On the Analogy between Field Experiments in Economics and Clinical Trials in Medicine

Judith Favereau

On the Analogy between Field Experiments in Economics and Clinical Trials in Medicine

Judith Favereau

Á paraître en juin 2016 dans le *Journal of Economic Methodology*

Abstract

Randomized experiments, as developed by Esther Duflo and Abhijit Banerjee at the Abdul Latif Jameel Poverty Action Lab (J-PAL), offer a novel, evidence-based approach to fighting poverty. This approach is original in that it imports the methodology of clinical trials for application in development economics. This paper examines the analogy between J-PAL's field experiments in development economics and randomized controlled trials in medicine. Randomized controlled trials (RCTs) and randomized field experiments (RFEs) are commonly treated as identical, but such treatment neglects some of the major distinguishing features that make each experiment specifically apt for use in its respective field. The central claim of this paper is that the analogy between medicine and development economics is incomplete because the central dimensions of RCTs are not simply different, but altogether lacking in J-PAL's approach. This weakens both the political and the theoretical power of such experiments in development economics.

Keywords: Randomized Experiments; Medical Clinical Trials; Philosophy of Economics; Development Economics; Interdisciplinary Transfers; Analogy

JEL Classifications: A120; B400; O200

1 Introduction

"In the 20th century, clinical trials have revolutionized medical practice. Unfortunately, the same observation does not apply to policies concerning education and health. Often, these policies are not subject to rigorous evaluation before being generalized (...) It is, however, possible to draw inspiration from clinical trials to conduct evaluations of pilot programs in terms of education and health.[1]" (Duflo, 2010, p. 17)

"These changes will be incremental, but they will sustain and build on themselves. They can be the start of a quiet revolution." (Banerjee & Duflo, 2011, p. 265)

In the last decade, randomized experiments (an experimental design centered on a random assignment of treatments) have achieved notable success, marked by a large increase in their use in economics and in diverse areas of the social sciences. Fisher (1935, 1926) originally designed this experimental procedure to test agricultural soils and fertilizers. However, this procedure achieved major success for its application not in the field of agriculture, but in medicine, as a result of the "expansion" of randomized controlled trials (RCTs) during the 1950s. More recently, randomized experiments have proven exceptionally—and increasingly—useful to researchers in development economics at the Abdul Latif Jameel Poverty Action Lab (J-PAL). Three economists—Abhijit Banerjee, Esther Duflo, and Sendhil Mullainathan— at the Massachusetts Institute of Technology (MIT) founded this laboratory with the aim of adapting an evidence-based approach to the fight against poverty. The idea was to produce reliable data on the efficacy of development programs and then to use these data to guide policy-makers. From this perspective, randomized experiments appear to be the preferred tool for producing these data because they offer strong internal validity. In the researchers' own words, their strategy essentially aimed "to mimic randomized trials that are used in medicine to evaluate the effectiveness of new drugs" (Banerjee & Duflo, 2011, p. 8) within development economics. This impressive use of randomized field experiments (RFEs) has compelled Angrist and Pischke (2010) to speak of an "empirical revolution." In addition to this "empirical revolution," which follows from the methodological revolution that occurred in medicine[2], Banerjee and Duflo (2011, p. 265) envisage a "quiet revolution," one which stands to substantially improve the life of the poor.

Clinical trials in medicine and field experiments in economics are usually considered analogous. In fact, many philosophers of science (Cartwright, 2011, 2010, 2009, 2007; Reiss & Teira, 2013; Teira, 2013, 2011)[3], physicians (Rothwell, 2005), and even development economists (Deaton, 2010; Rodrik, 2009; Barett & Carter, 2010; Ravallion, 2009a, 2009b; Harrison, 2011a, 2011b) do not distinguish between these two types of experiments in their analyses. Although RFEs and RCTs share many identical features, they also possess a number of distinguishing ones. Examining the differences between these two types of experiments can highlight many of the challenges faced in both domains. This paper examines the analogy between randomized experiments in medicine and development economics by exploring the similarities and differences in the transfer of this experimental design between these two fields. Contrary to the notion of identity, analogy[4], which identifies similarities and differences, allows one to scrutinize experiments made in medicine and development economics according to their disciplinary transfers.

The central claim of this paper is that an analogy between medicine and development economics is ultimately incomplete because the central dimensions of RCTs are not simply different from those of RFEs, but lacking alto-

gether. This lack, therefore, weakens both the political and the explanatory powers of RFEs in development economics. Indeed, although RCTs have worked reasonably well in medicine, they faced serious objections regarding their ability to measure the safety and efficacy of treatments.[5] Compared to RCTs, RFEs seem incomplete from two standpoints: elements are missing from the experimental design and their outcome is interpreted in a markedly different context. Consequently, objections to RCTs in medicine appear more serious in economics. To demonstrate this, the first section presents the common features of randomized experiments in medicine and development economics by mapping their one-to-one correspondences, as Gentner (1983) suggests. Doing so allows one to define precisely what these experiments are, thus breaking with any confusing definitions of this methodology. The second section illustrates the distinguishing features of these two uses of randomized experiments by mapping their one-to-one incongruities and returning to the domain in which these experiments are embedded. This study shows that RCTs are just the third stage in a long, four-phase research process, while RFEs in development economics are self-standing experiments. I contend that the three phases in RCTs that are missing in RFEs contribute crucial background and confirmatory knowledge that a proper assessment of the effects of a social intervention cannot dispense. I examine the implications of this incompleteness for research in development economics and show that although J-PAL's approach clearly contributes to an "empirical revolution," the approach must still complete the analogy by explicitly formulating prior theoretical frames in order to provide both the explanatory and political insights that would finally lead to a complete revolution.

2 Analogy: Mapping Clinical Trials in Medicine and Field Experiments in Development Economics

For at least two centuries,[6] researchers have introduced random allocations into experimental designs and statistical estimations. The addition of a random dimension presumably guarantees reliable assessments. However, introducing a random dimension is insufficient to define precisely what randomized controlled trials or randomized experiments entail. This section underscores the need to map the analogy between RCTs and RFEs in order to clarify the similarities and differences between how each field employs the experimental design. Identifying the features common to both RCTs and RFEs relates to the first step in the use of analogical reasoning. For that reason, I first present two case studies: a clinical trial in the field of medicine and a randomized experiment in development economics (2.1.). I then use these two experiments to map their common features (2.2.), which will serve as the core of a definition of a randomized experiment.

2.1 Rosuvastatin and Cardiovascular Events *versus* Pratham and Primary Education

The "**J**ustification for the **U**se of statins in **P**revention: an **I**ntervention **T**rial **E**valuating **R**osuvastatins" (JUPITER) (Ridker *et al.*, 2008) is an emblematic clinical trial in medicine. The JUPITER trial is highly representative in that it includes the essential methodological and practical elements of a medical trial. The objective of this trial was to determine whether the administration of 20 mg of rosuvastatin daily would prevent major cardiovascular events, the leading cause of death in the United States (Heron *et al.*, 2009). Such events correlate positively with a high level of low-density lipoprotein (LDL) cholesterol.[7] To prevent cardiovascular events, current treatments recommend statins[8] for individuals with high levels of LDL . However, Ridker *et al.* (2008) maintain that many cardiovascular events (e.g., strokes or myocardional infarctions) occur among apparently healthy men and women with stable levels of LDL cholesterol. In addition to high levels of LDL, the presence of high-sensitivity C-reactive protein (CRP) is a biomarker of future cardiovascular events (Ridker, Cushman, Stampfer, Tracy & Hennekens 1997; Ballantyne *et al.*, 2004). Statins can also lower CRP levels (Ridker, Rifai, Pfeffer, Sacks & Braunwald, 1999; Albert, Danielson, Rifai & Ridker, 2001). Thus, Ridker *et al.* (2008) hypothesize that treating seemingly healthy individuals, men and women with "levels of LDL cholesterol below the current treatment thresholds but with elevated levels of high-sensitivity C-reactive protein" (Ridker *et al.*, 2008, p. 2196) with 20 mg of rosuvastatin daily will protect them from cardiovascular events. The JUPITER trial tests this hypothesis by producing data about the benefits of rosuvastatins in healthy individuals. The trial's sample comprised 17 802 men (aged 50 years or older) and women (aged 60 years or older). Half of the sample was assigned to the treatment group, which received 20 mg of rosuvastatin daily, and the other half to the control group, which received a placebo daily. The trial was implemented at 1,315 sites in 26 countries. The first patient entered the trial in February 2003, and the last in December 2006; the trial ended in March 2008. The clinical endpoint (the target outcome of the trial) was the occurrence of a cardiovascular event. The trial had 520 confirmed endpoints: 142 cardiovascular events in the treatment group, and 251 events in the control group. Previous clinical trials of statins other than rosuvastatin—which included patients with high levels of LDL cholesterol—showed a 20% decrease in vascular risk, whereas rosuvastatin in the JUPITER trial reduced this risk by as much as 25%. As a consequence, the trial concluded that rosuvastatin prevents cardiovascular events and even other causes of death:

> In this randomized trial of apparently healthy men and women with elevated levels of high-sensitivity C-reactive protein, rosuvastatin significantly reduced the incidence of major cardiovas-

> cular events, despite the fact that nearly all study participants had lipid levels at baseline that were well below the threshold for treatment according to current prevention guidelines. Rosuvastatin also significantly reduced the incidence of death from any cause. (Ridker *et al.*, 2008, p. 2202)

Now, let me briefly describe the randomized experiment conducted in development economics before proceeding to analogically map the features (similarities and differences) of the two experiments. Abhijit Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden conducted this experiment in India from 2001 to 2004 in collaboration with the NGO Pratham[9] (see, Banerjee *et al.*, 2007; and J-PAL, Policy Briefcase, 2006). The program, titled *Balsakhi* and developed by Pratham, aimed to hire young women who had already finished secondary school to work with children who had not yet learned to read or write. As in the medical experiment, this particular experiment was not randomly selected. I chose this RFE for a number of reasons: first, Pratham is the largest NGO to collaborate with J-PAL since the latter's creation in 2003; second, educational programs are among J-PAL's main concerns and represent a major issue affecting many Indian cities, thus providing a large field for these experiments. Lastly, two of J-PAL's leaders, Banerjee and Duflo, initiated this experiment, thus making it significantly representative of J-PAL's approach. In this experiment, young women were hired to work with 15 or 20 students for two hours each week. This program had already been implemented in Mumbai in 1998 and was later implemented in Vadodara in 1999. In the first stage (from 2001 to 2002), the experimenters aimed to evaluate the impact of the *Balsakhi* program in Vadadora. Of the 122 schools in this city, 58 participated in the experiment; these 58 schools were then divided into two groups: group A (students in degree three[10]) and group B (students in degree four). Two groups were randomly selected from group A: one would serve as the treatment group, and the other as the control group. The same random assignments were carried out for group B. In the second stage, from 2002 to 2003, the assignments were reversed: students from group A who *had not* received the benefits of the *Balsakhi* program now *did* receive benefits, while those who previously received benefits no longer did; the same inversion was carried out for group B. The following year, from 2003 to 2004, none of the groups enjoyed the benefits of the program. The results of the experiment were considered highly positive:

> The Balsakhi program appears to be successful: in all years, for both subjects, and in both cities, and for all subgroups, the difference in post-test scores between treatment and comparison groups is positive and, in most instances significant. (Banerjee *et al.*, 2007, p. 10)

2.2 Mapping the One-to-One Correspondences: Toward a Precise Definition of Randomized Experiment

What do these two experiments share in common? To provide a rough answer, one might first consider that both involve randomized allocations of treatments. What does it mean for an experiment to involve a randomized allocation of treatment? Mapping all the one-to-one correspondences produces the following table (see Table 1).

Both experiments aim to (1) compare two states—one which includes the treatment (or the program) and one which does not—in order to identify a proper causal effect of the treatment. In the case of the JUPITER trial, the goal is to compare receiving 20 mg of rosuvastatin daily with receiving a placebo in order to determine the usefulness of rosuvastatin in preventing cardiovascular events; in the case of Banerjee *et al.* (2007), the goal is to compare the presence of the *Balsakhi* program with its absence in order to determine the program's benefits on the writing and reading skills of the children involved. One can determine a proper causal effect by (2) *randomly* assigning the treatment. The random assignment is defined by (3) the creation of at least two groups. These two dimensions define the randomized part of the experiment and are considered by Fisher to be the "essential safeguard" (Fisher, 1935, p. 19) for obtaining reliable statistical results. In addition, both experiments have been (4) replicated; the JUPITER trial was the first to assess the impact of rosuvastatins on the prevention of cardiovascular events. Since this first attempt, many trials, such as the AURORA, the ASTROMER, the METEOR or the ASTEROID trials, have assessed the effects of rosuvastatins on cardiovascular events.[11] The RFE in Vadodara is itself a replication, since the *Balsakhi* program was first assessed in Mumbai. Both experiments share (5) orthogonality, because both are characterized by random assignment. Furthermore, since both experiments aimed to test only one effect (rosuvastatin and the *Balsakhi* program, respectively), (6) no factorial plan was necessary. According to Fisher (1935, 1926), these six characteristics allow us to define the first element of a randomized experiment: its experimental design.[12] Both experiments share this experimental design.

For Fisher (1935, 1926), the crucial aspect of this specific experimental design is the statistical rigour that it offers, especially over other experimental designs. To obtain a reliable statistical result, we commonly refer to three kinds of statistical biases: selection bias, attrition bias, and monitoring bias. The first kind of bias, selection bias, exists when the difference between the two groups is linked not *only* to the treatment, but *also* to other factors apart from the program or the treatment.[13] Randomized (1) and controlled (2) dimensions allow us to remove this bias.[14] Thus, both experiments are controlled and randomized. Both experiments are also (3) multicentric, in that both implemented in different places at the same time: the JUPITER

trial was implemented at 1,315 sites, and to guarantee the validity of the results, the *Balsakhi* program was also implemented in Mumbai in 2001-2002 and in 2002-2003. The experiment is designed (4) with the *intention to treat*, that is to say the evaluation will take into account the effects the treatment had on patients who left the experiment, which indeed occurred in both of these experiments. Finally, to avoid a monitoring bias—an experiment can be (5) blind to both the experimenters and the participants, so as not to influence any party—only the JUPITER trial was *double-blind* (blind to both parties); the participants of the *Balsakhi* program, in contrast, knew of its goal, so it was not blind to any party.[15]

Both of these aspects, the experimental design and the statistical rigour, define a randomized experiment. Once we have mapped the analogy, we can grasp more precisely the similarities between the experimental design of the trials in both fields. A randomized experiment is a specific design characterized by random assignment into at least two groups, a process serving as a guarantee for reliable statistical results.

3 Mapping Dissimilarities between RCTs and RFEs: Toward an Incomplete Analogy

Put simply, medicine is concerned with sick people, and development economics, with poor people, so the domains in which RCTs and RFEs are embedded differ. RCTs are embedded in the therapeutic process, which aims to test a medication. RFEs in development economics do not branch from this two-fold process (therapeutic and clinical); in fact, they leave this two-fold process out altogether, thus rendering the analogy with medicine incomplete. The aim in this section is to define and demonstrate this analogical incompleteness. To do so, I will first map the one-to-one incongruities within the two trials, showing that RCTs and RFEs focus on different effects: RCTs focus on potential side effects, whereas RFEs aim to draw out different effects from the experimental results in order to learn and collect knowledge from the experiment (3.1.). This main incongruity could be attributed to the fact that a RCT is the last test in a long, two-fold process that requires the accumulation of a lot of background knowledge before the clinical-trial stage can begin. I will then map the one-to-one incongruities from outside the trial, by showing that this two-fold process is clearly absent from the RFEs, resulting in said incompleteness, which explains why J-PAL's researchers attempted to gather the background knowledge missing from the experiment (3.2.).

3.1 Inside the Trial: Plan of Experiments and Statistical Background

A randomized experiment is a set of scripts. Here, I attempt to describe these scripts in both medicine and economics in order to map the incongruities (see Table 2) of the two methodologies. Doing so will not only identify precisely what is at stake methodologically in both experiments, but also determine what both experiments aim to produce.

Both experiments had to define their experimental samples[16] depending on precisely what the experiment sought to test, the target sample could change. In medicine, such a sample is defined through a study population with explicit exclusion and inclusion criteria. In the JUPITER trial, the target population, as described above, comprised men aged 50 years and older and women aged 60 years and older[17] with no history of cardiovascular events, a level of LDL cholesterol below 130 mg, a high-sensitivity C-reactive protein level above 2.0 mg, and a triglyceride level below 500 mg. The level of LDL cholesterol, as defined for the sample, is considered "normal," whereas the level of high-sensitivity C-reactive protein is considered below the "normal" threshold, respecting in that sense the hypothesis of the trial. To join the trial, participants had to commit to the entire trial by providing their written, informed consent. To identify this willingness and to estimate participant compliance, all participants went through a run-in phase in which they all received the placebo. The exclusion criteria were quite specific[18] in order to avoid interaction effects and to guarantee the completion of the study.

In the *Balsakhi* program in development economics, the sample included 58 of the 122 schools in Vadodara city and comprised only children in degrees 3 and 4. The sample was then stratified by language, exam score, and gender. As seen in the previous section, this experiment aimed to assess the basic reading and writing skills of trial participants who should have just acquired an elementary ability to read and write. Including children in degrees 3 and 4 in the sample allowed the researchers to evaluate the effects of participation in the program for one and two years, respectively. No exclusion criteria were defined, and the experimental frame made no references to informed consent or willingness to participate. Thus, the target sample in medicine is much clearer and more explicit, thereby preventing interaction effects, while defining such a sample in economics seems to be much more flexible. In addition to these specificities, a statistical baseline, as well as clinical endpoints, was set for the participants in the medical trial. However, neither of these two aspects appeared in the field trial in economics, which weakens the analogy between the two methodologies: if one is not explicit enough about the eligibility criteria of the target population, how can one generalize the outcome?

Both experiments enabled measurement of the different effects of the

educational program and the treatment; these measures are also useful in mapping the incongruities between these two trials. RCT test both (1) the safety and (2) the efficacy of the treatment, whereas RFEs do not take safety into account. In order to meet the regulatory standards and to guarantee the safety of the trial participants, RCT commonly assess compliance effects, side effects, and the occurrence of adverse events during the trials. Compliance effects describe how closely participants follow the instructions of a trial (e.g., taking the pills in the JUPITER trial). Ridker *et al.* (2008) report that participants did indeed follow instructions carefully during the trial, which compared LDL cholesterol levels in the two groups; however, the researchers note that compliance decreased after the trial. In development economics, the researchers seldom account for such effects. One obvious reason could be that the program being tested is usually already visible; the only factor researchers might need to control for or monitor whether the field partners faithfully implement the program.[19] Furthermore, RCT list all of the adverse events occurring during the trials that could lead to potential drug-induced side effects. In the rosuvastatin trial, the adverse events observed in the treatment and control groups were similar; however, after the trial, more patients from the treatment group than from the control group reported diabetes to their physicians. Further trials have been conducted to determine whether diabetes was a side effect of rosuvastatin treatment. Because such studies do not exist in development economics experiment the interaction between different causal factors is not assessed in a comparable manner.

The experiment assessing the *Balsakhi* program distinguishes between several effects: short-term and long-term effects, distributional effects, and direct and indirect effects. In the-short term, as seen above, the *Balsakhi* program seems effective. In this case, the researchers assess such effects by comparing the exam scores of children before and after implementation of the program. The likelihood of these effects persisting after the experiment is over is low. To determine whether the observed short-term effects would persist in the long run, researchers compared the effects of exposure to the program over one versus two years, explaining that if the effects lasted two years, then they could be considered durable and cumulative (Banerjee *et al.*, 2007, p. 1255). Researchers found that, in the long run, the positive effects disappeared; on average short-term effects became insignificant. The *Balsakhi* program targeted children with lower exam scores in order to improve their basic reading and writing skills; however, researchers found that it might also be interesting to determine whether the program benefited all the children who participated in it and whether participation in the program harmed the children at the top of the class. Therefore, the researchers looked at the distributional effects among three sub-populations:[20] (1) the children with the highest exam scores, (2) the children with average exam scores, and (3) the children with the lowest exam scores. They conclude that the

program most benefited the children with the lowest exam scores. Finally, Banerjee *et al.* (2007) aim to open the box of their results by looking at the direct and indirect effects of the program, since in such an evaluation, two effects can conflate with each other: the program can affect children benefiting from the program as well as children not benefiting from it. In other words, the goal is to determine whether such a program might have spillover effects. For that purpose, Banerjee *et al.* (2007) use both an instrumental variable analysis and a discontinuity one. They claim that the effects of the *Balsakhi* program concerned mainly children who directly benefited from it; nonetheless, they affirm that they cannot reject the hypothesis that this program may have affected children who did not directly benefit from it.

Consequently, RCTs and RFEs focus on different types of effects. The medical trial tends to concentrate on (1) the safety of the treatment (the harms that a medication can cause by precisely demonstrating the adverse events that occur during trials as well as potential side effects) and (2) the efficacy of the treatment (the validity of the trial's results). RFEs, however, focus only on the efficacy of the program. They aim to distinguish between conflated effects. In this sense, because development programs often generate spillover effects, researchers tend to separate direct and indirect effects. Furthermore, researchers spell out the results of the experiments with further statistical analyses, such as specifying the distribution of the treatment for certain specific sub-populations. In this sense, because RFEs have less control over the participants eligibility, the number of unnoticed events is more likely to be higher than in medicine, which threatens the validity of the RFEs results.

3.2 Thinking Outside the Trial: the Therapeutic Phases

The fact that a RCT concerns only phase III of a therapeutic trial may explain the substantial methodological differences that arise from using similar experiments for significantly different purposes. I will therefore go beyond the analogy between the two experimental designs and attempt to explore the disanalogy between the context in which RFEs and RCTs produce their results. In medicine, a therapeutic trial consists of five phases: the preclinical phase, phase I, phase II, phase III, and phase IV (Bouvenot & Vray, 2006). Phase III (the RCT) is the last test carried out in the development of a new drug. Before reaching phase III, however, knowledge about the drug's effects and mechanisms will have already been gathered. In medicine, the preclinical phase aims to determine the principle effect of a treatment, often by running tests on animals. Beginning in the 1980s (prior to the JUPITER trial) several RCTs of other statins assessed their effects on the primary and secondary prevention of heart disease (for a review, see Greene (2007) and Brugts *et al.*, 2009). Some experiments involved animals to properly determine the effects of rosuvastatin (for a review, see Olsson, MacTagart,

and Raza, 2002) and to define the preclinical phase of rosuvastatin. Phase I entails the first administration of a treatment in a healthy human being and aims to identify the conditions of the patients tolerance beings. Previous studies on healthy volunteers have aimed to determine whether the effects of rosuvastatin in humans are comparable to those observed during animal testing (Martin, Mitchell & Schneck, 2002). The goal of phase II is to define the conditions under which a treatment is effective as well as the optimal dosage. Phase II involves a group of healthy volunteers. Studies have also determined the optimal dosage of rosuvastatins in healthy individuals (Olsson, Pears, McKellar, Mizan & Raza, 2001). Phase III studies the efficacy of the treatment in sick people and consists of a RCT (consider here the JUPITER trial). The last phase, phase IV, aims to define the long-term effects of a treatment. Researchers monitored rosuvastatin treatment to identify some of its potential side effects, such as diabetes, as tested in the METEOR, CORONA[21], and AURORA trials. The Food and Drug Administration (FDA)[22] finally approved rosuvastatin in 2003, and a steering committee accepted the JUPITER trial's recommendations, leading to the end of the trial in March 2008. Rosuvastatin, branded Crestor, is now the third top-selling drug in the United-States.

In this context, a RCT is just one piece of a larger trial,[23] and this larger trial has a specific goal: to determine the effectiveness of a new medication to treat a specific disease. The therapeutic process follows the clinical process. The latter helps to diagnose and then define a remedy, whereas the former is concerned only with the remedy. The RCT aims to reliably determine the effects of the remedy, and the last phase assesses the efficacy of this treatment in daily life outside the experimental design. We commonly distinguish efficacy from effectiveness (Cartwright, 2009; Rothwell, 2005; El-Serag, 2007). The notion of efficacy concerns the effects of a treatment within the experimental design (i.e., within the RCT).The notion of effectiveness concerns the effects of the treatment outside the experiment, which involves generalizing and translating these effects into another context. In other words, efficacy and effectiveness translate to the common distinction between internal and external validity.[24] If we wish to increase the effectiveness of a treatment or its external validity, one must first pursue some qualitative inquiries.[25] In sum, a lot of background knowledge is accumulated before reaching the third phase of a therapeutic trial (Teira, 2011, 2013; Reiss & Teira 2013); data gathered after the trial serve to determine the long-term effects of a treatment. Phase IV tests the effectiveness of the treatment. In medicine, phase III is fallible, but its conclusions are supported by previous and subsequent studies.

As Table 3 shows, J-PAL's approach borrows from medicine only its third phase and not the four other phases of a therapeutic trial. Thus, in Banerjee *et al.*'s experiments, the researchers begin with no explicit hypothesis, but with an empirical fact, namely a wide-ranging study of child

literacy in India (see: Pratham, 2005). This study provided evidence that many Indian students in degree three had not yet acquired the basic reading and writing skills that they should have acquired in degrees one and two. Many students simply were not learning to read and write at school. Balancing this empirical fact with one of the Millennium Goals[26] (making primary school universal), researchers showed that increasing the number of children at school risked undermining what children learn at school. The objective of the experiment was quite clear: to test the impact of *one possible way* to increase learning among children, and the researchers cited only a few references: the Millennium Goals (UN, 2000), the World Bank report (2004), and Hanushek (1995, 1986)—two empirical works on schooling and its "quality" (regardless of whether children learn at school). These works do not serve to explain the results of Banerjee *et al.*'s experiment, however, and none of these works deals with the impact of hiring young women to improve learning at school. It seems that the experiment begins with a puzzle (children are not learning at school) and then shows that the *Balshaki* program is cost effective, since it is really inexpensive and increases students' exam scores (during the experiment, at least). J-PAL's researchers began with an unexplained (and empirically grounded) puzzle and aimed not first to theoretically explain it, but rather to solve the problem it represents. In this sense, by importing only phase III of the therapeutic trial and ignoring the other phases, which entail accumulated background knowledge, J-PAL's RFEs simply bring to development economics the last test phase of a medical trial. Thus, RFEs not only share incongruities, but miss out on the advantages which these rigorous, fundamental phases of RCTs provide. RFEs in development economics have a weaker grasp of the causal knowledge at play than do RCTs in medicine. Because RFEs do not test to determine an adequate dosage, the conclusions one should draw from previous outcomes in development economics, as well as what counts as replication, remain unclear. As a result of these incongruities, any analogy between these two methodologies is incomplete. This incompleteness explains some of the differences between RCTs and RFEs: namely that neither experiment focuses on the same effects. J-PAL's researchers focus on effects that allow them to unpack and spell out their experimental results in order to draw from the experiment only the knowledge they are missing.

4 Implications of the Incomplete Analogy: Theoretical Frames and Political Recommendations

According to J-PAL's researchers, the best way to solve empirical puzzles is through randomized experiments. But can randomized experiments alone succeed at such a daunting task, especially with an issue as complex as poverty, where many often interrelated causal factors are at play? The in-

completeness of the analogy raises two main issues: a theoretical and a political one. This section shows that by refusing to formulate and drive the experiments on any explicit *a priori* theoretical frames, J-PAL's RFEs struggle to produce both univocal explanations of their results and clear political recommendations. I will show that this struggle is the main implication of the incomplete analogy. For that purpose, I will first show that, contrary to the main criticisms of J-PAL's approach, this approach is not a-theoretical, but rather provides two statuses to the theory: (1) the refusal of any prior theories driving the experiments and (2) the desire to build a unified theory from the experimental results. However, creating this unified theoretical framework appears difficult without first establishing clear theoretical guidance (4.1.). I will show that the incompleteness of the analogy with RCTs carries not only implications for a future theoretical framework, but important political implications for the fight against poverty also. Without accumulating background knowledge beforehand, drawing clear policy recommendations from the results of the RFEs will likely be difficult (4.2.). These two implications tend to undermine the "quiet revolution" to which Banerjee and Duflo aspire.

4.1 Toward an Explicit Formulation of an *Ex-ante* Theory

Development programs are evaluated in order to determine *what works* (i.e., what is effective) in combating poverty.[27] The trial enabled J-PAL's researchers to put aside any *a priori* assumptions and explain why they focused only on phase III and not the others, thereby rendering their approach "neutral;" in this sense, the researchers focused their knowledge-gathering efforts on experimentation. Thus, the epistemology underlining RFEs in development economics is explicitly empiricist. The epistemic stance of J-PALs researchers can be defined as follows: one should gather first evidence in the field and then build a theoretical frame based on this evidence. Such an epistemic stance implies that RFEs work in isolation, which threatens the construction of a theoretical frame.

By assessing an intervention, the experiment would create the theory. Banerjee (2005) explains that this approach entails two phases: in the first, the experiment should produce the theory, and in the second, a theoretical framework of "new development economics" should emerge from the experimental results. Thus, both development economics and development policies are based on evidence. To emphasize this point, Banerjee (2005) claims that the theoretical advances in behavioral economics are based on contradictory frames, which lead to theoretical confusion. Thus, for Banerjee (2005) the experiment should not be driven on such a theoretical frame, as it shows no clear direction; rather, one should assess specific programs in order to determine what is most effective in the fight against poverty and then build on these insights. In the same vein, Duflo (2006a, 2006b, 2010) and Banerjee

and Duflo (2009, 2011) show that prior and explicit theoretical insights are usually misleading and unhelpful:

> What this research was making clear is that, at the level of the efficacy of individual ingredients of the educational production function, our intuition (or economic theory per se) was unlikely to be very helpful—how could we possibly know, a priori, that deworming is so much more effective than hiring a teacher. (Banerjee & Duflo, 2009, p. 153)

Many critiques (Deaton, 2010; Rodrik, 2009; Leamer, 2010, Harrison, 2011a; Ravallion, 2009b; Barett and Carter, 2010) of RFEs in development economics have pointed out the a-theoretical nature of J-PAL's approach. However, the epistemic stance of the theory surrounding J-PAL's approach seems more complex than a simple refusal. To re-frame it, it is necessary to understand the theory surrounding J-PALs approach at two levels: *ex-ante* and *ex-post*. The *ex-ante* theory expresses explicit prior theoretical frames before implementing, and in a sense guiding, the experiment; the *ex-post* status represents a desire to build a new theoretical framework based on reliable results from RFEs. The *Balsakhi* experiment, however, has no explicit theoretical framework, as explained above, but only empirical references, since the aim is to begin with an empirical fact (or puzzle) and then to theorize from it. However, no real theoretical framework develops from this experiment; rather, some "plausible explanations" (Banerjee *et al.*, 2007, p. 1262) emerge. Therefore, the results on which "new development economics" should be based resemble a "black-box test of 'what works"' (Deaton, 2010, p. 451) or what does not work to reduce poverty. To base precise guidance for both policy-makers as well as a unified theoretical framework on these results seems complicated. The theoretical insights from the *Balsakhi* experiment appear unclear, and lack background knowledge: Why do the *Balsakhi* program's effects fail to endure in the long run? Why do all sub-populations seem to benefit from the *Balsakhi* program and not only the target population (i.e., the most marginalized children)? An *ex ante* theory or accumulated background knowledge before the RFE appears to be missing, but this is not the case with RCTs, where background knowledge is accumulated prior to testing.

Without a doubt, background knowledge about the relationship between education and development economics in the *Balsakhi* experiment is implicit. For instance, the *Balsakhi* program targets basic education skills and remedial education.[28] Although Banerjee *et al.* (2007) never referred explicitly to that field, the background knowledge of such a field is at least relevant to the experiment, even if it does not drive it. However, an explicit prior theoretical framework in such a field (consider Kozeracki (2002), who highlights seven ways to improve students' remedial education)[29] could have

permitted the testing of an explicit hypothesis or a prior theoretical frame, especially given that Duflo and Banerjee (2009, p. 156) raised the fact that "experiments are emerging as a powerful tool for testing theories." However, the absence of an explicit *ex-ante* theory to test weakens that power. The wish for an *ex-post* theory based on results with strong internal validity prevents researchers from starting with a prior theory because, in doing so, they would no longer be neutral. In this sense, the only part to import from medicine is phase III of the therapeutical trial. In this way, randomized experiments in development economics work in isolation from previous efforts to accumulate background knowledge, which tends to weaken their results, thereby rendering them non explicit, languishing in a black-box from which one must extract plausible explanations. Consequently, the "challenge for a new development economics" (Banerjee, 2005, p. 4340) is to compensate for the shortcomings of early experiments with RFEs, efforts which may help to develop more rigorous RFEs in development economics.

4.2 From Local to Global Policy Recommendations: Toward a "Quiet Revolution"

One of the main objectives of J-PAL is to offer guidance to policy-makers in developing countries.[30] If Banerjee and Duflo (2011) target a "quiet revolution," this policy guidance has to function on a two-fold level: locally and globally. On the local level, evaluations of the development program should offer clear recommendations for local policy and refer to the location where the program has been assessed. On the global level, these policy recommendations should more generally improve the lives of the poor. For Banerjee and Duflo, as for the creation of an *ex-post* theory, such policy guidance should emerge from the experiments and its results. Nonetheless, on the local level, policy recommendations from the *Balsakhi* experiment seem somewhat indefinite. For instance, the main policy lesson drawn from the experiment is that "general school-quality improvements must be complemented by specific strategies to improve the performance of marginalized children" (J-PAL Policy Briefcase, 2006, p. 7). Such specific strategies are not so straightforward. Are the researchers referring to the *Balsakhi* program as a specific strategy or are they referring to others strategies? The researchers insist on two dimensions: (1) the low cost of the *Balsakhi* program and (2) the need to focus on marginalized children. By running a cost-benefit analysis, researchers show that the *Balsakhi* program is cheaper than the other five programs in education[31]. For instance, the *Balsakhi* program costs 2.25$ per child per year for a 0.27 increase in test scores, whereas the "Scholarship for Girls" program costs 3.53$ per child per year for a 0.53 increase in test scores. However, how one should use that cost analysis remains unclear. Should Indian cities implement a cheaper program with fewer gains or a more expensive program with larger gains? More generally, what

policy lessons should be drawn from this research to help improve Indian cities? The recommendation from the JUPITER trial was unequivocal (to prevent primary cardiovascular events, one should use rosuvastatin) and led to the end of the trial in March 2008. The univocity of the recommendation does not imply that the JUPITER trial was free of controversy; as a matter of fact, the JUPITER trial did face some controversies (see Andreoletti & Teira, forthcoming; Ridker, 2009; De Lorgeril, 2010), but the accumulation of background knowledge and the entire frame in which RCTs are embedded undoubtedly supported the recommendation.

Concerning the global level, Banerjee and Duflo (2011) offer a strong diatribe of what they call "political economy" represented mainly in the work of Acemoglu and Robinson (2012), who claim that institutions are key in development and, thus, target both structural and global changes. In contrast for Banerjee and Duflo (2011, p. xi), global approaches to fighting poverty are only "magic bullets" with no proof of their efficacy; instead, Banerjee and Duflo propose a change of perspective: focusing not on Institutions (with a capital "I"), but on institutions (with a lower case "i"). In other words, the accumulation of small changes should lead to incremental changes. By only focusing on phase III of a trial, however, the *Balsakhi* evaluation agrees completely with this proposition. However, Banerjee and Duflo fail to define the passage from local changes (institutions) that have proven their efficacy through RFEs to more global changes (Institutions) in the lives of the poor remains, leaving them rather vague. Thus, in addition to the challenge embodied in the *ex-post* theory, there is also the challenge of Institutional change. Banerjee, Banerji, Duflo, Glennerster and Khemani (2010), again in partnership with Pratham, assess a vast nationwide program known as "Read India". This program aims to increase both basic reading skills and collective cooperation in India. The goal of this program is to encourage volunteers to teach children in their city how to read by offering pedagogical training. This program was implemented in only a few cities in India. The results, like those in the *Balsakhi* program, were considered very positive and very context dependent, yet no real explanation for them was offered.[32] Banerjee *et al.* (2010) interpret these results as an indication that small institutional changes can be extremely powerful, thus emphasizing the idea that a transformation of the public sector to improve education is unnecessary:

> This suggests that in settings in which the public service delivery system is entirely unresponsive to beneficiaries, identifying innovative ways to foster and channel local action may be the most effective way to improve the final outcomes. Pratham's 'Read India' program is a particularly powerful example of such an intervention. (Banerjee *et al.*, 2010, p. 29)

Even though small changes, such as the "Read India" program, can have

positive impacts, how these programs should be implemented and how one can scale up from the local scale to a more general level remains unclear. If Banerjee and Duflo (2011) want to achieve this "quiet revolution," they will have to accumulate (from the explicit formulation of prior *ex-ante* theory) background knowledge and then complete the analogy with medicine. Nevertheless, to do so, J-PAL has to relinquish some internal validity by focusing on more than the experimental procedure and its statistical rigour. In order to make sense of these results, J-PAL, in alignment with the empirical revolution that it created, must generate the missing knowledge that will ultimately lead to a quiet revolution. One has to go from the empirical facts to an *ex-ante* theory and then test it and finally build the *ex-post* theory that would express clearly both local and global political recommendations. Indeed, Jeffrey Sachs, one of the main figures in development economics, highlights the necessity of such clinics by defining the role of a differential approach and aiming to define a "clinical economy" (Sachs, 2005, p. 75). Dani Rodrik (2010) also insists on the role of making a diagnosis before offering a prescription. Both of these development economists make a clear reference to the clinic, and both are in same sub-field as J-PAL. The desire for neutrality also diverts J-PAL from other disciplines which have tried to offer a qualitative view of poverty, as have many anthropological works (Collins, Murdoch, Rutherford & Ruthven, 2009; Farmer, 2004), or even works from the World Bank (Chambers, Narayan, Shah & Petesch, 2000), which conducted more than 20 000 interviews with the poor in order to understand their needs. In this sense, the field should not only be one of experiments, but one apt for the explicit formulation of an *ex-ante* theory and its test. This urges J-PAL to adopt a much more general interdisciplinary approach, one developed not only by borrowing one element from medicine, but by trying to build a "new development economics" with the help of other disciplines (and even with help from within the discipline) that can help to provide access to the missing knowledge and then to complete the analogy.

5 Conclusion

In contrast with the notion of identity, the notion of analogy, which recognizes both similarities and differences, led me to examine the experiments made in medicine and development economics through their disciplinary transfers. By mapping their one-to-one correspondences, their similarities served to define what RFEs borrow from medicine; this, in turn, enabled a precise definition of randomized experiments. The differences between the two domains break with the common idea that RCTs and RFEs are identical. The mapping of the one-to-one and incongruities shows that the analogy is incomplete, since RFEs borrow only one of the empirical phases of the

therapeutic trial. J-PAL's approach tends mainly to describe poverty rather than to fully explain it, let alone solve it. Although J-PAL's experiments may indeed constitute an "empirical revolution," the more substantial revolution promised by Banerjee and Duflo (2011, p. 265) can only be achieved by more consistently opening the disciplinary boundaries to recent work in both development economics and other disciplines, such as anthropology or political science.

Notes

[1]Note: This is my translation from (Duflo, 2010, p. 17) ; the original quotation reads: "*Au XXéme siécle, les essais cliniques ont révolutionné la pratique de la médecine. Malheureusement, il n'en est pas de même pour les politiques relatives á l'éducation et la santé. Bien souvent, elles ne sont pas évaluées rigoureusement avant d'être généralisées. (...) Il est cependant possible de s'inspirer des essais cliniques pour conduire des évaluations de programmes pilotes en matiére d'éducation et de santé.*

[2]See Marks (1997) for a history of this methodological revolution in medicine.

[3]Vincent Guillin's work (2013) is one of the only attempts to analyze RFEs in development economics as distinct from randomized controlled trials in medicine. Guillin assesses the existence of causal capacities within RFEs in development economics and concludes that RFEs do not yet allow the production of such causal capacities.

[4]For a study of analogical reasoning, see Achinstein (1964); Bartha, (2010, 2013); Hesse (1966). For a study of analogical arguments, see Juhte (2005).

[5]The "replicability crisis" is one illustration of such objections; for an overview of this crisis, see (Andreoletti & Teira, forthcoming).

[6]According to Hacking (1988), Charles Sanders Peirce and his assistant were the first to introduce a random dimension to an experimental design at the end of the 19th century in order to test low stimuli; see (Peirce & Jastrow, 1885).

[7]Transporting cholesterol (as well as all fat molecules) around the body is the task of two lipoproteins: high-density lipoprotein (HDL) and low-density lipoprotein (LDL). LDL is often labeled as "bad cholesterol" because high levels of this lipoprotein lead to the deposition of LDL in the arteries, which increases ones risk for cardiovascular events. On the other hand, HDL is considered "good cholesterol," because this lipoprotein removes oxidized cholesterol from the arteries and moves it to the liver, which degrades it.

[8]Statins are molecules that decrease cholesterol levels. Several kinds of statins are currently available on the market, such as atorvastatin (sold as Lipitor), fluvastatin (sold as Lescol), and rosuvastatin (sold as Crestor); I will focus only on rosuvastatin.

[9] NGO Pratham was created in 1994 in Mumbai with the support of UNICEF with the aim of improving primary education in India. http://www.pratham.org/ (accessed 23 July 2015)

[10]Degree three corresponds to the third year of primary school. The Indian educational system is based on eight years of primary school and two years of secondary school. The first five years are devoted to basic education. Students are divided into five degrees, with degree three corresponding to the first year.

[11]For the AURORA trial, see Fellstrom *et al.* (2009); for the ASTROMER trial, see Chan, Teo, Dumesnil, Ni and Tam (2010); for the METEOR trial, see Crouse *et al.* (2007); lastly, for the ASTEROID trial, see Nissen *et al.*(2006)

[12]Fisher (1926) first defines these six characteristics: the comparison and the random dimension (p. 87), the randomized blocks (p. 90), the replication (p. 86), the controlled aspect (p. 92) and the need for a factorial plan when testing several aspects (pp. 92-93).

Fisher (1935) then systematizes these six characteristics in his landmark book *The Design of Experiment*.

[13]A selection bias is present when the treatment is not the only explanation for the differences between the two groups. Randomization is a unique method enabling the removal of such a bias; removing such a bias is in fact its primary appeal. For an analysis of randomization and selection bias, see Heckman, Ichimura, Smith and Todd (1998).

[14]However, James Heckman and Jeffrey Smith (1995) have showed that instead of removing the selection bias, randomized experiments only balance it between the two groups: "Randomized social experiments solve the problem of selection bias (...) Finally note that random assignment does not remove selection bias, but instead balances the bias between the participant and non-participant samples." (Heckman and Smith, 1995, pp. 88-89) Furthermore, Worrall (2010, p. 292) notes "that nothing in the argument shows that the *only* way that selection bias can be eliminated is through randomization."(Emphasis in the original)

[15]See Teira (2013) for an analysis of blinding in both RCTs in medicine and RFEs in the social sciences.

[16]The size of the samples in our two experiments are similar (each consisting of more than 15 000 individuals);

[17]The age of the sample is important: Cardiovascular events usually occur after the age of 50 for men and after the age of 60 for women (Ridker, 2003; Ridker *et al.*, 2007).

[18]The terms included no use of hormones, no diabetes, no incidence of cancer in the last five years of the participant's life, no history of alcohol or drug abuse, and no inflammatory conditions

[19]Most experiments carried out by the J-PAL pay close attention to this aspect; in fact, Cohen and Dupas's (2010) experiment on bed nets in Kenya even introduced an incentive system to ensure that field partners meticulously follow the allocation of the program.

[20]The results of randomized experiments are the average treatment effects for the treatment group and the average effects for the control group. Thus, the researchers had no access to the individual distribution. They could distinguish between only a few subpopulations using the data collected from the experiment. This leads to a major limitation of this methodology: the heterogeneity treatment effect, which means that, on average, results may appear positive while in fact harming many individuals in the sample. For a review of this limitation in medicine, see Baslow, Duran and Kravitz (2004); for a review in development economics, see Deaton (2010), Barett and Carter (2010), Harrison (2011a), Ravallion (2009a); and more generally in economics, see Heckman and Smith (1995).

[21]See Rogers *et al.*, 2014

[22]For a review of rosuvastatin from the FDA, see Roberts (2009)

[23]It is worth noting that in medicine, some researchers carry out experiments focusing only on phase III. My intention here is simply to illustrate the larger frame in which RCTs are usually embedded.

[24]For definitions and distinctions between these two concepts, see (Campbell, 1957); for a precise distinction of the two concepts and its stakes in medicine, see (Rothwell, 2005); for a precise distinction and the stakes of the two notions within experimental economics, see (Guala, 2005, 2003, 1999; Guala & Mittone, 2005).

[25]Nancy Cartwright, for instance, shows that to go from efficacy to effectiveness, randomized experiments should define causal capacities. Therefore, to increase external validity, randomized experiments should focus on the intrinsic qualities of the intervention (Cartwright, 1989).

[26]The Millenium Goals are eight developmental objectives defined by the United Nations in 2000 that aim to achieve the goals by 2015.

[27]For a precise presentation of how development programs should be evaluated, see (Duflo, Glennester, & Kremer, 2007).

[28]Remedial education is a recent branch of education that targets basic literacy and numeracy learning.

[29]In the seven ways that Kozeracki (2002) emphasizes, tutoring comes in the fifth position. A prior and explicit reference to this work could have placed Banerjee *et al.*'s experiment in the field of remedial education, and thus reinforced the power of their experiment.

[30]The idea here is exactly the same as in evidence-based medicine, which aims to inform medical practices about the results of RCTs. The analogy between such a movement and J-PAL's approach is beyond the scope of this paper, but represents, in itself, a fruitful line of investigation.

[31]The researchers compared the Balsakhi program to (1) the Pratham CAL program (Banerjee *et al.*, 2007), which assessed the effects of a computer assistance learning program; (2) the Scholarship for Girls program (Kremer, Miguel & Thornton, 2009), which provided merit-based scholarships for girls to attend primary school; (3) the Teacher Incentives program (Glewwe, Kremer, Moulin & Zitzewitz, 2002), which rewarded teachers; (4) external monitoring (Duflo & Hanna, 2012), which aimed to reduce teacher absences; and (5) the Tennessee's Project Star program, which tested the effect of class size reduction.

[32]"Whatever the explanation, it seems clear that the current faith in participation as a panacea for the problems of service delivery is unwarranted. The results seem to depend, in very complex ways, on the details of the intervention and the contexts. An intervention designed according to the best practice rules failed to have any impact on the public education sector in India." (Banerjee *et al.*, 2010, p. 29)

References

Acemoglu, D.& Robinson, J. (2012). *Why Nations Fail, The Origins of Power, Prosperity, and Poverty*. New-York, NY: Crown Business.

Achinstein, P. (1964). Models, Analogies and Theories. *Philosophy of Science*, 31, 328-349.

Albert, M.A., Danielson E., Rifai, N. & Ridker, P. (2001). Effect of Statin Therapy on C-Reactive Protein Levels: the Pravastatin Inflammation/CRP Evaluation (PRINCE), a Randomized Trial and Cohort Study. *Journal of the American Medical Association*, 286, 64-70.

Andreoletti, M., & Teira, D. (forthcoming). Statistical Evidence and the Reliability of Medical Research, In: H. Kincaid, J. Simon & M. Solomon (Eds.), *The Routledge Handbook of Philosophy of Medicine*. London: Routledge.

Angrist, J., & Pischke, J-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking Con out of Econometrics. *Journal of Economic Perspectives*, 24(2), 3-30.

Ballantyne, C., Hoogeveen, R., Bang, H., Coresh, J., Folsom, A., Heiss, G. & Sharrett, A. (2004). Lipoprotein-associated phospholipase A2, high-sensitivity C-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the Atherosclerosis Risk in Communities (ARIC) study. *Circulation*, 109, 837-842.

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

Banerjee, A., & Duflo, E. (2009). The Experimental Approach to Development Economics. *Annual Review of Economics*, 1, 151-178.

Banerjee, A., & Duflo, E. (2011). *Poor Economics. A Radical Rethinking of the Way to Fight Global Poverty*. New York, NY: Public Affairs.

Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics*, 122(3), 1235-1264.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R. & Khemani, S. (2010). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. *American Economic Journal: Economic Policy*, 2(1), 1-30.

Barett, C., & Carter, M. (2010). The Power and the Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspective and Policy*, 32(4), 515-548.

Bartha, P. (2010). *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. New York, NY: Oxford University Press.

Bartha, P. (2013). *Analogy and Analogical Reasoning*. The Stanford Encyclopedia of Philosophy. E.N. Zalta (Ed.)

Baslow, J., Duran, N. & Kravitz, R. (2004). Evidence-Based Medicine, Heterogeneity of Treatment Effects, and The Trouble With Average. *he Milbank Quaterly*, 82(4), 661-687.

Bouvenot, G., & Vray, M. (2006). *Essais cliniques, théories, pratique et critique* [Clinical Trials, Theories, Practice and Critic]. Paris: Flammarion.

Brugts, J., Yetgin, T., Hoeks, S., Gotto, A., Shepherd, J., Westendorp, R., ... Deckers, J. (2009). The Benefits of Statins in People Without Established Cardiovascular Disease but with Cardiovascular Risk Factors: Meta-analysis of Randomized Controlled trials. *British Medical Journal*, 338, b2376.

Campbell, D. (1957). Factors Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin*, 54(4), 297-312.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Cartwright, N. (2007). Are RCT's the Gold Standard?. *Biosocieties*, 2(1), 11-20.

Cartwright, N. (2009). What is This Thing Called Efficacy?. In C. Mantzavinos (Ed.), *Philosophy of the Social Science. Philosophical Theory and Scientific Practice* (pp. 185-206). Cambridge: Cambridge University Press.

Cartwright, N. (2010). What Are Randomized Controlled Trials Good For?. *Philosophical Studies*, 147(1), 59-70.

Cartwright, N. (2011). A Philosopher's View of the Long Road From RCT's to Effectiveness. *The Lancet*, 377, 1400-1401.

Chambers, R., Narayan, D., Shah, M.K, & Petesch, P. (2000). *Crying Out for Change: Voices of the Poor*. Washington DC: World Bank.

Chan, K.L., Teo, K., Dumesnil, J.G., Ni, A., & Tam, J. (2010). Effect of Lipid Lowering with Rosuvastatin on Progression of Aortic Stenosis: Results of the Aortic Stenosis Progression Observation: Measuring Effects of Rosuvastatin (ASTRONOMER) Trial. *Circulation*, 121(2), 306-314.

Collins, D., Morduch, J., Rutherford, S., & Ruthven, O. (2009). *Portfolios of the Poor. How the World's Poor Live on $2 a Day*. Princeton, NJ: Princeton University Press.

Cohen, J. & Dupas, P. (2010). Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment. *Quaterly Journal of Economics*, CXXV(1), 1-45.

Crouse, J.R., Raichien, J.S., Riley, W.A., Evans, G.W., Palmer, M.K., O'Leary, D.H., ... Bots, M.L. (2007). Effect of Rosuvastatin on Progression of Carotid Intima-Media Thickness in Low-Risk Individuals with Subclinical Atherosclerosis: the METEOR Trial. *Journal of the American Medical Association*, 297(12), 1344-1353.

Deaton, A. (2010). Intruments, Randomization, and Learning About Development. *Journal of Economic Literature*, 48, 424-455.

De Lorgeril, M., Salen, M.P., Abramson, J., Dodin, S., Hamazaki, T., Kostucki, W., ... Rabaeus, M. (2010). Cholesterol Lowering, Cardiovascular Diseases, and the Rosuvastatin-Jupiter Controversy. A Critical Reappraisal. *Archives of Internal Medicine*, 170(12), 1032-1036.

Duflo, E. (2006a). Poor but Rational?. In A. Banerjee, D. Mookherjee & R. Benabou (Eds.), *Understanding Poverty* (pp. 367-378). New York, NY: Oxford University Press.

Duflo, E. (2006b). Field Experiments in Development Economics. In R. Blundell, W. Newey & T.Person (Eds.), *Advances in Economics and Econometrics: Theory and Applications* (pp. 322-348). New York, NY: Cambridge Universty Press.

Duflo, E. (2010). *Lutter contre la pauvreté, le développement humain, tome 1* [Fighting Poverty, Human Development, Volume 1]. Paris: La république des idées.

Duflo, E. & Hanna, R. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, 102(4), 1241-1278.

Duflo, E., Glennerster, R., & Kremer M. (2007). Using Randomization in Development Economics Research: A Toolkit. In P. Schultz & J. Strauss (Eds.), *Handbook of Development Economics* (pp. 3895-62). Amsterdam, Elsevier Science.

El-Serag, H. (2007). Comments from the Editors. *Gastroenterology*, 123, 8-10.

Farmer, P. (2004). *Pathologies of Power: Health, Human Rights, and the War on the Poor*. San Francisco, CA: University of California Press.

Felstrom, B.C., Jardine, A.G., Schmieder, R.E., Holdaas, H., Bannister, K., Beutler, J., ... Zannad, F. (2009). Rosuvastatin and Cardiovascular Events in Patients Undergoing Hemodialysis. *The New England Journal of Medicine*, 360(14), 1395-1407.

Fisher, R.A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33, 503-513.

Fisher, R.A. (1935/1960). *The Design of Experiments*. New York, NY: Hafner Publishing Company Inc.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7, 155-170.

Glewwe, P., Kremer, M., Moulin, S. & Zitzewitz, E. (2002). Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya. *Journal of Development Economics*, 74(1), 251-268.

Greene, J.A. (2007). *Prescribing by Numbers: Drugs and the Definition of Disease*. Baltimore, MD: John Hopkins University Press.

Guala, F. (1999). The Problem of External Validity (or 'Parallelism') in Experimental Economics. *Social Science Information*, 38, 555-573.

Guala, F. (2003). Experimental Localism and External Validity. *Philosophy of Science*, 70, 1195-1205.

Guala, F. (2005). *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Guala, F., & Mittone, L. (2005). Experiments in Economics: External Validity and the Robustness of Phenomena. *Journal of Economic Methodology*, 12(4), 495-515.

Guillin, V. (2013). De quoi les essais contrôlés randomisés sont-ils capables? Evaluation, mécanismes et capacités en sciences sociales [What Are the Randomized Controlled Trials Capable of? Evaluation, Mechanisms and Capacities in Social Sciences]. *Cahiers Philosophiques*, 133(2), 79-102.

Hacking, I. (1988). Telepathy: Origins of Randomization in Experimental Design. *Isis*, 79(3), 427-451.

Hanushek, E. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, 24(3), 1141-1177.

Hanushek, E. (1995). Interpreting Recent Research on Schooling in Developing Countries. *The World Bank Researcher Observer*, 10(2), 227-246.

Harrison, G. (2011a). Randomization and its Discontents. *Journal of African Economies*, 20(4), 626-652.

Harrison, G. (2011b). Field Experiments and Methodological Intolerance. *Journal of Economic Methodology*, 20(2), 103-107.

Heckman, J. & Smith, J. (1995). Assessing the Case for Social Experiments. *The Journal of Economic Perspectives*, 9(2), 85-110.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing Selection Bias Using Experimental Datas. *Econometrica*, 66(5), 1017-1098.

Heron, M., Hoyert, D., Murphy, S., Xu, J., Kochanek, K. & Tejada-Vera, B. (2009). Deaths: Final Data for 2006. *National Vital Statistics Reports. U.S. Department of Health and Human Services*, 57(14), 1-136.

Hesse, M. (1966). *Models and Analogies in Science*. Southbend, IN: University of Notre Dame Press.

The Abdul Latif Jameel Poverty Action Lab (J-PAL). (2006). Making Schools Work for Marginalized Children: Evidence from an Inexpensive and Effective Program in India. *Policy Briefcase*. MIT.

Juthe, A. (2005). Argument by Analogy. *Argumentation*, 19, 1-27.

Kozeracki, C. A. (2002). ERIC Review: Issues in Developmental Education. *Community College Review*, 29, 83-101.

Kremer, M., Miguel, E. & Thornton, R. (2009). Incentives to Learn. *The Review of Economics and Statistics*, 91(3), 437-456.

Leamer, E. (2010). Tantalus on The Road to Asymptopia. *Journal of Economic Perspectives*, 24(2), 31-46.

Marks, H.. (1997). *The Progress of Experiment: Science and Therapeutic Reform in the United-States, 1900–1990* . New York, NY: Cambridge University Press.

Martin P.D., Mitchell, P.D. & Schneck, D.W. (2002). Pharmacodynamic Effects and Pharmacokinetics of a New HMG-CoA Reductase Inhibitor, Rosuvastatin, After Morning or Evening Administration in Healthy Volunteers. *British Journal of Clinical Pharmacology*, 54, 72-77.

Nissen, S.E., Nicholls, S.J., Sipahi, I., Libby, P., raichlen, J.S., Ballantyne, C.M., ... Tuzcu, E.M. (2006). Effect of Very High-Intensity Statin Therapy on Regression of Coronary Atherosclerosis: the ASTEROID Trial. *Journal of the American Medical Association*, 295(13), 1556-1565.

Olsson, A.G., Pears, J., McKellar, J., Mizan, J. & Raza, A. (2001). Effect of Rosuvastatin on Low-Density Lipoprotein Cholesterol in Patients with Hypercholesterolemia. *The American Journal of Cardiology*, 88(5), 504-508.

Olsson, A.G., McTaggart, F., & Raza, A. (2002). Rosuvastatin: A Highly Effective New HMG-CoA Reductase Inhibitor. *Cardiovascular Drug Reviews*, 20(4), 303-328.

Peirce, C.S., & Jastrow, J. (1885). On Small Differences in Sensation. *Memoirs of the National Academy of Sciences*, 3, 73-83.

Pratham Organization. (2005). Annual Status of Education Report. *Pratham Resources Center*. Mumbai.

Ravallion, M. (2009a). Evaluation in The practice of Development. *World Bank Research Observer*, 24(1), 29-53.

Ravallion, M. (2009b). Should the Randomistas Rule ?. *The Economist' Voice. Berkeley Electronic Press*, 6(2), 1-6.

Reiss, J., & Teira, D. (2013). Causality, Impartiality and Evidence-Based Policy. In H-K. Chao, S-T Chen, & R.L. Millstein (Eds.), *Mechanism and Causality in biology and Economics* (pp. 207-224). New York, NY: Springer.

Ridker, P., Cushman, M., Stampfer, M., Tracy, R., Hennekens, C. (1997). Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *New England Journal of Medidicine*, 336, 973-979.

Ridker, P., Rifai, N., Pfeffer, M.A., Sacks, F. & Braunwald, E. (1999). Long-Term Effects of Pravastatin on Plasma Concentration of C-Reactive Protein. *Circulation*, 100, 230-235.

Ridker, P. (2003). Rosuvastatin in the Primary Prevention of Cardiovascular Disease Among Patients With Low Levels of Low-Density Lipoprotein Cholesterol and Elevated High-Sensitivity C-Reactive Protein: Rationale and Design of the JUPITER Trial. *Circulation*, 108, 2292-2297.

Ridker, P., Fonseca F., Genest, J., Gotto, A., Kastelein, J., Khurmi, N., ... Glynn, J. (2007). Baseline Characteristics of Participants in the JUPITER Trial, a Randomized Placebo-Controlled Primary Prevention Trial of Statin Therapy Among Individuals With Low Low-Density Lipoprotein Cholesterol and Elevated High-Sensitivity C-Reactive Protein. *The American Journal of Cardiology*, 100, 1659-1664.

Ridker, P.M., Danielson, E., Fonseca, F., Genest, J., Gotto, A., ... Kastelein, J. (2008). Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein. *The New England Journal of Medicine*, 359(21), 2195-2207.

Ridker, P. (2009). The JUPITER Trial: Results, Controversies and Implications for Prevention. *Circulation*, 2, 279-285.

Roberts, M.D. (2009). *New Drug Application 21-366/S016: CRESTOR (rosuvastatin calcium)*. Clinical Briefing Document: Endocrine and Metabolic Drugs Advisory Committee Meeting .

Rodrik, D. (2009). The New Development Economics: We Shall Experiment, But How Shall We Learn. In J. Cohen, & W. Easterly (Eds.), *What Works in Development, Thinking Big and Thinking Small* (pp. 24-48). Washington, DC: Brookings Institution Press.

Rodrik, D. (2010). Diagnostics Before Prescription. *Journal of Economic Perspectives*, 24(3), 33-44.

Rogers, J.K., Jhund, P.S., Perez, A., Bohm, M., Cleland, J.G., Gullestad, L., ... Pocock, S.J. (2014). Effect of Rosuvastatin on Repeat Heart Failure Hospitalizations. The CORONA Trial (Controlled Rosuvastatin Multinational Trial in Heart Failure). *Journal of the American College of Cardiology: Heart Failure*, 2(3), 289-297.

Rothwell, P. (2005). External Validity of Randomised Controlled Trials: To Whom Do the Result of This Trial Apply?. *The Lancet*, 365, 82-93.

Sachs, J. (2005). *The End of Poverty: Economic Possibilities for Our Time*. New York, NY: Penguin Press.

Teira, D. (2011). Bayesian versus Frequentist Clinical Trials. In F. Gifford (Ed.), *Philosophy of Medicine (Handbook of Philosophy of Science, volume 16)* (pp. 255-299). Oxford, Elsevier.

Teira, D. (2013). Blinding and the Non-Interference Assumption in Medical and Social Trials. *Philosophy of the Social Sciences*, 43(3), 358-372.

United Nations (UN). (2000). *United Nations Millennium Declaration*. General Assembly: New York, NY.

World Bank. (2004). *World Development Report 2004: Making Services Work for the Poor*. New York: Oxford University Press.

Worrall, J. (2010). Do We Need Some Large, Simple Randomized Trials in Medicine?. In M. Suárez, M. Dorato, M. Rédei (Eds.), *EPSA Philosophical Issues in the Sciences* (pp. 289-301). Dordrecht, Springer.

List of Tables

Rosuvastatins and Heart Disease	*Balsakhi* Program and Primary Education
Experimental Design	
(1) Comparison	= (1) Comparison
(2) Randomized	= (2) Randomized
(3) Two randomized Blocks	= (3) Two Randomized Blocks
(4) Replication	= (4) Replication
(5) Orthogonality	= (5) Orthogonality
(6) No Factorial Plan	= (6) No Factorial Plan
Statistical Rigor	
(1) Randomized	= (1) Randomized
(2) Controlled	= (2) Controlled
(3) Multi-centric	= (3) Multi-Centric
(4) Intent to Treat	= (4) Intent to Treat
(5) Double-Blind Study	$\neq$ (5) Open Study

Table 1: Mapping One-to-One Correspondences

JUPITER trial	*Balsakhi* Evaluation
Creating the box: the target population	
- eligibility criteria: explicit inclusion and exclusion criteria	$\emptyset$ targeted a sample without explicit eligibility criteria
- run-in phase	$\emptyset$ no run-in phase
- baseline characteristics	$\neq$ stratified the sample by languages, test scores, and gender
Entering the box: the treatment-program effects	
- compliance effect	$\neq$ short-term and long-run effects
- adverse events	$\neq$ distributional effects
- side effects	$\neq$ direct and indirect effects

Table 2: Inside the Trial: Mapping One-to-One Incongruities

Rosuvastatin and Heart Disease Prevention	Hiring Teachers and Primary Education
Phases of a Therapeutic Trial	
Preclinical Phase Studies on animals (Olsson *et al.*, 2002)	
Phase I Previous tests on healthy individuals (Martin *et al.*, 2002)	
Phase II Optimal dose (Olsson *et al.*, 2001)	
Phase III RCT (Ridker *et al.*, 2008)	**Phase III**: RFE (Banerjee *et al.*, 2007)
Phase IV Monitoring (METEOR, CORONA and AURORA trials)	

Table 3: Outside the Trial: Mapping One-to-One Incongruities