chew wood from pencil (**R**), exposing lead. Emergency knife (**S**) is always handy ir case opossum or the woodpecker gets sick anc can't work.

Banerjee and Duflo (2017) concede this point (but only in the same footnote referenced earlier) – that RCTs evaluating programs are indeed "black boxes" which do not reveal the true mechanisms of interest. Critics emphasize this as a crucial weakness, the inability of most RCTs to uncover the underlying theory of *why* things work (due to their design), which would be the key to generalizability. And further by Angus Deaton: "For an RCT to produce 'useful knowledge' beyond its local context, it must illustrate some general tendency, some effect that is the result of a mechanism that is likely to apply more broadly." (Deaton, 2010, p. 448). Vivalt (2015) finds in a systematic review of randomized trials that most papers do not specify a model or "causal chain" of the mechanism through which the intervention is supposed to work and fail to provide basic information about the context of the intervention. In this light, the "collection of random results" that Duflo (2006) identified as a result of a series of RCTs might be rather a collection of random contexts, which were poorly (or not at all) accounted for by stating theories about how the mechanisms that were studies would create an effect.

Thus, RCTs would be able to look into the "black box" if they were better designed to do so. In this light, the critique can also be understood as a reply to economists running RCTs in an unstructured manner. Many RCTs these days are conducted, not because they attempt to test a theory, but simply because they are possible to do, or "easy" to implement.

A thorough diagnostic of the status quo ahead of implementing a project is often lacking (see e.g. Rodrik, 2008; de Souza Leão and Eyal, 2019). After a diagnostic process, ideally, a theory would be formulated about what intervention is needed to improve a certain situation, what mechanisms would bring about the change, and why it could work in the given context. The contextual factors are important, as they can render an intervention more or less effective. A tire might be more slippery in the rain or on ice than on solid ground. Ignoring the context of a test in this situation would be a blunder. In a way, in many cases, an experiment can say more about the context, than it does about the intervention. Thus, if these steps are avoided, Deaton (2010, p. 429) accuses randomistas of letting the light fall where it may, "and then proclaim that whatever it illuminates is what we were looking for all along." These are essentially calls for better rooting studies in economic theory and making the theories that are being tested explicit (and potentially even registering it in a trial registry) before starting a study. Earp and Trafimow (2015, p. 7) say that a theory should be stated as a *prediction* in this form: "If the theory is true and a set of auxiliary assumptions is true, an observation should occur." This contains 3 crucial elements, as illustrated in Figure 6.3 (which is based on the above quote by Earp and Trafimow), namely setting out a theory of how a project or program (or mechanism works), describing the impact and shape of the context to the extent possible, and predicting an outcome (which enables falsification).[117]

Figure 6.3 – Example of a Predictive Theory



In this line of reasoning, Kremer and Holla (2009) question the value of socio-scientific predictive power in general if we do not use theories as the starting point for research:

> If our theories are not very good and the impact of treatment depends on context in a way that is complicated, subtle, and difficult to predict, results from one setting are unlikely to generalize in other settings that may look similar to reasonable people. If indeed it is so difficult to generalize, then this raises questions not simply about randomized evaluations

[117] DellaVigna et al. (2019) recently published an article in Science that stresses the value of making ex ante predictions of outcomes in the social science and even propose a prediction registry.

> but more generally about the extent to which one can learn from social science. For example, if treatment effects vary across countries, then cross-country estimates of the impact of different policies or institutions will typically yield biased estimates. (Kremer and Holla, 2009, p. 94)

Deaton argues further, emphasizing the value of theory:

> (…) [T]he validity of evidence-based policy depends on the weakest link in the chain of argument and evidence, so that by the time we seek to use the experimental results, the advantage of RCTs over matching or other econometric methods has evaporated. In the end, there is no substitute for careful evaluation of the chain of evidence and reasoning by people who have the experience and expertise in the field. The demand that experiments be theory-driven is not guarantee for success, though the lack of it is close to a guarantee of failure. (Deaton, 2010, p. 450)

World Bank RCTs are mostly done in collaboration with project managers, which are responsible to design and run the projects to which the researchers then link their studies. One of the first questions I ask a project manager when embarking on such a study is always: "What would a success for this project would look like in your eyes?" Or: "Which indicator would you expect to go up (or down) for this project to be a success?" These questions force the project manager to think through the theory of change for the project, something that is – to my surprise – often not done in a thorough way.[118] For a (hypothetical) education project, the mandate might be to build schools, hire and train teachers in a remote area. If this is achieved, it still not might translate into *impacts* that we are most interested in, which is that students are actually learning more or better when compare to the counterfactual. Student learning could be approximated by test scores in math or reading going up. If a theory is constructed in this way, including "testable content", it can be easily subjected to a test by means of an RCT.

Deaton (2010) actually argues to conduct *more* experiments, but of a different kind, namely with a smaller focus that are not evaluating an entire project, but the underlying mechanisms and links to theory (see e.g. Ludwig et al., 2011 for the use of "mechanism experiments").[119] Deaton (2010, p. 450) argues further: "Randomized experiments, which allow the researcher to induce controlled variance, should be a powerful tool in such programs and make it possible to construct tests of theory that might otherwise be difficult or impossible." Thus, instead of trying to test if

[118] Legovini et al. (2015) find that linking an IE to a project speeds up the disbursement rate of the project as a benefit in addition to the value of the research. This could be linked to researchers actively engaging with project managers, which leads to more efficient implementation.

[119] Earp and Trafimow distinguish semantically between "conceptual" and "direct" replications whereas conceptual replications study a theory and direct replications investigate a "fact or finding".

projects or programs “worked”, randomized trials could also be used to test theories around behavioral mechanisms such as loss aversion, price elasticities for demand, procrastination etc. (as argued by Deaton, 2010; Ludwig et al., 2011). They should be focused on *why* something works (the mechanism/theory), not *whether* it works or not (which ignores the “black box”). Then, “the project (…) is an embodiment of the theory that is being tested and refined, not the object of the evaluation, in its own right” (List, 2007; cited in Deaton, 2010, p. 451).

The mere replication of IEs on project will likely not produce a lot of learning: “This means that if the World Bank had indeed randomized all of its past projects, it is unlikely that the cumulated evidence would contain the key to economic development” (Deaton, 2010, p. 442). Mechanisms, however, can be tested in different circumstances, so Deaton argues, and maybe more consistent patterns emerge.[120] Going back to Figure 6.3, mechanism experiments basically attempt to “cut out” (or at least reduce) the middle element, since the auxiliary assumptions about contexts are notoriously hard to measure or even to grasp. I argued earlier that the use of randomized trials is a reduction of scope from larger questions (e.g how countries develop) to smaller questions (how do we ensure children go to school). This argument for mechanism experiments can be understood as a further reduction of dimensionality, and an attempt to answer even smaller questions.

The scattered results found in the meta study by Vivalt (2019) referred to earlier may be an expression of this lack of focus on smaller mechanisms and theories. Mechanism experiments may have greater potential to find common threads in behavior, and therefore to pin down part of a structure (Deaton and Cartwright, 2018) which is more coherent than just grouping larger project evaluations together. Ludwig et al. (2011) illustrate an example of a mechanism experiment: Obesity is a problem in a certain neighborhood, and some believe this is caused by lack of fresh fruit vendors in the area (i.e. the problem of a "food desert"). Instead of providing incentives for fruit/vegetable vendors to open shops, researchers could provide randomly assigned families with a basket of fresh fruits every week and analyze the effect on obesity. If this high “dose” of "treatment" shows no effect on obesity, we can assume that incentives for new businesses were a waste. Through minimalizing the scope of the experiment, we can draw conclusions about the theory behind, in this case, inner-city obesity, which in turn might generate a greater input for policymaking than a policy experiment on the grand scale. If an unrealistically high dose of the

[120] Earp and Trafimow (2015) speak of the challenge in social psychology that theories are often formulated “too loosely”. This, mechanism experiments would be helpful, because they imply that the theories that are being tested are “tighter”.

proposed treatment shows no effect, we can confidently rule out the hypothesis and divert resources towards other questions.[121]

This value of smaller focused experiments is echoed by Woolcock (2013) who defines the "causal density" or complexity of development problems. The lower the causal density of an intervention – i.e. the closer it is to a mechanism experiment – the better suited it is to be evaluated via an RCT as it is possible to discern or isolate components more easily. In contrast, if an experiment shows that a larger project works, we are often left guessing which of the many mechanisms therein was responsible for the effect we see. Woolcock (also) also points out that recent "successes" are stemming from experiments studying phenomena with relatively low causal density (textbooks, malaria nets, deworming pills etc.; see section 1.2 for examples of "successes" of RCTs). The obvious problem is that it remains hard to assess which development problems are of low and which ones are of high complexity (for which Woolcock proposes a grading structure). Some seemingly smaller issues might persist being hard to solve but learning about this in a structured manner would also be of value.

Deaton (2010) summarizes: "(…) RCTs of 'what works' (…) are unlikely to be helpful for policy or to move beyond the local unless they tell us something about why the program worked, something to which they are often neither targeted nor well suited" (Deaton, 2010, p. 448). He mentions some positive recent examples of the use of theory on smaller mechanisms (Deaton, 2010, p. 450-1):

- Karlan and Zinman (2008), who are concerned with the price elasticity of the demand for credit;
- Bertrand et al. (2010), who take predictions about the importance of context from the psychology laboratory to the study of advertising for small loans in South Africa;
- Duflo, Kremer, and Robinson (2009), who construct and test a behavioral model of procrastination for the use of fertilizers by small farmers in Kenya; and
- Giné, Karlan, and Zinman (2010), who use an experiment in the Philippines to test the efficacy of a smoking-cessation product designed around behavioral theory.[122]

Similarly Kremer and Glennerster (2011) set up a series of theories on price sensitivity for take-up of preventative health products, of which some were upheld, and others refuted by Dupas and Miguel in a recent summary of the evidence (2017; see also Banerjee et al., 2017 for a summary). Deaton and Cartwright (2018) point towards the "Graduation from Ultra Poverty" (GUP) studies (Banerjee et al., 2015) as an excellent example of a contribution to the theory of economic

[121] This example was also mentioned in Dunsch (2012).

[122] The examples are also mentioned in Banerjee (2005), five years earlier, which maybe illustrates the dearth of theory-driven experimental papers.

development – which is probably not how the authors themselves conceived the series of studies (rather focusing on the "what works" idea). The project tested a theory (large injections of capital can lead to development) and failed to falsify it – which does not mean that the theory is conclusively proven – but still advances our thinking.

Thus, theories and experiments should reinforce each other, they should not be substitutes. Theories should be based on past experimental and observational research (quantitative and qualitative) and can then be refuted by experiments, which helps to shape updated and falsifiable theories:

> Technique is never a substitute for the business of economics. (…) It took scientific understanding to overcome the heterogeneity of experience, which ultimately defeats trial and error. As was the case then, so it is now, and I believe we are unlikely to banish poverty in the modern world by trials alone unless they are guided by and contribute to theoretical understanding. (Deaton, 2010, p. 452)

6.4.3 Falsification

Humans seem to be naturally inclined to "look for proof", rather than for counterarguments to their position or world view. Current incentives in science are not aligned with the importance of falsificationism. Related concepts are the psychological concept of "confirmation bias" (Nickerson, 1998), or the "narrative fallacy" (Taleb, 2007; Taleb 2011). Related to confirmation bias, publication bias is a large problem in science (Dickersin, 1990). Basu (2014) gives and example: If 10 research teams investigate the impact from C on I and 9 teams find no effect, these 9 teams won't be able to publish a paper. The 10th team which finds an effect will be able to publish a paper even though it is likely that their result is the result of chance or a funky random draw (in a RCT setting). However, this paper will be widely cited and might become seminal. Maniadis et al. (2014) also confirm that "surprising" results are necessary in order for researchers to get published, and they go on to argue that these are, more often than not, false. DellaVigna et al. (2019) recently proposed to systematize ex ante expert predictions of trial outcomes, so that 1) researchers are less inclined to say that their findings where what they were looking for all along ("hindsight bias"), and 2) research results can be compared to these recorded predictions, rather than to the null hypothesis of no effect (which is currently the scientific standard and hampers growth of knowledge), which would mitigate publication bias. It is easy to obtain confirmations for nearly every theory – if that's what one's out to find (Popper, 2014). Yet, negative or "zero" results can advance our knowledge no less, and sometimes even more, than those that display positive impacts. Their power then is to *reject* hypotheses rather than *buttressing* them. By *falsifying* mechanisms that are less likely to work, we inch closer to *what really works* in development. It is, for that reason, important that no-results-IEs lose their stigma and attain the attention they deserve.

The focus on the value of falsification is stressed by Popper and linked to “piecemeal” rather than “holistic” or “utopian” social engineering (the latter which is more in line with the “what works” agenda):

> The piecemeal engineer knows, like Socrates, how little he knows. He knows that we can learn only from our mistakes. Accordingly, he will make his way, step by step, carefully comparing the results expected with the results achieved, and always on the look-out for the unavoidable unwanted consequences of any reform; and he will avoid undertaking reforms of a complexity and scope which makes it impossible for him to disentangle causes and effects, and to know what he is really doing. Holistic or Utopian social engineering, as opposed to piecemeal social engineering ... aims at remodeling the 'whole of society' in accordance with a definite plan or blueprint. (Popper, 2002, p. 61; cited in: Easterly, 2008, p. 13)

The essence of piecemeal reform is "searching for, and fighting against, the greatest and most urgent evils of society" in contrast to "searching for, and fighting for, its greatest ultimate good", which utopian social engineering entails (Popper, 1971).

Thus, IEs might be most valuable if they assist in disproving or deconstructing theories, in line with Smith and Ebrahim (2002) who state that observational studies “propose” and RCTs have the power to “dispose”. In this light, chapters 2 and 3 of this dissertation can also be seen as disproving certain aspects of the projects under study. They show for example that effects for both projects wane with time. This form of falsification shows that the next project would need to be designed differently (a new theory), to see if effects can persist over time. This point links to the argument presented earlier on formulating theories before conducting RCTs. Theories should be explicitly stated in ways that allow for their refutation in the future: “I shall require that [the] logical form [of the theory] shall be such that it can be singled out, by means of empirical tests, in a negative sense: it must be possible for an empirical scientific system to be refuted by experience” (Popper, 2005, p. 18). Theories that are stated in a way that cannot be falsified are in the realm of non-science (or pseudo-science) rather than science.

The idea of moving away from publication and confirmation bias is expressed through slowly progressing (or niche) efforts to create journals for “negative results” only, e.g. the Journal of Negative Results in BioMedicine. Basu (2014, p. 463) also calls for this – saying that such as journal would have a “sobering effect on economics” – and recognizes the Campbell Collaboration for its efforts to create repositories that store all experiments, whether they show publishable results or not. Box 1 shows examples of “failed” experiments, which are nevertheless very valuable for learning.

Historically, the (mostly non-scientific) trial-and-error method drove human progress (see e.g. Cohen and Easterly, 2010). This process happened at different speeds at different times, but, due to economic necessities, or democratic processes, countries shifted policies to a path of more sustainable governance. Ravallion (2008) cites the example of China that uses unstructured experimentation (not RCTs) to advance its reform progress (see also Rodrik, 2010). In most developed countries, the principle institution in the process of governing and forming societies are open “markets of ideas”: "We need freedom because we are ignorant." (Manzi, 2012, p. xxii). I would argue that more focus on the value of falsification can assist this natural "trial-and-error" process of policymaking by speeding up the feedback cycle. As laid out, IEs can never reliably verify whether a policy is worth implementing. However, they possess the potential to *falsify* existing theories (or mechanisms) so that the trial-and-error process is accelerated, and better working policies can be identified faster in the political process. Akin to IBM founder Thomas Watson’s premise: “If you want to succeed, raise your failure rate” (cited in Rodrik, 2010, p. 40). There is no guarantee that theories that have not been proven wrong yet, will continue to be upheld in future trials. However, theories (or policies) that have not been falsified yet hold an advantage of theories that have been falsified. Mokherjee (2005, p. 4330) suggests the “use of the least unsuccessful theory from the standpoint of empirical verification for purposes of prediction and policy evaluation.” By the same token this means that policies that have been *verified* are not better than those that have not been verified. This point ties in with the argument on context variance. An experiment that shows “what works” – a verification – we can never be sure whether the effect was caused by the treatment, or by the contextual factors in the given situation, even after multiple replications. This is well illustrated by Vivalt (2019), who shows the wide dispersion of experimental results for the same type of treatment. Falsifications, however, provide certainty: If a treatment does not lead to a discernible difference in outcomes, we know for sure that either the treatment doesn’t work, or that the context mitigated the treatment effect. A falsification necessitates the alteration of the existing theory that was tested, either by tweaking the treatment, or applying the same treatment in another context. This is markedly different from current practice where replications are usually conducted for “successful” studies, also because the field is often not aware of falsifications, i.e. studies with null results. Even worse: Researchers currently have strong economic incentives to tweak results of “accidental” falsifications, so that some positive result can be rescued, e.g. for a subgroup, just to avoid not having anything to show for. These types of incentives actively prohibit faster scientific progress.

In sum, IEs are a great research tool that should be used when possible. RCTs can be useful for program evaluations, hypothesis testing (theory testing), or establishing proof of concept (Mookherjee, 2005; Deaton and Cartwright, 2018). However, the proponents of the method have been overstating its value as a tool to provide immediate and actionable policy-advice. While RCTs are very valuable, they unfortunately won't be able to tell us "what works" with certainty. Rather than searching for "what works", it would be sufficient to 1) state "what worked" instead of "what works", 2) set new explicit theories around programs and smaller mechanisms, and 3) repeatedly try to falsify these. Ideally this proposed process speeds up the trial-and-error process of policymaking through faster and highly credible dispositions, which can lead to better project designs (and therefore outcomes) sooner.

Box 1: Examples – Falsifications

Below are some examples of *successful falsifications of prior beliefs*.

1. Distribution of Schoolbooks in Sierra Leone

IEs show that simply providing an (otherwise valuable) input is not enough. Input *use* is apparently what matters: Research in Sierra Leone found that free textbook provision had only modest impacts on teacher behavior and no effects on students' learning results. It was found that a large majority of new books were not used but rather stored and locked-up.
The tendency to store is apparently cortable related with uncertainties of head-teachers towards future government spending on these items. Refuting the old theory – providing schoolbooks raises educational scores – gave way for a new hypothesis: The indication that uncertainty can play a stark role for educational achievements. Other impact evaluations corroborate these results as they show that simple school *inputs* often fail to achieve the desired goals (Sabarwal and Evans, 2014).

2. Community-driven development (CDD)

The rationale for community-driven development programs is straightforward: Community members working cooperatively to select, manage and monitor development projects is meant to increase accountability, competence and inclusiveness of local institutions. Especially in post-conflict situations, CDD is believed to restore social cohesion.

But not so fast. Recent IE research sheds light on the fact that CDD programs often fail to cause better development outcomes. For example, the evaluation of a World Bank CDD program in South Sudan (Sudan Community Development Fund, CDF) showed no significant differences in villagers' behavior – measured through behavioral games as well as self-reported (Avdeenko and Gilligan, 2015). Social capital in targeted communities has not increased. Also, the perception of social cohesion did not differ in treated versus control communities. These findings confirm prior findings about poor CDD effectiveness.

This does not mean that the CDF was worthless. The results show that we need to think harder about what mechanisms work and which do not since CDD programs, as they were ran in the past, were not showing the expected results in all settings, at least not in the short run.

3. Youth Employment – Training

For Malawian youth – notably young women – low skill levels are often coupled with high rates of unemployment, extreme poverty, high-risk sexual behavior, and HIV prevalence. However, to date, rigorous analysis of the impact of youth employment programs in developing countries is scant. The best available evidence comes from few randomized impact evaluations in Latin America, where these programs were found to have some positive effects on labor outcomes, but only for certain groups; for example, female trainees. The Apprenticeship Training and Entrepreneurial Support Program (ATESP) aims to provide vulnerable youth in Malawi an opportunity to enhance their employability and earning potential, thus reducing high risk behavior that increases vulnerability to HIV infection. The project was accompanied by the first IE of a skills training youth employment program in Africa (Cho et al., 2013). The gender equity analysis finds – contrasting results from Latin America – negative effects on labor outcomes and business activities of female trainees in the short-term (4 months after the training). Although health outcomes were positive, the primary goal of the project was not achieved (for women). This example teaches us, that positive labor outcomes by providing training per se is not a given. The results indicate that training is not enough. Quality and design matter.

6.5 References

Alcott, H., & Mullainathan, S. (2012). External Validity and Program Selection Bias. *NBER Working paper*, *18373*.

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, *24*(2), 3-30.

Athey, S., & Imbens, G. W. (2017). The Econometrics of Randomized Experiments. In *Handbook of Economic Field Experiments* (Vol. 1, pp. 73-140). North-Holland.

Avdeenko, A., & Gilligan, M. J. (2015). International interventions to build social capital: evidence from a field experiment in Sudan. American Political Science Review, 109(3), 427-449.

Bacon, F., (1859.) Novum organum. In: Ellis, R.L., Spedding, J. (Eds.), The Philosophical Works of Francis Bacon. Longmans, London, England.

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

Banerjee, A. V. (2007). *Making aid work*. MIT press.

Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. *Annu. Rev. Econ.*, *1*(1), 151-178.

Banerjee, Abhijit, 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, no. 6236 (2015): 1260799.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M. and Walton, M. (2017a). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, *31*(4), 73-102.

Banerjee, A., Duflo, E., Imbert, C., Mathew, S., & Pande, R. (2017b). DP11761 E-governance, Accountability, and Leakage in Public Programs: Experimental Evidence from a Financial Management Reform in India. CEPR Discussion Paper.

Banerjee, A. & Duflo, E., (Eds.). (2017). *Handbook of Field Experiments* (Vol. 2). Elsevier.

Banerjee, A. V. (2005). 'New Development Economics' and the challenge to theory. *Economic and Political Weekly*, *40*(40), 4340-4344.

Banerjee, A. V., Duflo, E., & Kremer, M. (2016). The influence of randomized controlled trials on development economics research and on development policy. *Mimeo MIT*.

Barrett, C. B., & Carter, M. R. (2014). A retreat from radical skepticism: rebalancing theory, observational data, and randomization in development economics. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, 58-77.

Basu, K. (2005). New empirical development economics: remarks on its philosophical foundations. *Economic and Political Weekly*, *40*(40), 4336-4339.

Basu, K. (2014). Randomisation, causality and the role of reasoned intuition. Oxford Development Studies, 42(4), 455-472.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., & Zinman, J. (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, *125*(1), 263-306.

Bold, T., Kimenyi, M., & Mwabu, G. (2013). Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American educational research journal*, *5*(4), 437-474.

Burke, E. (1854). The works of the right honourable Edmund Burke (Vol. 1). Henry G. Bohn.

Cameron, D. B., Mishra, A., & Brown, A. N. (2016). The growth of impact evaluation for international development: how much have we learned?. *Journal of Development Effectiveness*, *8*(1), 1-21.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, *54*(4), 297.

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio Books.

Cartwright, N. (2007). Are RCTs the gold standard?. *BioSocieties*, *2*(1), 11-20.

Cartwright, N. (2010). What are randomised controlled trials good for?. *Philosophical studies*, *147*(1), 59.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on psychological science*, 9(1), 40-48.

Cho, Y., Kalomba, D., Mobarak, A. M., & Orozco, V. (2013). Gender differences in the effects of vocational training: Constraints on women and drop-out behavior. The World Bank.

Chong, A., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2014). Letter grading government efficiency. *Journal of the European Economic Association*, *12*(2), 277-298.

Clasen, T. F., Alexander, K. T., Sinclair, D., Boisson, S., Peletz, R., Chang, H. H., ... & Cairncross, S. (2015). Interventions to improve water quality for preventing diarrhoea. *The Cochrane Library.*

Cohen, J., & Easterly, W. (Eds.). (2010). *What Works in Development?: Thinking Big and Thinking Small*. Brookings Institution Press.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Davis, J., Guryan, J., Hallberg, K., & Ludwig, J. (2017). *The Economics of Scale-Up* (No. w23925). National Bureau of Economic Research.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature*, 424-455.
Deaton, A., & Cartwright, N. (2016). *Understanding and Misunderstanding Randomized Controlled Trials* (No. w22595). National Bureau of Economic Research.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science & Medicine, 210, 2-21.

DellaVigna, S., Pope, D., & Vivalt, E. (2019). Predict science to improve science. *Science*, *366*(6464), 428-429.

de Souza Leão, L., & Eyal, G. (2019). The rise of randomized controlled trials (RCTs) in international development in historical perspective. Theory and Society, 1-36.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Jama*, *263*(10), 1385-1389.

Duflo, E. (2006). Field experiments in development economics. *Econometric Society Monographs*, *42*, 322.

Duflo, E., Kremer, M., & Robinson, J. (2009). Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya.

Duflo, E., & Kremer, M. (2005). Use of randomization in the evaluation of development effectiveness. *Evaluating development effectiveness*, *7*, 205-231.

Duflo, E. (2005). Evaluating the impact of development aid programmes: the role of randomised evaluations. *Development Aid: Why and How? Towards strategies for effectiveness*.

Dunsch, F. A., Evans, D. K., Eze-Ajoku, E., & Macis, M. (2017). *Management, Supervision, and Health Care: A Field Experiment* (No. w23749). National Bureau of Economic Research.

Dupas, P., & Miguel, E. (2017). Impacts and determinants of health levels in low-income countries. In *Handbook of economic field experiments* (Vol. 2, pp. 3-93). North-Holland.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, *6*, 621.

Easterly, W. (2008). *Reinventing foreign aid* (Vol. 1). The MIT Press.

Elster, Jon. 2000. Rational Choice History: A Case of Excessive Ambition. American Political Science Review 94: 685-695.

Evans, D., & Yuan, F. (2017). The economic returns to interventions that increase learning. *Background paper, World Bank, Washington, DC*.

Eyben, R. (2010). Hiding relations: the irony of 'effective aid'. The European Journal of Development Research, 22(3), 382-397.

Eyben, R. (2012). Relationships for aid. Routledge.

Fisher, R. A. (1925, July). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 22, No. 5, pp. 700-725). Cambridge University Press.

Giné, X., Karlan, D., & Zinman, J. (2010). Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics*, *2*(4), 213-35.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, *13*(3), e1002106.

Hume, D. (2000). *An enquiry concerning human understanding: A critical edition* (Vol. 3). Oxford University Press.

Jamison, J. C. (2017). *The entry of randomized assignment into the social sciences*. The World Bank.

Kahneman, D. (2002). Nobel Prize Lecture.'. Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice,'December, 8, 2002.

Karlan, D. S., & Zinman, J. (2008). Credit elasticities in less-developed economies: Implications for microfinance. *American Economic Review*, *98*(3), 1040-68.
Keynes, J. M. (1939). 1973." Professor Tinbergen's Method.". The Collected Writings of John Maynard Keynes, 14.

Kremer, Michael und Michael Holla. 2009. Pricing and Access: Lessons from Randomized Evaluations in Education and Health. In: Jessica Cohen und William Easterly (Hrsg.), *Thinking Big and Thinking Small.* Washington D.C.: Brookings Institution Press, 91-119.

Kremer, M., & Glennerster, R. (2011). Improving health in developing countries: evidence from randomized evaluations. In *Handbook of Health Economics* (Vol. 2, pp. 201-315). Elsevier.

Kristof, Nicholas D. 2011. Getting smart on aid. In: *New York Times*, 18.05.11: A27.

Laplace, P. S. (1902). A philosophical essay on probabilities. Wiley.

Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the growth of knowledge: Volume 4: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Cambridge University Press.

Legovini, A., Di Maro, V., & Piza, C. (2015). Impact evaluation helps deliver development projects.

List, J. A. (2007). Field experiments: a bridge between lab and naturally occurring data. *The BE Journal of Economic Analysis & Policy*, *5*(2).

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. Science, 355(6325), 584-585.

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives*, *25*(3), 17-38.

Majid, M. F., Kang, S. J., & Hotez, P. J. (2019). Resolving" worm wars": An extended comparison review of findings from key economics and epidemiological studies. *PLoS neglected tropical diseases*, *13*(3), e0006940.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. American Economic Review, 104(1), 277-90.

Manzi, J. (2012). *Uncontrolled: the surprising payoff of trial-and-error for business, politics, and society*. Basic books.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean?. American Psychologist, 70(6), 487.

McKenzie, D., & Woodruff, C. (2013). What are we learning from business training and entrepreneurship evaluations around the developing world?. *The World Bank Research Observer*, *29*(1), 48-82.

McKenzie, D. (2016). Have RCTs taken over development economics?. *World Bank Blog on Impact Evaluations*, *13*.

Meager, R. (2015). Understanding the impact of microcredit expansions: A bayesian hierarchical analysis of 7 randomised experiments. *arXiv preprint arXiv:1506.06669*.

Miguel, E., & Kremer, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72(1), 159-217.

Mill, J. S. (1884). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation* (Vol. 1). Longmans, green, and Company.

Mookherjee, D. (2005). Is there too little theory in development economics today?. Economic and Political Weekly, 40(40), 4328-4333.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2), 175-220.

Ogden, T. N. (Ed.). (2017). *Experimental Conversations: Perspectives on Randomized Trials in Development Economics*. MIT Press.

Parker, Ian. 2010. "The Poverty Lab. Transforming development economics, one experiment at a time." The New Yorker, May 17, 2010.

Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the Tropics–Development Policy, Randomized Controlled Trials, and External Validity. *The World Bank Research Observer*, *33*(1), 34-64.

Popper, K. (1971). The Open Society and Its Enemies. V. I, II. Princeton, New Jersey.

Popper, K. (2002). The poverty of historicism. Routledge.

Popper, K. (2005). The logic of scientific discovery. Routledge.

Popper, K. (2014). Conjectures and refutations: The growth of scientific knowledge. Routledge.

Pritchett, L., Samji, S., & Hammer, J. S. (2013). It's all about MeE: Using Structured Experiential Learning ('e') to crawl the design space.

Pritchett, L., & Sandefur, J. (2014). Context Matters for Size: Why External Validity Claims and Development Practice do not Mix. *Journal of Globalization and Development*, *4*(2), 161-197.

Pritchett, L., & Sandefur, J. (2015). Learning from experiments when context matters. *American Economic Review*, *105*(5), 471-75.

Prowse, M. (2007). Aid effectiveness: the role of qualitative research in impact evaluation. *ODI Background Note, December*.

Ravallion, M. (2008). Are there lessons for Africa from China's success against poverty?. The World Bank.

Roberts, R. (2015). Morten Jerven on African Economic Growth. *EconTalk Episode with Morten Jerven.* Weblink: https://archive.li/zPMeV*.* Accessed on 18 April, 2019.

Rodrik, D. (2008). The new development economics: we shall experiment, but how shall we learn? Harvard Working Paper.

Rodrik, D. (2010). Diagnostics before prescription. Journal of Economic Perspectives, 24(3), 33-44.

Rothstein, J., & Von Wachter, T. (2017). Social experiments in the labor market. In Handbook of Economic Field Experiments (Vol. 2, pp. 555-637). North-Holland.

Rothwell, P. M. (2005). External validity of randomised controlled trials:“to whom do the results of this trial apply?”. *The Lancet*, *365*(9453), 82-93.

Russell, B. (2001). The problems of philosophy. OUP Oxford.

Sabarwal, S., Evans, D. K., & Marshak, A. (2014). The permanent input hypothesis: the case of textbooks and (no) student learning in Sierra Leone. The World Bank.

Silver, N. (2012). The signal and the noise: why so many predictions fail--but some don't. Penguin.

Smith, G. D., & Ebrahim, S. (2002). Data dredging, bias, or confounding: they can all get you into the BMJ and the Friday papers. *BMJ: British Medical Journal*, *325*(7378), 1437.

Steinmetz, G. (1998). Critical realism and historical sociology. A review article. *Comparative studies in society and history*, *40*(1), 170-186.

Stockmann, R. (2012). Von der Idee zur Institution Institut für deutsche Entwicklungsevaluierung gegründet. *Zeitschrift für Evaluation*, *11*(1), 85.

Stokes, S. C. (2014). A defense of observational research. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, 33-57.

Taleb, N. (2005). *Fooled by randomness: The hidden role of chance in life and in the markets* (Vol. 1). Random House Incorporated.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable* (Vol. 2). Random house.

Teele, D. L. (Ed.). (2014). *Field experiments and their critics: essays on the uses and abuses of experimentation in the social sciences*. Yale University Press.

Tetlock, P. E., & Gardner, D. (2015). Superforecasting: The art and science of prediction. Signal.

The Prize in Economic Sciences. (2019). NobelPrize.org. Nobel Media AB 2019. Mon. 14 Oct 2019. Retrieved from: https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/

Tsang, E. W., & Kwan, K. M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management review*, *24*(4), 759-780.

Vivalt, E. (2015). Heterogeneous treatment effects in impact evaluation. *American Economic Review*, *105*(5), 467-70.

Vivalt, E. (2019). How Much Can We Generalize from Impact Evaluations? Unpublished.

Wiblin, R., & Harris, K. (2018, May 15). Dr Eva Vivalt's research suggests social science findings don't generalize. So evidence-based development – what is it good for? [Audio podcast]. Retrieved from https://80000hours.org/podcast/episodes/eva-vivalt-social-science-generalizability/

Woolcock, M. (2013). Using case studies to explore the external validity of 'complex' development interventions. *Evaluation*, 19(3), 229-248.

6.6 Annex

Table 6.1 - Sources of External Invalidity (Bracht and Glass, 1968)

Population Validity	
A. Experimentally Accessible Population vs. Target Population	Generalizing from the population of subjects that is available to the experimenter (the accessible population) to the total population of subjects about whom he is interested (the target population) requires a thorough knowledge of the characteristics of both populations. The results of an experiment might apply only for those special sorts of persons from whom the experimental subjects were selected and not for some larger population of persons.
B. Interaction of Personological Variables and Treatment Effects	If the superiority of one experimental treatment over another would be reversed when subjects at a different level of some variable descriptive of persons are exposed to the treatments, there exists an interaction of treatment effects and personological variable.
Ecological Validity	
A. Describing the Independent Variable Explicitly	Generalization and replication of the experimental results presuppose a complete knowledge of all aspects of the treatment and experimental setting.
B. Multiple-Treatment interference:	When two or more treatments are administered consecutively to the same persons within the same or different studies, it is difficult and sometimes impossible to ascertain the cause of the experimental results or to generalize the results to settings in which only one treatment is present.
C. Hawthorne Effect	A subject's behavior may be influenced partly by his perception of the experiment and how he should respond to the experimental stimuli. His awareness of participating in an experiment may precipitate behavior which would not occur in a setting which is not perceived as experimental.
D. Novelty and Disruption Effects	The experimental results may be due partly to the enthusiasm or disruption generated by the newness of the treatment. The effect of some new program in a setting where change is common may be quite different from the effect in a setting where very few changes have been experienced.
E. Experimenter Effect	The behavior of the subjects may be unintentionally influenced by certain characteristics or behaviors of the experimenter. The expectations of the experimenter may also bias the administration of the treatment and the observation of the subjects' behavior.
F. Pretest Sensitization	When a pretest has been administered, the experimental results may partly be a result of the sensitization to the content of the treatment. The results of the experiment might not apply to a second group of persons who were not pretested.
G. Post-test Sensitization	Treatment effects may be latent or incomplete and appear only when a post-experimental test is administered.
H. Interaction of History and Treatment Effects	The results may be unique because of "extraneous" events occurring at the time of the experiment.
I. Measurement of the Dependent Variable	Generalization of results depends on the identification of the dependent variables and the selection of instruments to measure these variables.
J. Interaction of Time of Measurement and Treatment Effects	Measurement of the dependent variable at two different times may produce different results. A treatment effect which is observed immediately after the administration of the treatment may not be observed at some later time, and vice versa.

7. Conclusion

This dissertation combines 4 self-contained empirical studies from West Africa and closes with a more theoretical chapter on external validity, which has relevancy for how the results and future value of each empirical study can be assessed. Chapter 2 and 3 can be categorized as evaluations of full projects whereas chapter 4 is a survey experiment, and chapter 5 is more akin to a "mechanism experiment" as described by Ludwig et al. (2011; see also chapter 6 for details). While each study is self-contained, they are connected by the fact that

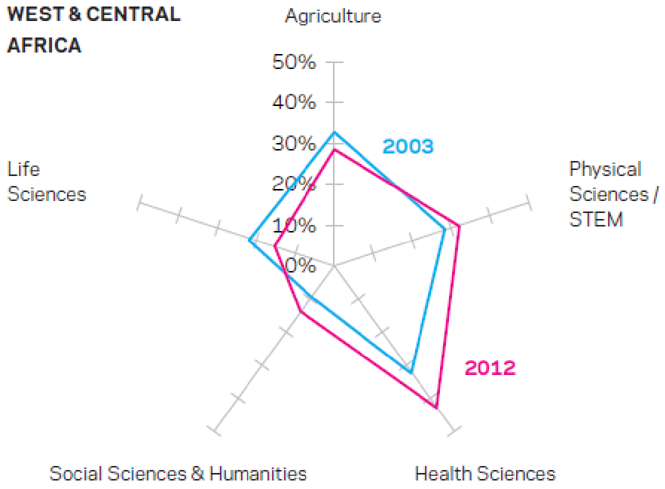
- they are all experimental studies studying cause-and-effect relationships,
- they were all conducted in West Africa (Nigeria and Ghana) – which remains woefully understudied – using primary data, and
- they are investigating policy-relevant questions for which there is to date not a lot of causal research, especially in this region.

The main advantage of IEs is short and powerful: „The core purpose of RCTs is to use random assignment in order to ensure that the unconfoundedness assumption essential to identifying an average treatment effect holds" (Barrett and Carter, 2010, p. 522). With their work using the RCT method, Esther Duflo, Abhijit Banerjee, and Michael Kremer won the Nobel Prize in Economics for having "introduced a new approach to obtaining reliable answers about the best ways to fight global poverty" (The Prize in Economic Sciences, 2019).

In each of the four empirical studies it was possible to causally identify treatment effects – with a high level of internal validity. It was shown that an edutainment movie about safe saving and responsible borrowing in Lagos, Nigeria, can increase the sign-up for savings accounts in the short-term, but behavior was not affected in the longer run (chapter 2), a quality enhancement/supportive supervision intervention in Nigeria had measurable impacts on improving clinical practices that were relatively easy to change in the short term, but effects did not remain in the longer term (chapter 3), patient satisfaction surveys in Nigeria were severely biased by positive framing of questions and should therefore be reconsidered as a method to assess quality of care (chapter 4), and, for community-health officers in Ghana, working in better equipped facilities and having better career opportunities can compete with higher salaries as an incentive to accept (and remain in) rural postings (chapter 5).

Chapter 6 is relevant for all 4 empirical studies and can therefor also be perceived as a connecting chapter (in addition to this conclusion). There I argue that it cannot be claimed that these studies show definitively whether the projects or tested mechanism would work in the same way in other contexts (due to the problem of context variance). The research nevertheless provides value (as all well-done research does in equal measure), as future policymakers and academics can take the results into account when defining theories to be tested, planning new policies or additional research. In fact, chapter 2 and 3, for example, disprove the notion that the found effects persist over time (as they disappear in the longer run).[123] This is valuable as it incentivizes the next project coordinator to potentially tweak the mechanism (set a new theory) to see whether effects can last longer (which is then open for falsification). The articles are particularly important and relevant as experimental (identifying cause-and-effect relationships) research in these fields is still scarce, especially in West Africa. Figure 7.1 shows that this is especially true for the social sciences, which only account for 14% of studies originating in the region (Lan et al., 2014; see also Figure 1.1 in chapter 1).

Figure 7.1 – Percentage of total article output by subject grouping for SSA and South Africa, 2003 vs. 2012 (Lan et al., 2014)



Some of my personal experiences when working on the projects corroborate points that were made in the introductory chapter and in chapter 6. To successfully conduct each study, a large amount of coordination and project management work on our team's behalf was required. The main goal of these efforts was to ensure *internal validity* of the studies, i.e. that the results are indeed accurate causal estimates of (past) impacts.

As a result, one main preoccupation when running these experiments in the field is to make sure that the project under scrutiny is well implemented by the partner organization. Researchers (like me) have a great interest in making sure the project is implemented as planned and the research protocol is adhered to precisely so that the pitfalls explained in chapter 1 are avoided to the extent possible (attrition, spillovers, data accuracy, etc.). Deaton and Cartwright (2018, p. 8)

[123] The RCTs in this dissertation answered rather "small" development questions but did so – in my opinion – convincingly. Whether or not the time and money spent on those was "worth it" is hard to judge now as it is not clear how the results are perceived by other researchers or policymakers.

call this the "policing" of the experiment." It required weekly calls, lots of emails, and frequent mission travel to the project sites to provide some capacity building and to make sure the local partner understands and is on board with the research design. This experience is echoed by de Souza Leão and Eyal (2019, p. 392-3): "A complex organizational effort is required to coordinate the multiple parties and, most importantly, to control the control group and prevent attrition." While I am confident that the studies were conducted with high levels of internal validity, it is of course hard to say whether the same outcomes would be achieved if someone else (e.g. a local NGO) would have been in charge of the research project management (akin to the "piloting bias" referred to in Bold et al., 2013). (This problem, however, is not unique to RCTs, but all quantitative studies that involve primary data collection.)

Each project took multiple years to complete from inception to the final working paper (often up to 5 or even 6 years), which is an additional counterpoint to the "what works" agenda discussed in chapter 6 as the research results could often not be delivered in a timely manner. Cameron et al. (2016) find that it takes 3.9 years from the endline collection to the final publication *on average.* Habitually this is related to the time it takes for us researchers to clean and analyze the data and write up the academic papers (which then need time to be peer-reviewed for quality assurance). So, while the policy issue might have been pressing at the time the research was designed, after 5-years it might not be as pressing anymore, or the Government has changed altogether. Some organizations have realized this and are now focusing on more "nimble" IEs (however often still ignoring the problem of context variance).

There are many examples where the project itself is poorly managed, which leads to a disruption in the research process. Sometimes well- and long-planned research must be halted or canceled because the project itself is not being implemented coherently (by a national government, an NGO, or an international organization like the World Bank). During my time at the World Bank resource-extensive projects that I worked on (for example on a Government bureaucracy reform projects in Nepal and in Guinea) had to be canceled after sometimes years of preparatory work. Due to publication bias and misaligned incentives, learning from these failures is often lost as there is no publication record.

The work presented in this dissertation are the studies that "survived". While there was a lot of supervision work involved, the projects themselves were reasonably well implemented by the Governments and local partners. I can also confirm that the research was only possible by involving an international team, including a local field coordinator for each project. This necessity of a massive international effort unfortunately confirms parts of the critique made by Besley (2012, p. 162) earlier: RCTs can be perceived as an "outside intervention by those who know best", often with a lack of local ownership. The work requires extensive data collection, data preparation, and data analysis expertise. Usually – while I tried where possible – there is also very little room for extensive capacity building.

Another point that likely is prohibitive for developing countries that are interested in conducting these types of studies themselves is that the studies were very expensive, a concern also voiced by Deaton (2010) and others. The main cost driver is the data collection in the countries, for which local survey firms are contracted (a process that I also led for all studies). The combined cost for all 4 empirical studies lies likely between $750,000 and $1,250,000 in total. Gertler et al. (2016) find in a review that the average study costs ca. $1 million (with a range from $130,000 to $2.78 million). Lant Pritchett (2002, p. 267) counters this point: "(…) [S]ince evaluation costs are a tiny fraction of program costs and the potential gains are enormous it is difficult to believe this is a compelling reason for substantial areas of public intervention".

Despite these drawbacks, I made an effort in this dissertation to show that impact evaluations are useful in a variety of ways. They can test the effectiveness of a completed project, they can test and falsify theories and contribute to the creation of new theories, and through the process of falsification, expedite the trial-and-error principle that is and always has been central to policy-making. Some even argue that implementing an IE alongside a project even improves the actual project design quality and in the case of the World Bank, increases the fund disbursement speed (see Pritchett et al. 2013; Legovini et al., 2015).

Good RCTs are important as they have the potential to constitute *good research* and can provide *clues* as to how successful development programs can be designed (for example those described in this dissertation). They "do not need to be 'conclusive' in order to be *informative"* (Earp and Trafimow, 2015). Their application is a first step in incorporating empiricism into evaluation practices, which in the past were too often based on arbitrariness. RCTs are not applicable in all circumstances, but if they can be used, they should be considered (after weighing costs and potential benefits). Each development project is a rich opportunity to systematically learn through research (not just by means of RCTs), which is unfortunately not done enough today, especially by large bi- and multi-lateral donors (including Germany's GIZ). One important takeaway from two of the studies (chapters 2 and 3) for example is that while effects can be measured in the short-run, they might dissipate in the longer run.[124] This might be an indication that it can be worth letting interventions run longer, or that commitment devices should be employed to "lock" people into "more efficient" behavior (such as commitment savings accounts). The DCE presented

[124] This issue is related to a problem raised in chapter 1: Outcomes are sensitive to the *timing* of data collection. Depending on when (3, 6, 9, or 12 months for example) outcome data is collected, results might differ.

in Chapter 5 ("lab-in-the-field experiment") showed that providing frontline health workers with better equipment and facilities might incite them to stay longer at their posting. These are examples of a new theories originating from this research that could be further tested by means of future studies.

The current "hype" around RCTs and the "credibility revolution" (Angrist and Pischke, 2010) also raises the quality standards for other evaluation techniques, and it helps to focus on issues such as endogeneity of influences, selection bias, and inadequate separation of correlation and causality (Cohen and Easterly, 2010). However, there is also a risk to go too far. There is a trend in economics to dismiss descriptive, correlational, or qualitative research and to acknowledge only studies that convincingly identify causal effects as serious research. Stokes (2014) warns in this context of "radical skepticism".[125] There has indeed been a shift in the (academic) field of development economics towards demanding more and more proof to establish the independence of key outcome variables from potentially confounding factors ("impact"), which RCTs do inherently well. The shift has been so strong, it often seems that development economists are now much more [126]occupied with the statistical features of their papers and eliminating biases, trying to improve *internal validity* in order to not be accused of having no valid "identification strategy", than they are thinking about how economic development comes about (Mookherjee, 2005).[127]

One key take-away from the debate around external validity and most importantly its limits (chapter 6), is that we cannot know everything. But if there is a "free market of ideas", it might not even be necessary to be able to ascertain "what works". To some extent, the trial-and-error process weeds out bad ideas and set up new theories over (long periods of) time. This does not mean that science, and in the framework of this paper – experiments – cannot be helpful to inform policymakers. Quite the opposite: RCTs, as all well-made research, can provide insights and clues, they can falsify theories and create new ones and therefore have the potential to speed-up the often sluggish trial-and-error principle, akin to IBM's founder Thomas Watson's premise: "If you want to succeed, raise your failure rate" (cited in Rodrik, 2010). A smart policymaker would look at all the research and then make an informed decision, also using her intuition: "A new drug might do better than a placebo in an RCT, yet a physician might be entirely correct in not

[125] In this context, the recently created German Evaluation Office for Development Cooperation (DEval) should be perceived as a further building block in resolving methodological weaknesses, self-evaluations, and only partial independence (Stockmann, 2012).

[126] As a side note: Many randomistas publish research on a plethora of many different issues in short periods of time. One might wonder why this is and how deep the researchers can understand any one issue by this focus on the RCT method rather than a subject matter.

[127] This is confirmed by the author's participation in e.g. research seminars of the Development Economics Research Group at the World Bank, where the technicalities of papers are more central to the discussions than the development problem at hand.

prescribing it for a patient whose characteristics, according to the physician's theory of the disease, might lead her to suppose the drug would be harmful" (Deaton, 2010, p. 441).[128]

There is an organic element to development that RCT dogmatists may overlook. They might implicitly believe that, if we just mix the right ingredients, e.g. educate nurses, provide bed nets against malaria, build wells and well-equipped schools, and we find out "what works", "development" will follow. This line of thinking is rooted in the thought that development is mainly a "technical" problem that can be "solved", much like an equation (de Souza Leão and Eyal, 2019). This corresponds to the thinking of Jeffrey Sachs (2005) and the "big push" models earlier years and is in contrast with what Popper (1971) called "piecemeal engineering" (see section 6.4.3). I believe that functioning societies are more like metaphorical "prairies", as illustrated by the economist Russ Roberts:

> How do you build a prairie? (…) Well, we know what's in a prairie. It's a certain set of plants. But if I start with a bare patch of ground outside of O'Hare Airport in Chicago, which used to be a prairie, so I'm going to recreate that, I'm going to do very poorly. Because if I just sort of mix all the ingredients together, the right plants, because even though those are the right plants to make a prairie, I don't understand the process by which the prairie emerged. I don't understand – certain things had to be put in place first. Certain things had to come at the same time. That there's a dynamic, organic nature to a prairie that's also true of an economy and institutions. (Roberts, 2015).

Although they might think otherwise, "technocrats" do not know how to build this "prairie" as deep-rooted development problems are almost always or institutional (Acemoglu and Robinson, 2012). Institutions are complex and cannot be "built" by just putting a framework in place and some tools. Relations matter tremendously for development, even if these can hardly be measured (Eyben, 2010). Broad and Cavanagh (2006, p. 24) call this (in a rebuttal to the ideas of Jeffrey Sachs) an "ahistorical focus on technology". Furthermore:

> Rather, the history of most parts of the world suggests a more violent process of poverty creation rooted in unequal power relations and manifested through slavery, the colonial legacy of export economics, the presence of extraction industries, and the sale of natural resources by governments to the highest corporate bidders. (Broad and Cavanagh, 2006, p. 24)

[128] A related quote from Deaton and Cartwright (2018, p. 17) is the following: "If *your* physician tells *you* that she endorses evidence-based medicine, and that the drug will work for *you because* an RCT has shown that 'it works", it is time to find a physician who knows that *you* and the *average* are not the same."

If the counter-technocratic position is true, this indeed leaves more room for the *political* decision-maker to deviate from what social scientists believe is right (in addition there is a lot of herd behavior in social science as well). A clever politician might (think to) know more about the complex relationships and the fiber of her community, city, or country and might consequently judge that her situation differs from the *average,* which is the basis for the results of most RCTs. The policymaker might decide to copy a policy from a neighboring district, where it enjoys success, even if a study shows that on average the policy has no results. Thus, a good decision might mean going against the grain, or even going against a widely held belief. This is not anti-science (which is dangerous). It is imperative that policymakers take into account studies that are available. Yet, they should not be entirely based on just these (Basu, 2014).[129]

This is also the essence of a representative democracy where the people choose who make decisions for them: "Your representative owes you, not his industry only, but his judgment; and he betrays, instead of serving you, if he sacrifices it to your opinion" (Burke, 1854). As stated by the former World Bank Chief Economist Kaushik Basu: "To get to policy conclusions requires combining the findings of randomized experiments with human intuition, which being founded in evolution, has innate strengths" (Basu, 2014, p. 455). This is especially true when policymakers are not aware of the full story and quality of any given study, where the researchers (like me) had to face "the muddy realities of field applications" (Barrett and Carter, 2010, p. 522). When implementing RCTs day-by-day it becomes clear that they are by no means a method of "mechanical objectivity" (de Souza Leão and Eyal, 2019, p. 404). Usually a lot of information around how the project or study was implemented does not make it into the final papers, due to the rigid customs of scientific style.

Lastly, the study of the RCT landscape shows that the effectiveness debate currently seems to be dominated by a selected group of development economists.[130] However, this "agenda monopoly" of economics in development is by no means justified. The apparent knowledge-vacuum (mentioned in chapter 1, and the associated micro-macro-paradox, in addition to weak links to theory and a collection of far rather "random results" (Duflo, 2006) rather prove that development economics as we know it could be perceived to be spinning its wheels. In fact, the new focus on "micro-evidence" (which is becoming ever smaller with the most recent focus on mechanism experiments) might be a chance for other scientific disciplines (like political science) to make

[129] Basu (2014, p. 465) also states that, "the fact that we should use the best available evidence does not imply that policies must invariably be based on evidence."

[130] De Souza Leão and Eyal (2019) confirm that the "2nd wave" of development RCTs is dominated by economists, whereas a prior and smaller "1st wave" of development RCTs was more inclusive of other disciplines and saw very few economists contribute.

contributions. Economic thinking will remain important, but development also encompasses many other components such as education, health, conflict studies, institutions, intra-group trust, etc. Experts for these fields should be more involved in the debates around development effectiveness and should not leave the field to economists. In the technocratic worldview what's often neglected are power relations in both, developing countries and donor countries that shape the course of how development projects and policies are implemented. Economists often overlook the importance of power, special interest groups, and intra-society complexities that may hinder growth and development and are frequently too focused on the "what works" aspect. Ogden (2017) calls this the "Policy Sausage Critique" of RCTs. Knowledge gains do not automatically lead to adoption by decisionmakers – a line of thought humorously dubbed the "Cro-Magnon simple model of policy adoption" by Lant Pritchett (in Ogden, 2017, p. 136). Thus, next to re-empowering policymakers to not delegate all the thinking and decision-making to economists (even if they win Nobel Prizes), this situation is also an opportunity for political science as a discipline to contribute to the debate. Our field has innate strengths of methodological pluralism and the emphasis on the importance of power relations within nations and the international order. These may matter as much as economic factors, if not more, for development and the enduring effort for a world where all people can live in dignity.

7.1 References

Acemoglu, D., & Robinson, J. A. (2012). Why nations fail: The origins of power, prosperity, and poverty. Crown Books.

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, *24*(2), 3-30.

Barrett, Christopher B. und Michael R. Carter. 2010. The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspectives and Policy* 32: 515-548.

Basu, K. (2014). Randomisation, causality and the role of reasoned intuition. Oxford Development Studies, 42(4), 455-472.

Besley, T. (2012). Poor Choices: Poverty from the Ground Level. *Foreign Affairs*, *91*, 160.

Bold, T., Kimenyi, M., & Mwabu, G. (2013). Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education.

Broad, R., & Cavanagh, J. (2006). The hijacking of the development debate: how Friedman and Sachs got it wrong. World policy journal, 23(2), 21-30.

Burke, E. (1854). The works of the right honourable Edmund Burke (Vol. 1). Henry G. Bohn.

Cameron, D. B., Mishra, A., & Brown, A. N. (2016). The growth of impact evaluation for international development: how much have we learned?. *Journal of Development Effectiveness*, *8*(1), 1-21.

Cohen, J., & Easterly, W. (Eds.). (2010). *What Works in Development?: Thinking Big and Thinking Small*. Brookings Institution Press.

de Souza Leão, L., & Eyal, G. (2019). The rise of randomized controlled trials (RCTs) in international development in historical perspective. Theory and Society, 48(3), 383-418.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature*, 424-455.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science & Medicine, 210, 2-21.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, *6*, 621.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice*. The World Bank.

Lan, G., Blom, A., Kamalski, J., Lau, G., Baas, J., & Adil, M. (2014). A Decade of Development in Sub-Saharan African Science, Technology, Engineering and Mathematics Research. World Bank, Washington, DC.

Legovini, A., Di Maro, V., & Piza, C. (2015). Impact evaluation helps deliver development projects.

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives*, *25*(3), 17-38.

Ogden, T. N. (Ed.). (2017). *Experimental Conversations: Perspectives on Randomized Trials in Development Economics*. MIT Press.

Popper, K. (1971). The Open Society and Its Enemies. V. I, II. Princeton, New Jersey.

Pritchett, L. (2002). It pays to be ignorant: a simple political economy of rigorous program evaluation. The Journal of Policy Reform, 5(4), 251-269.

Pritchett, L., Samji, S., & Hammer, J. S. (2013). It's all about MeE: Using Structured Experiential Learning ('e') to crawl the design space.

Roberts, R. (2015, June 22). Morten Jerven on African Economic Growth. [Audio Podcast]. Retrieved from: https://www.econtalk.org/morten-jerven-on-african-economic-growth/

Rodrik, D. (2010). Diagnostics before prescription. Journal of Economic Perspectives, 24(3), 33-44.

Sachs, J. (2005). *The end of poverty: How we can make it happen in our lifetime*. Penguin UK.

Stockmann, R. (2012). Von der Idee zur Institution Institut für deutsche Entwicklungsevaluierung gegründet. *Zeitschrift für Evaluation*, *11*(1), 85.

Stokes, S. C. (2014). A defense of observational research. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, 33-57.

The Prize in Economic Sciences. (2019). NobelPrize.org. Nobel Media AB 2019. Mon. 14 Oct 2019. Retrieved from: https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/

A1. List of Publications – Liste der Einzelarbeiten

Chapter 2. The Nollywood nudge: An entertaining approach to saving

Authors: Felipe A. Dunsch, Aidan Coville, Vincenzo Di Maro, Siegfriend Zottel

Published as: World Bank Policy Research working paper; no. WPS 8920. Washington, D.C.: World Bank Group.

Link: *http://documents.worldbank.org/curated/en/468251561642192667/The-Nollywood-Nudge-An-Entertaining-Approach-to-Saving*

Chapter 3. Management, Supervision, and Health Care: A Field Experiment

Authors: Felipe A. Dunsch, David Evans, Ezinne Eze-Ajoku, Mario Macis
Available as National Bureau of Economic Research (NBER) and Institute of Labor Economics (IZA) working papers (2017).
Link to NBER paper: *http://www.nber.org/papers/w23749*
Link to IZA paper: *https://www.iza.org/publications/dp/10967/management-supervision-and-health-care-a-field-experiment*

Chapter 4. Bias in patient satisfaction surveys: a threat to measuring healthcare quality

Authors: Felipe A. Dunsch, David Evans, Mario Macis, Qiao Wang
Published in: *British Medical Journal – Global Health* – 2018; 3(2)
Link: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5898299/*

Chapter 5. Job Preferences of Frontline Health Workers in Ghana: A Discrete Choice Experiment

Authors: Felipe A. Dunsch, Edit Velenyi (World Bank Senior Health Economist)
Unpublished.

Chapter 6. Experimentation in the Social Sciences& the Problem of External Validity

Author: Felipe A. Dunsch (single-authorship)
Unpublished.

A2. Summary

In the last two decades years, experimental impact evaluations (IEs; particularly RCTs) became a very popular tool to measure the effectiveness of development projects. In 2019, Esther Duflo, Abhijit Banerjee, and Michael Kremer even won the Nobel Prize in Economics for their use of RCT, having "introduced a new approach to obtaining reliable answers about the best ways to fight global poverty". The goal of this dissertation was to conduct 4 independent RCTs to measure and evaluate the effectiveness of development projects in West Africa while at the same time assess the general value and utility of IEs in the introduction, the article 5 (chapter 6), as well as in the conclusion/the look ahead.

1. The Nollywood nudge: An entertaining approach to saving

Can edu-entertainment be an effective tool to strengthen financial inclusion? In collaboration with a local NGO (Credit Awareness) and a Microfinance Bank (Accion) we explored the short- and medium-term savings decisions of a group of micro-entrepreneurs in Lagos, Nigeria by inviting business owners to one of four randomly allocated events: (i) A movie screening of "The Story of Gold" - a Nollywood film encouraging entrepreneurs to save responsibly; (ii) an event where business owners are shown a "placebo" screening of a movie with no financial education content and offered "on-the-spot" micro savings accounts through Accion; (iii) a combined event, screening The Story of Gold and offering "on-the-spot" accounts; and (iv) a screening of the placebo film only as our control group. We find that entrepreneurs watching The Story of Gold were 5 percentage points more likely to open a savings account on the spot than those in placebo screenings, and this effect was mostly driven by male business owners. In contrast, less than 1% of entrepreneurs who were not offered "on-the-spot" signed up for a savings account after the screening. In the longer run, only moderate changes in attitudes and perceptions were found, while savings and borrowing behavior was un-changed four months after the screening. This suggests that, while influencing short-term decisions is possible, longer-run behavior is far less malleable through once-off events. This paper contributes to the literature by directly testing the importance of linking emotional stimulus to financial messages to influence short-term savings decisions and identifying the important interaction be-tween emotional stimulus and the opportunity to act on this stimulus.

This article was published as a World Bank Policy Research Working Paper: *http://documents.worldbank.org/curated/en/468251561642192667/The-Nollywood-Nudge-An-Entertaining-Approach-to-Saving*

2. Management, Supervision, and Health Care: A Field Experiment

If health service delivery is poorly managed, then increases in inputs or ability may not translate into gains in quality. However, little is known about how to increase managerial capital to generate persistent improvements in quality. We present results from a randomized field experiment in 80 primary health care centers (PHCs) in Nigeria to evaluate the effects of a health care management consulting intervention. One set of PHCs received a detailed improvement plan and nine months of implementation support (full intervention), another set received only a general training session, an overall assessment and a report with improvement advice (light intervention), and a third set of facilities served as a control group. In the short term, the full intervention had large and significant effects on the adoption of several practices under the direct control of the PHC staff, as well as some intermediate outcomes. Virtually no effects remained one year after the intervention concluded. The light intervention showed no consistent effects at either point. We conclude that sustained supervision is crucial for achieving persistent improvements in contexts where the lack of external competition fails to create incentives for the adoption of effective managerial practices.

This article was published as a working paper of the National Bureau of Economic Research (NBER) and the Institute of Labor Economics (IZA):

IZA: https://www.iza.org/publications/dp/10967/management-supervision-and-health-care-a-field-experiment

NBER: https://www.nber.org/papers/w23749

3. Bias in patient satisfaction surveys: a threat to measuring health care quality

Patient satisfaction surveys are an increasingly common element of efforts to evaluate the quality of health care. Many patient satisfaction surveys in developing countries frame statements positively and invite patients to agree or disagree, so that positive responses may reflect either true satisfaction or bias induced by the positive framing. In an experiment with more than 2,200 patients in Nigeria, we distinguish between actual satisfaction and survey biases. Patients randomly assigned to receive negatively framed statements expressed significantly lower levels of satisfaction (87 percent) than patients receiving the standard positively framed statements (95 percent). Depending on the question, the effect is as high as a 19-percentage point drop. Thus, high reported patient satisfaction likely overstates the quality of health services. Providers and policy makers wishing to gauge the quality of care will need to avoid framing that induces bias and to complement patient satisfaction measures with more objective measures of quality.

This article was published in the British Medical Journal (BMJ) – Global Health:

https://www.ncbi.nlm.nih.gov/pubmed/29662696

4. Job Preferences of Frontline Health Workers in Ghana: A Discrete Choice Experiment

The lack of supply of adequately skilled and motivated health workers especially in rural areas poses a major obstacle to health service delivery in Ghana, as well as in many other developing countries. In this paper, we present the results of a discrete choice experiment (DCE) conducted with community health officers (CHO) and community health volunteers (CHV), which are extension workers that provide health services and consultations to mostly rural populations.

CHOs and CHVs completed the cadre-specific discrete choice experiment that elicited preferences for attributes of potential job postings. Data was collected from 404 CHOs and 206 CHVs in 8 Ghanaian districts in 4 regions. For CHOs, next to increases in salary, the choice of job posting was most strongly influenced by facility quality, followed by career development opportunities and transport subsidies. Additional supervision showed now effects. For CHVs, next to receiving a monthly stipend, facility quality was also most important, followed by training opportunities.

We are corroborating the notion that other non-financial incentives can have strong effects on job preferences, for example the equipment of the facilities, which includes housing, as well as training and career development and training opportunities. Given that Ghana's health wage bill is already very high, this may open new policy avenues for health workers recruitment and retention to converge towards the aspired universal health coverage.

5. Experimentation in the Social Sciences & the Problem of External Validity

The quantity of impact evaluations (IEs) in the field of development economics utilizing experimental or quasi-experimental methods has picked up considerably over the past 2 decades. This is partly because impact evaluations are being welcomed by governments and researchers as a tool to evaluate the effectiveness of their policies and to provide policy guidance for future decisions. However, IEs struggle with "external validity", i.e. the generalizability of results to other contexts. I argue that the problem of context variance is an overlooked aspect for lack of external validity of IE results as this age-old principle has been neglected by economists and policymakers alike. No matter how many instances are observed, it is theoretically not possible to verify that an effect holds over time, and much less when applied in another context. However, IEs can still be a valuable tool for policymakers. Instead of seeking to find "what works", RCTs should be used to try to falsify explicitly stated theories – which is rare in current practice. In addition, experiments might be better suited to identify commonalities across context by testing more contained "mechanisms" rather than large policies or programs. Through rapid falsification, the trial-and-error process of policymaking can be accelerated, and potentially better working policies can be implemented faster, even if RCTs cannot tell us conclusively "what works".

Conclusion

Conducting good RCTs is very important as they can provide clues as to how successful development programs can be designed (for example those described in this dissertation) – even if they don't have the power to tell us "what works". Their application is a first step in incorporating empiricism into evaluation practices, which in the past were too often based on arbitrariness. RCTs are not applicable in all circumstances, but if they can be used, they should be considered. Each development project is a rich opportunity to systematically learn through research, which, despite many opportunities, is unfortunately not done enough today. Each of the empirical case studies are stand-alone research products. Policymakers should take these research results into consideration but must also understand that past results cannot provide certainty for future outcomes (due to the lack of external validity). Results from social science cannot lift the burden of independent *political* decision-making from policymakers – even if scientists might claim that they can (in line with the "what works" agenda).

A3. Zusammenfassung

In den letzten zwei Jahrzehnten ist eine rasante Zunahme der Nutzung sogenannter Impact Evaluations (IEs) – darunter vor allem Randomized Controlled Trials (RCTs) – zur Messung der Effektivität von Projekten im Rahmen der Entwicklungsarbeit zu verzeichnen. 2019 haben Esther Duflo, Abhijit Banerjee, and Michael Kremer sogar den Nobelpreis in den Wirtschaftswissenschaften für die Verbreitung der RCT Methode gewonnen, mit der Begründung, dass sie „einen neuen Ansatz eingeführt haben, um verlässliche Antworten auf die Probleme der globalen Armutsbekämpfung zu finden." Das Ziel dieser Dissertation war einerseits die Durchführung von 4 eigenständigen RCTs zur wissenschaftlichen Messung und Evaluierung der Effektivität verschiedener Projekte in Westafrika und gleichzeitig die Beurteilung der Nützlichkeit dieser Methode in der Einleitung, dem 5. Artikel (Kapitel 6), sowie in der Zusammenfassung/dem Ausblick.

1. The Nollywood nudge: An entertaining approach to saving

Kann "Edu-Entertainment" ein effektives Mittel zur Stärkung der finanziellen Inklusion in Nigeria sein? In Zusammenarbeit mit einer lokalen Nichtregierungsorganisation und einer Mikrofinanzbank wurde in dieser Studie das kurz- und mittelfristige Sparverhalten einer Gruppe von Kleinunternehmern in Lagos, Nigeria, untersucht. Dazu wurden sie per Zufallsauswahl zu einer von 4 Veranstaltungen eingeladen:

- Eine Filmvorführung des Filmes „Story of Gold", ein Nollywood Spielfilm, der neben dem unterhaltenden Effekt auch zum verantwortungsvollen Sparen anrät.
- Eine Vorführung eines „Placebo-Films" ohne Informationen zum Sparverhalten.
- Eine Vorführung des Filmes Story of Gold mit anschließender optionaler Beratung einer Mikrofinanzbank bei der die Teilnehmer auch ein Konto eröffnen konnten.
- Eine Vorführung des Placebofilms mit anschließender optionaler Beratung der Mikrofinanzbank.

Die Ergebnisse zeigen, dass sich für die Unternehmer, die den Film „Story of Gold" gesehen haben, die Wahrscheinlichkeit, dass sie direkt nach der Veranstaltung ein Sparkonto eröffnen, um 5 Prozentpunkte erhöht (im Vergleich zur Placebo Gruppe). Dieser Effekt wird hauptsächlich durch die männlichen Teilnehmer erklärt bzw. hervorgerufen. Weniger als 1% der Unternehmer, denen kein Sparkonto direkt nach der Veranstaltung angeboten wurde, entschied sich noch später eines zu eröffnen. Längerfristig sind nur moderate Veränderungen in den Einstellungen und Wahrnehmungen im Bezug auf das Spar- und Leihverhalten der Teilnehmer zu verzeichnen und das tatsächliche Verhalten der Unternehmer war nach 4 Monaten nach den Vorstellungen unverändert und ohne messbare Unterschiede zwischen den Gruppen. Diese Ergebnisse suggerieren, dass zwar kurzfristig das Verhalten beeinflusst werden kann, aber dass dies beim langfristigen

Verhalten schwieriger ist, vor allem mit einer nur einmaligen Veranstaltung. Dieser Artikel ist ein Beitrag zur wissenschaftlichen (verhaltensökonomischen) Debatte, die sich mit der Verbdingung von emotionalen Stimuli und daraus folgenden Handlungsmustern beschäftigt („nudging for public policy“).

Dieser Artikel wurde in der World Bank Policy Research Working Paper Series veröffentlicht: *http://documents.worldbank.org/curated/en/468251561642192667/The-Nollywood-Nudge-An-Entertaining-Approach-to-Saving*

2. Management, Supervision, and Health Care: A Field Experiment

Wenn Gesundheitsdienstleistungen schlecht *organisiert* sind, dann reichen Verbesserungen der Infrastruktur oder des Wissens der medizinischen Fachkräfte oft nicht aus, um die Qualität der Gesundheitsversorgung messbar zu erhöhen. Man weiß jedoch bisher nur kaum, wie die Qualität des Managements erhöht werden kann, um anhaltende Qualitätsverbesserungen im Gesundheitsbereich zu erzielen. Mit dieser Studie werden die Ergebnisse eines randomisierten Feldversuchs mit 80 medizinischen Grundversorgungszentren (PHCs) in Nigeria präsentiert , um den Einfluss eines Projekts zur Verbesserung der Qualität des Managements von PHCs zu messen. Eine Gruppe der 80 PHCs erhielt einen detaillierten Verbesserungsplan und neun Monate Implementierungsunterstützung („vollständiges Projekt“), eine andere Gruppe erhielt nur eine allgemeine Schulung, eine Gesamtbewertung des Status-Quo und einen Bericht mit Verbesserungsempfehlungen („Light-Projekt“) und eine dritte Gruppe diente als Kontrollgruppe. In der kurzen Frist hatte das vollständige Projekt große und bedeutende Auswirkungen auf die Übernahme mehrerer Maßnahmen, welche innerhalb der Kontrolle des PHC-Personals waren. Ein Jahr nach dem Abschluss des Projekts blieben jedoch praktisch keine Effekte mehr bestehen. Das „Light-Projekt“ zeigte zu beiden Zeitpunkten keine konsistenten Effekte. Die Studie kommt zu dem Schluss, dass eine nachhaltige Supervision von entscheidender Bedeutung sein kann, um dauerhafte Verbesserungen in Kontexten zu erzielen, in denen mangelnder externer Wettbewerb keine Anreize für die Einführung wirksamer Managementpraktiken schafft.

Dieser Artikel wurde als Working des National Bureau of Economic Research (NBER) und des Institute of Labor Economics (IZA) veroeffentlicht:

IZA: https://www.iza.org/publications/dp/10967/management-supervision-and-health-care-a-field-experiment

NBER: https://www.nber.org/papers/w23749

3. Bias in patient satisfaction surveys: a threat to measuring health care quality

Patientenzufriedenheitsbefragungen sind ein zunehmend angewandtes Mittel, um die Qualität von Gesundheitsversorgung zu evaluieren. Viele dieser Befragungen in Entwicklungsländern formulieren diese Fragen affirmativ/positiv und Patienten sind dazu angehalten zuzustimmen oder

ablehnend zu antworten, sodass positive Antworten wahre Zustimmung signalisieren oder ein Verzerrungseffekt sein könnten, der durch das positive „Framing“ der Fragen entsteht. In einem Experiment mit mehr als 2200 Patienten in Nigeria quantifiziert diese Studie diesen Verzerrungseffekt. Patienten, die nach dem Zufallsprinzip für negativ „geframten“ Aussagen ausgewählt wurden, zeigten eine signifikant geringere Zufriedenheit (87 Prozent) als Patienten, die die positiv Standardaussagen (95 Prozent) erhielten. Je nach Frage erreicht der Effekt bis zu 19 Prozentpunkte. Die Studie schlussfolgert, dass eine hohe Zufriedenheit der Patienten daher wahrscheinlich überbewertet ist und damit auch die so gemessene Qualität der Gesundheitsdienstleistungen. Versorger und politische Entscheidungsträger, die die Qualität der Versorgung beurteilen möchten, müssen diese Verzerrungen vermeiden und Befragungen zur Patientenzufriedenheit durch objektivere Qualitätsmaßnahmen ergänzen.
Dieser Artikel wurde im British Medical Journal (BMJ) – Global Health veroeffentlicht:
https://www.ncbi.nlm.nih.gov/pubmed/29662696

4. Job Preferences of Frontline Health Workers in Ghana: A Discrete Choice Experiment

Der Mangel an ausreichend qualifizierten und motivierten Gesundheitsfachkräften, insbesondere in ländlichen Gebieten, stellt ein großes Hindernis für die Erbringung von Gesundheitsdiensten in Ghana und in vielen anderen Entwicklungsländern dar. Dieser Artikel präsentiert die Ergebnisse eines „Discrete Choice Experiments“ (DCE), das mit „community health officers“ (CHO) und Freiwilligen (CHV) durchgeführt wurde. Diese Kader versorgen die vorwiegend ländlichen Bevölkerungsgruppen mit Gesundheitsdiensten und Konsultationen. Das DCE testete die Präferenzen für Attribute potenzieller Stellenausschreibungen. Die Daten wurden von 404 CHOs und 206 CHVs in 8 ghanaischen Bezirken in 4 Regionen gesammelt. Für CHOs war neben Gehaltserhöhungen die Präferenz des Stellenangebots am stärksten von der Qualität der Einrichtungen abhängig, gefolgt von Karriereentwicklungsmöglichkeiten und Transportsubventionen. Zusätzliche Supervision zeigte keine Effekte. Für CHVs war neben dem monatlichen Stipendium vor allem die Qualität der Infrastruktur von Bedeutung, gefolgt von Weiterbildungsmöglichkeiten. Die Studie bekräftigt die Auffassung, dass andere, nichtfinanzielle Anreize, starke Auswirkungen auf die Berufspräferenzen haben können, z. B. die Ausstattung der Einrichtungen, einschließlich eines Wohnraums sowie Ausbildungs- und Karriereentwicklungsmöglichkeiten. Angesichts der Tatsache, dass die Löhne im Gesundheitssektor in Ghana bereits sehr hoch sind (relativ), könnten diese Ergebnisse neue Möglichkeiten für die Rekrutierung und die verlässliche und dauerhafte Anstellung von Fachkräften eröffnen und somit einen Beitrag für eine verbesserte generelle Gesundheitsversorgung im Land leisten.

5. Experimentation in the Social Sciences & the Problem of External Validity

Die Anzahl der Impact Evaluations (IEs) im Bereich der Entwicklungsökonomie (experimentelle oder quasi-experimentellen Methoden) hat in den letzten zwei Jahrzehnten erheblich zugenommen. Dies ist teilweise darauf zurückzuführen, dass die Regierungen und Forscher IEs als Instrument zur Bewertung der Wirksamkeit ihrer Politik und Projekte vermehrt begrüßen. IEs kämpfen jedoch mit Problemen im Bezug auf "externe Validität", d. H. Der Generalisierbarkeit von Ergebnissen für andere Kontexte. Ich stelle fest, dass das Problem der „Kontextinvarianz“ ein übersehener Aspekt ist. Das uralte „Induktionsproblem“ wird in der Debatte um die Nützlichkeit von IEs bisher häufig übersehen bzw. vernachlässigt: Unabhängig von der Anzahl der beobachteten Fälle kann theoretisch nicht verifiziert werden, ob ein Effekt über einen längeren Zeitraum hält, und noch viel weniger, wenn ein Projekt in einem anderen Kontext angewendet wird. Trotz dieser Einschränkung können IEs dennoch ein wertvolles Instrument für politische Entscheidungsträger sein. Anstatt zu versuchen herauszufinden „was funktioniert“ („what works“), sollten RCTs dazu genutzt werden zu versuchen explizit formulierte Theorien zu falsifizieren. Zusätzlich wären RCTs besser geeignet „Mechanismen“ zu untersuchen, die aufgrund ihres geringeren Fokus evtl. höheres Potential haben eine gewisse Allgemeingültigkeit über Kontexte hinweg zu erreichen, anstelle großer Projekte mit vielen Komponenten. Durch schnellere Falsifizierung kann der Prozess des „Versuchs und Irrtums“ beschleunigt werden und potenziell bessere Programme können schneller eingesetzt werden, auch wenn uns RCTs nicht direkt wissen lassen, „was funktioniert.“

Zusammenfassung und Ausblick

Die Durchführung guter RCTs ist sehr wertvoll und wichtig, da sie *Hinweise* darauf geben können, wie erfolgreiche Entwicklungsprogramme gestaltet werden können. Ihre Anwendung ist ein erster Schritt, um Empirismus in Evaluationspraktiken zu integrieren, die in der Vergangenheit zu oft auf Willkür beruhten. RCTs sind nicht unter allen Umständen anwendbar, aber wenn sie verwendet werden können, sollten sie berücksichtigt werden. Jedes Entwicklungsprojekt ist eine Gelegenheit, durch Forschung systematisch zu lernen, was heutzutage in Rahmen der Entwicklungspolitik leider nicht genug getan wird. Die empirischen Fallstudien dieser Arbeit sind jeweils eigenständige Forschungsprojekte. Die politischen Entscheidungsträger sollten diese Ergebnisse berücksichtigen, müssen jedoch auch verstehen, dass einzelne Studien (aufgrund der mangelnden externen Validität) keine Garantien für zukünftige Ergebnisse bieten können. Den politischen Entscheidungsträgern bleibt somit die Verantwortung, unabhängige und eben *politische* Entscheidungen zu treffen, die (Sozial-)Wissenschaft allein ihnen nicht abnehmen kann, selbst wenn Wissenschaftler dies vermehrt für möglich halten (im Einklang mit der „what works“ Agenda).

A4. Erklärung

Hiermit erkläre ich, Felipe Alexander Dunsch, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die Arbeit wurde nicht schon einmal in einem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

Hamburg, 20.10.2019

Ort/Datum Unterschrift Doktorand/in

****

A5. Eidesstattliche Versicherung:

Ich, Felipe Alexander Dunsch, versichere an Eides statt, dass ich die Dissertation mit dem Titel:

„Randomized Controlled Trials in West Africa – Practice and Theory“

selbst und bei einer Zusammenarbeit mit anderen Wissenschaftlerinnen oder Wissenschaftlern gemäß den beigefügten Darlegungen nach § 6 Abs. 3 der Promotionsordnung der Fakultät für Wirtschafts- und Sozialwissenschaften vom 24. August 2010 verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht benutzt.

Hamburg, 20.10.2019

Ort/Datum

Unterschrift Doktorand/in

Unterschrift Verwaltung

A6. Selbstdeklaration

Für Kapitel 6: „Experimentation in the Social Sciences & the Problem of External Validity” liegt die Eigenleistung bei 100%

Für Kapitel 2: „The Nollywood nudge: An entertaining approach to saving“ liegt die Eigenleistung für

das Konzept / die Planung bei	50%
die Durchführung bei	75%
der Manuskripterstellung bei	40%

Für Kapitel 3: „Management, Supervision, and Health Care: A Field Experiment“ liegt die Eigenleistung bei

das Konzept / die Planung bei	50%
die Durchführung bei	80%
der Manuskripterstellung bei	35%

Für Kapitel 4: “Bias in patient satisfaction surveys: a threat to measuring health care quality” liegt die Eigenleistung bei

das Konzept / die Planung bei	30%
die Durchführung bei	80%
der Manuskripterstellung bei	25%

Für Kapitel 5: „Job Preferences of Frontline Health Workers in Ghana: A Discrete Choice Experiment“ liegt die Eigenleistung bei

das Konzept / die Planung bei	80%
die Durchführung bei	90%
der Manuskripterstellung bei	95%

Folgend ergänze ich (in englischer Sprach) die obigen Informationen mit weiteren Erklärungen, die mit den jeweiligen Koautoren abgestimmt sind:

This work started in 2012. I have been a co-principal investigator for articles/chapters 2-5. Chapter 6 was written in single authorship. I have made ca. 10 field trips to conduct the field work necessary for chapters 2-5 to multiple locations in Nigeria and Ghana.

Chapter 2

For the article “The Nollywood Nudge, and Entertaining Approach to Saving” I worked under the leadership of the Principle Investigators Aidan Coville and Vincenzo Di Maro.

I independently lead the implementation of the extensive endline survey in Lagos, Nigeria, which included being the main liaison between the local field coordinators and the Principle Investigators (during mission travel to Lagos, Nigeria). I also contributed significantly to the design and programmed the endline survey for tablet computers (which resulted in the first data collection using tablets instead of pen-and-paper for the World Bank’s Development Impact Evaluation (DIME) Unit), and was in charge of ensuring the collection of high-quality data (supervising and training a team of local enumerators). I also liaised extensively with the local NGO that implemented the project that was the subject of the study during this time.

I contributed significantly to the write-up of the paper, especially the sections on the background, literature base, motivation of the study and the details about the field work.

I independently presented the draft paper at the Oxford University Centre for the Study of African Economies (CSAE) Conference in 2015.

Chapters 3 and 4

The principal investigators for the “Management, Supervision, and Health Care: A Field Experiment” (chapter 3) and “Bias in satisfaction surveys: a threat to measuring health care quality” (chapter 4) articles were David Evans and Mario Macis.

As a Co-Principal Investigator, I contributed to the project design from the beginning, starting with a research design workshop in Nigeria in 2013 where the team was formed. Over a span of 3-4 years, I served as the research project manager, which included the hiring and daily supervision of a local field coordinator, a survey firm, research assistant, and the procurement of necessary equipment of the study.

This work also included regularly (often weekly) liaison with representatives of the Government of Nigeria in order to ensure the research design was being adhered to, as well as regular mission travel to discuss with our partners in-country. I also led the survey programming (on electronic tablet computers) and training of local enumerators together with the hired survey firm. In the role of research project manager, I also served as the connector between the principal investigators and the other team members and Nigerian counterparts.

At all stages I led and contributed in large parts to the strategic decisions of the project (study & questionnaire design, study execution/project management, analysis & write-up, and dissemination). Dissemination efforts included (among others) a presentation at the What Works Global Summit (2017) in London. I also had a lead role in procuring funding from the Bill & Melinda Gates Foundation, which was of central importance to enable this work.

For both papers I am the first author.

Chapter 5

For the article, "Job Preferences of Frontline Health Workers in Ghana" I served as the Co-Principle Investigator (and I am first author) alongside Dr. Edit Velenyi.

Dr. Velenyi had the initial study idea but then trusted me to lead large parts of the design and execution of the research project independently. This included leading on the study design, designing and programming the questionnaire, supervision of the local field coordinator and a research assistant, as well as multiple field trips to discuss the project with the Government and to interact with World Bank staff in Ghana to ensure the project can be executed as planned. I also supervised and trained the enumerators of a firm that was commissioned to collect the data.

I was independently responsible for the cleaning and econometric analysis of the data, as well as for the full write-up of the final paper draft.

I also presented the results to the Government of Ghana in a workshop (at which the study results were received with great interest).

Chapter 6

Article 6, "Experimentation in the Social Sciences & the Problem of External Validity" was written in single-authorship.

Die vorliegende Einschätzung über die von mir erbrachte Eigenleistung wurde mit den am Artikel beteiligten Koautoren einvernehmlich abgestimmt.

Hamburg, 11.11.2019