RANDOMIZED CONTROLLED TRIALS AT WAR

Jules Salomone

However increasingly lauded for their supposed objectivity and in spite of the tremendous hopes they are arousing among economists who more and more welcome them as the undreamed-of way out of the bleak skepticism affecting nowadays economics, and especially development economics, Randomized Controlled Trials (RCT) are not without their own biases and blind spots. A close examination of The Abdul Latif Jameel Poverty Action Lab (J-PAL) discourse and practice casts light on the methodological hegemony to which RCTs aspire as well as the striking warlike exercise of power on which the use of such a tool seems to rest. Although surprising, this second result is actually corroborated by the history of this methodology for which World War II researches in psychology played a decisive role: never before the existence of the Experimental Branch of the Army's Information and Education Divsion had anyone hoped to that extent that an as rigorous technique as RCTs would be increasingly used so as to build indisputable theories and solve all kinds of social problems. Consequently, Johnson Administration's "*War* on Poverty", which heralded the Golden Age of RCTs on the basis of the WWII experience and the institutions created in its aftermath, as well as the more recent J-PAL endeavor appear as nothing more than a conquering policy of neutralization of politics allegedly made necessary by an urgent war to be waged.

Keywords: Randomized Controlled Trials – War on poverty – Social engineering – Economic imperialism

Codes: A12 – B23

Acknowledgments

I would like to warmly thank Jean-Yves Grenier for his unfailing support and the friendly attention he paid to my endeavor, Jérôme Bourdieu for having kindly accepted to read this paper, Romain Huret for his pertinent remarks, Cédric Durand for his comments on a first version of this work, Maha Jafri and Nicholas Taylor for their time, precious comments, and decisive proofreading.

I am highly indebted to Grégoire Chamayou for the enthusiastic interest he manifested for this work, and for having encouraged me to explore the links between randomized controlled trials and behaviorism. I could not have foreseen that this was going to be such a fertile direction.

I also owe a lot to Pierre Pénet who, in the course of my researches, gave me, as one the most precious gifts, one of the N-gram Viewer-based graphs that I discussed in my lexicographical study. Without him, I would have probably not discovered this powerful tool.

I am very grateful to Agnès Labrousse for her reading advise, the accuracy of her remarks, and her repeatedly invitation to the most demanding rigor in writing this paper.

I finally have a special thought for Clémentine Van Effenterre, Arthur Jatteau, Thomas Cortado, Françoise Arqué, Christian Salomone, and Simon Bittman.

CONTENTS

INTRODUCTION

The increasingly widespread use of Randomized Controlled Trials (RCT) in the evaluation of public policies – and most notably in development economics – is probably one of the most important transformations to which economics has been recently bearing witness. Their success and rapid popularity owes in large part to their reputation for objectivity, which is seen as an antidote to the biases and methodological problems that plague conventional econometrics. Many researchers welcomed them with enthusiasm, considering them the spearhead of the "credibility revolution in empirical economics" which "better research designs" made possible (Angrist and Pischke, 2010). Over the last few years, RCTs have become the yardstick against which any empirical strategy has to be compared. For instance, the methodology of instrumental variables (IV), which was designed to address the issue of endogeneity without any direct reference to the experimentalist paradigm, is increasingly reduced to the status of second-best solution, employable only when RCTs are not available. Furthermore, if the IV is a binary variable – which is often the case – the analogy between the two methods, whether well-funded or not, is made obvious: the sample can be split into the "control" and the "treatment" groups of a so-called *natural experiment*. The gap between economics and experimental sciences would then be on the verge of being bridged.

Nowadays, RCTs are widely used in development economics, especially by the Jameel Poverty Action Lab (J-PAL). Since its creation in 2003, it has coordinated more than two hundred programs and has become a prominent actor in the fight against poverty, often in partnership with NGOs and governments. However, according to the head of J-PAL, Esther Duflo, who has recently been awarded the John Bates Clark Medal, "creating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize *social policy* [emphasis added] during the 21st century, just as randomized trials revolutionized medicine during the 20th" (Deaton, 2009). In other words, RCTs are very likely to become the gold standard methodology of public policy evaluation in general. The present crisis is not likely to reverse the trend. Thanks to their relative independence from the classical questions of standard economic theory and

to their openness to the unexpected questions which may arise from data and field work, RCTs would indeed help to pave the way out of the bleak pessimism affecting economics and policymaking. As stated by an article of the Bloomberg Businessweek quoted on the J-PAL website, "the financial crisis blew a hole in big-think economics, raising the profile of a new breed of skeptical empiricists committed to assiduous testing and tangible results, no matter how tiny. Even lentils can lead to little miracles."[1] Therefore, lauded for their supposed modesty, RCTs may end up being bolstered by the skepticism cast on standard models. One could even argue that the ongoing recession has aggravated preexisting poverty – in developing as well as in developed countries – and made the resources required to combat it increasingly scarce. The need for such evaluations would then be all the more urgent. At any rate, if the J-PAL's call for a more systematic use of RCTs in policy evaluation was heard, not only in developing areas, but more generally in any socio-economic context, this would constitute an important date in the century-long history of this methodology.

Development economists working for this organization do not claim to have discovered the core principles of randomized experiments, and they usually trace their origins back to Ronald Fisher's seminal works in statistics and biometrics (1926; 1935), as well as they often mention their increasingly widespread use in the United States from the 1960s onwards. But they are probably right when they credit themselves with a renewed use of this methodology since no one before the J-PAL, except of course RCTs practitioners in the medical field, had envisioned such an international destiny for this kind of evaluation technique. However, the historical reconstruction in which they dab turns out to be incomplete, if not partial. For instance, historians of RCTs (Oakley, 1998; 2000; Dehue, 1997; 2001; Hacking, 1988), when their studies do not exclusively focus on the medical field and the related birth of evidence-based medicine in the decades following the end of WWII (Marks, 1999; Keel, 2011), emphasize the too often overlooked decisive role played by psychologists at the very beginning of the 20th century, arguing in favor of a close link between controlled experiments and early behaviorism. Similarly the often alluded to idea that medical research made possible the use of RCTs in non medical fields does not seem to be supported by the actual facts. Interestingly though, few are the studies

1 Bloomberg Businessweek, July 2, 2010 (http://www.businessweek.com/magazine/content/10_28/b4186056393103.htm)

which aim at uncovering the historical conditions of possibility of the routinized use of RCTs. Dehue (1997; 2001), Hacking (1988) and Danziger (2000), even though challenging the traditional epistemology of randomization, do not uncover the architecture of power relationships which was necessary to its systematic use. All in all, the transformation of RCTs into a genuine political tool is rarely studied. Besides, what little criticism currently exists of RCTs[2] mostly focuses on issues related to the generalization of results to different places and periods – external validity[3] – (Rodrik, 2008) or bemoans the lack of a coherent theoretical structure (Deaton, 2009 ; Labrousse, 2010) in which the results of the RCTs could find their proper place. Ethical questions are often referenced but are generally left to medical journals to pursue more thoroughly (see Worrall, 2007).

In light of the way they are used by their main promoter, the J-PAL, this paper argues that randomized experiments, far from being the self-evident methodology to which anyone willing to fight poverty should appeal, are made possible by a certain approach to under-development, which is not without its own biases and blind spots. The very notion of objectivity that characterizes RCTs, rather than simply reflecting the validity of the method, serves to obscure some of its problems. I argue that, in spite of their reputation for localized initiatives and on-the-ground results that prioritize the concerns of specific populations, RCTs in fact aspire to and participate in a hegemonic, potentially neo-colonial enterprise by engaging in a politically problematic war on poverty. I would like to suggest that the increasingly dominant role of RCTs in economic research threatens to further disenfranchize poor populations and to continue an imperialist venture under the guise of research-based, locally driven solutions to poverty.

To so do, not only do I discuss the logic of RCT methodology in terms of the current scholarly debate in economics over the relative superiority of RCTs to other research methods, but I also intend to highlight the invariants of the architecture of power relationships in which such a methodology becomes a routinized proof production technique. Of course, identifying those invariants is an arduous task for which a mere historical investigation, although inevitable, is far from being enough. Indeed, by doing so,

2 In a recent article, Banerjee and Duflo (2009) take stock of the "rising tide of criticism" generated by their approach: they "list these objections and then [discuss] each one" (152, 159). Unfortunately, their bibliography does not systematically refer to the articles from which those objections arise.

3 As opposed to internal validity which has to do with inference.

one might just end up gathering a collection of isolated uses of RCTs whose connection would then be disputable. Therefore, my argument is two-fold. First, thanks to a close reading of both research and popularization articles, I point to the somehow puzzling incoherences to which RCTs, as put into practice nowadays by the J-PAL and other institutions, seem to necessarily lead, and I then try to interpret them as the pieces of a systematic set of power relationships. To do so, I only had to bear in mind that, as noticed by Oakley (1998; 2000), the 1960s-1980s was nicknamed, at least in the United States, the Golden Age of such a methodology, and that the War on Poverty, launched in 1964 by Johnson, seems to have been its decisive condition of possibility. Indeed, if taken at face value, such an expression gives a surprisingly fertile key to the interpretation of what RCTs require in order to be widely used, that is, a warlike architecture of power relationships. At the same time, this result seemed both unlikely and weak, as compared to all its unsettling consequences, not to mention the fact that the kind of war to which it points is now out-of-date. Second then, I explore the history of RCTs so as to verify if their first widespread use coincided with a particularly warlike period of time. Surprisingly, the implicit assumption that the 1960s-1980s Golden Age had been preceded by only some isolated attempts turns out to be wrong. Indeed, the watershed, in the history of RCTs, appears to have occurred during WWII, some twenty years before the period of time on which Oakley focused, and a couple of decades after their first and hesitant use in the field of psychology. Strikingly, my first rigorous description of the architecture of power relationships on which such a methodology rests, description which did not require more than a quick glance at the most prominent features of its history, receives a blatant confirmation of its validity. As a consequence, warlike exercise of power turns out to be more than a mere coincidental characterization of the contexts in which RCTs are widely employed. Had I inverted the steps of my reasoning, one could have criticized this analysis for having over-interpreted the present context in which field trials are now used on the basis of a mere historical account. Conversely, the order of presentation I chose, which is in strict conformity with the chronology of my discoveries, conveys the idea that the peculiar functioning of power architecture on which the use of RCTs depends may be *invariant*.

The remainder of the article proceeds as follows. Section 1 assesses the strengths of this methodology and explains the reasons why it has remained so far out of reach of

the existing criticisms. Still, they unveil some caveats and paradoxes which are discussed more thoroughly in section 2. Section 3 argues that the war on poverty seems to be the proper name of the technology of power in which those paradoxes are made innocuous to the very practice of randomized experiments. Section 4 concludes the first part. Section 5 explores the origins of RCTs, which were to become a routinized instrument during WWII, as thoroughly discussed in section 6. Section 7 portrays the political thought which led to the "War on Poverty," and section 8 challenges some of the historiographical views about the alleged period of decline which followed the 1960s-1980s Golden Age of randomization.

I

THE POLITICS OF RANDOMIZED CONTROLLED TRIALS

1

The End of Autism?

The greatest strength of RCTs – or at least in the way they are used by the J-PAL – lies in the new role they give to empirical studies in economics. Instead of testing existing theories with already-collected data, field research is likely to discover jarring facts at odds with economic intuitions, which are not reduced to mere anomalies but in turn call for new experiments and innovative explanations. Banerjee and Duflo (2009) refer to Seva Mandir, a NGO in Rajasthan, which wanted to improve its informal schooling system. They designed a program in which children had to write in a diary what they had done in class everyday. It was hypothesized that parents, newly aware of teachers and their children's absenteeism, would go on to take a more active part in their children's educations. The program did not bring about those expected results: parental opinion about the schooling system was not sensitive to the number of missed days of class. However, it turned out that diaries still had a positive effect on test scores. Since this surprising outcome was not part of the initial protocol, it might have been a statistical accident. So Seva Mandir and the J-PAL are now conducting a new experiment in order to evaluate the efficiency of diaries as a pedagogical tool. Results are forthcoming.[4] Such a methodological choice would have been hard to justify in the traditional econometric context, in which unexpected results can be dismissed as mere anomalies when they conflict with accepted theories or initial intuitions. Conversely, the surprising discovery, in this case, provoked a new round of evaluations. Thanks to this example, it is first made obvious that, as stated by Duflo (2009), "to be valid, an experiment does not require a valid theory."[5] For their being reliable, RCTs prevent econometricians from questioning first and foremost the data or the methodology. Thus, mere intuitions and popularly

4 Here is the link to the presentation of this ongoing experiment: http://www.povertyactionlab.org/evaluation/teacher-and-student-motivation-family-participation-and-student-achievement-rural-udaipur

5 "La validité de l'expérience ne repose pas sur la validité de la théorie." (*Ibid*:67) Quoted by Labrousse (2010). Note that "expérience" could actually be translated by "experience" - with the meaning of "to experience something".

accepted theories run the risk of being more easily challenged when tested with RCTs. Second, new ideas which emerge during the process of evaluation can be easily tested, as shown in the example of Seva Mandir, by conducting additional experiments. Finally, empirical studies can identify the most cost-efficient program among a set of different policies. For example, *a priori* knowledge would not have been useful for predicting that deworming children leads to better school attendance than providing their family with financial incentives (Duflo, 2010a:37 ; Banerjee and Duflo, 2009:153). All in all, it seems that observation had never been so important in economics as it has with the advent of RCTs.

Comparing Lucas's seminal article on the driving role of education in economic growth (1988) to the results obtained by the J-PAL, Labrousse (2010) concludes that "one cannot content himself with an abstract and intangible approach: the analysis has to carefully deal with social micro-structures and must rest on a realistic consideration of agents' behaviors and environment."[6] This criticism especially applies to endogenous growth models, which are characterized by disputable hypotheses and a crucial lack of specific recommendations. In contrast, the J-PAL promises to herald an era of tangible results in development economics, in the aftermath of the failure of the Washington Consensus. One may then wonder why Labrousse criticizes RCTs for their lack of epistemological foundations. How can a method be lauded as a major contribution to development economics due to its concrete results and experimental design and yet at the same time be critiqued as ill-founded? Labrousse and Deaton both agree on the fact that randomized experiments focus on "*whether* projects work instead of on *why* they work." (Deaton, 2009:4) According to them, the mechanisms through which policies produce (or fail to produce) their effects remain unexplored. As a consequence, RCT's structurally fail to grasp any *causal* relationships and the knowledge they extract cannot be unified into a coherent theoretical structure. Deaton even goes a step further by asserting that the absence of models which are normally supposed to encapsulate causality jeopardizes the internal as well as external validity of randomized experiments. According to him, "policy requires a causal model; without it, we cannot understand the welfare consequences of a

6 "On ne peut donc se contenter d'une approche surplombante et dématérialisée : l'analyse doit examiner de près les microstructures sociales et reposer sur une prise en compte réaliste des comportements des agents et de leur cadre de vie." (*Ibid*:7)

policy, even a policy where causality is established and that is proven to work on its own terms." (*Ibid*:43) In light of these concerns, Seva Mandir's experiment calls into question the very premise of both authors' reasoning: the whole experiment has been conducted in order to discover *concrete* mechanisms that will improve education. As such, the claim that causality has nothing to do with the findings seems inaccurate, no matter how localized in space and time the causal links the results unveil may be. As argued previously, field research is probably the best way to explore complex interactions of concrete heterogeneous processes.[7]

Deaton and Labrousse's reasoning seems to be actually driven by the sharp distinction both seem to make between economics as a *science* and econometrics as a *technique*. Indeed, Deaton concludes his article by saying that "technique is never a substitute for the business of doing economics" (*Ibid*:47) while Labrousse notices that the way RCTs are used by the J-PAL leads to a "reduction of economics to a mere technique."[8] The economic *discourse*, the structure of its axioms, theorems and propositions should then not be mistaken for the systematic *practices* either directly associated to its development, like statistical proof techniques, or aimed at its political applications. However, both authors seem to oscillate between this sharp theoretical distinction between science and technique, and an undifferentiated mixture of both. Referring to studies led by Duflo, Kremer and Robinson (2009) on the use of fertilizers and by Giné and Karlan (2010) on smoking-cessation, Deaton explains that in both cases, "the project (...) is the *embodiment* [emphasis added] of the theory that is being tested and refined, not the object of evaluation in its own right, and the field experiments are a bridge between the laboratory and the analysis of 'natural' data." (*Ibid*:46) Does "embodiment" mean that, according to Deaton, those experiments allow then for the unification of technique and science? Does it imply that good evaluations, according to Deaton's standards, require the distinction between discourse and practice to be blurry? More importantly, causality,

7 Labrousse (2010:10-11 and n. 18) firmly distinguishes causality proofs (that clearly establish a causal link between two or more given events) from efficiency proofs (which answer the following question: "is the measure implemented suitable to achieve a set of defined goals, given the available resources?") only about which RCTs have something to say. If the process of experimentation is as "creative" as J-PAL advocates say, such distinction might be called into question.

8 "Cette dissolution de l'économie dans la technique (...)" (*Ibid*:22)

which seemed to be an end in itself in his thinking, is necessary to "understand the welfare consequences of a policy". (*Ibid*:43) In other words, economics which primarily explores complex interactions of causal links between various phenomena would then not be more than the technical task of comparing expected outcomes of alternative policies. Even though this obvious mixture of discourse and practice seemed at first to be under the rule of scientific reason ("not the object of evaluation in its own right"), its final purpose could actually be quite *technical*. Indeed, Deaton asserts that econometricians should design "experiments to test predictions of theories that are *generalizable* [emphasis added] to other situations." (*Ibid*:45) Scientific knowledge would then be meant to tailor policies and to "understand the[ir] welfare consequences" before being generalized to a whole population. The theoretical distinction between science and technique has been so rigidly defined and the mixtures of both are so obviously unavoidable that both economics *and* econometrics seem to have to shift toward one and only one of the terms of the alternative. Deaton's solution to this problem is a good example of *technocratic* thought in which science has to be *tied* to technical endeavors. In other words, for those mixtures to exist and to produce their effects and for the distinction between discourse and practice to maintain its hermeticism, technocracy requires economics to always be an *applied* science.

Conversely, Labrousse's article may embody the symmetrically positivist situation. Like Deaton, she notices that science and technique are easily mistaken one for the other. She refers to Desrosières (2008), who often describes economics as social engineering and writes that "statistics was a way to promote both desideologization and objectivation [of social issues] (…)"[9] when it is no more than a mere technique. But unlike Deaton, she seems to be less interested in making science more technical than making techniques more scientific. In other words, Labrousse's criticism tends to neutralize RCTs , understood as a "technology to govern populations",[10] by integrating their results into a more unified theory of behaviors. Indeed, she explains that the "reduction of economics to a mere technique, the relative self-restriction to mundane matters, leave the room for a *de facto* appropriation [of this technique] by mainstream economics. This is quite surprising because a lot of the results obtained by the new development economics do not fit with the

9 "La statistique a donc été investie d'un rôle (...) de désidéologisation et d'objectivation (…)." (Desrosières, 2008, first volume:22; quoted by Labrousse, 2009:17)

10 "Une technologie de gouvernement des populations" (*Ibid*:20)

very core of standard economics (stable preferences, independence from the context, etc.)."[11] At the risk of overinterpreting Labrousse's goal, it seems that her epistemological critique aims at making the economic discourse hold sway over econometric practices: eventually, economists could then reject some experimental protocols for their lack of scientific foundations. And if those foundations were heterodox, the hegemony of standard economics would be, for once, seriously called into question.

However, Labrousse's criticism – as well as Deaton's – may have little influence on J-PAL's arguments and practices. It seems that neither Duflo nor her coworkers suggested drawing such rigid distinction between econometrics and economics. Instead, the process of "creative experimentation", as shown with the example of Seva Mandir, interweaves rigorous estimations and innovative intuitions. This methodology cannot be accused of being technocratic because unexpected observations uncovered by field work are carefully taken into account and are given the opportunity to deeply regenerate the theory. Economics is not reduced to a catalog of matter-of-course techniques that simply distinguish good generalizable policies from bad – which is often to say, costly – programs. Instead, science remains quite independent from techniques insofar as new intuitions formulated during the experimental process can *disrupt* well accepted practices. Conversely, the theoretical structure in which the results from RCTs find their place is not unified enough to abide by positivist principles. But it is actually one of the strengths of this method. Although the results they obtain may not be ever-lasting and usually only apply to small areas, they are still true and potentially effective. RCTs would then be putting an end to the autism of mainstream economics often criticized by heterodox scholars. According to those critics, standard theory deals less with real economic phenomena than with imaginary worlds incorporated in traditional models and their simplifying – simplistic? – assumptions. In addition, those fictitious narratives would be far from being innocuous because they would hold sway over economic practices which would, in turn, make them eventually true (Armatte, 2010). A rigid distinction between discourse and practice seems to allow for this oscillation between the positivism of

11 "Cette dissolution de l'économie dans la technique, cette relative autolimitation à des enjeux en apparence très prosaïques, permettent une agglomération de facto aux courants dominants dont la théorie standard. Et ce, alors même que nombre de résultats de ces *new development economics* sont incompatibles avec le noyau dur de la théorie standard (stabilité des préférences, indifférence de celles-ci par rapport au contexte etc.)." (*Ibid*:22)

systematic but fictitious models and the technocratism of performative theories which characterizes autistic mainstream economics. The importance given to *pragmatic* empirical observation in RCTs is seen as a potential route out of this stalemate. This pragmatism has allowed randomization to remain relatively unchallenged by either orthodox or heterodox economists.

Still, however blunt their general argument is, some of Deaton and Labrousse's criticisms shed light on paradoxes inherent to this methodology. Although these critiques do not substantially undermine the discourse and practices related to RCTs, they contribute to the understanding of their logic.

2
The End of Skepticism?

Deaton, in the same article, focuses on somewhat overlooked issues related to the internal validity of results obtained through the use of RCTs. As repeatedly explained by advocates of RCT methodology, randomization makes the program being tested orthogonal to the error term. As a consequence, they argue, RCTs make possible the exclusion of endogeneity biases, which are always likely to alter the results in standard econometrics. Most of the time, then, debates focus on the degree to which policies are generalizable to other regions and periods of time. In other words, external validity is more often discussed than internal validity. The significance of Deaton's article lies in his effort to challenge RCTs by focusing on the foundations of their supposed superiority over standard methods.

For instance, Deaton explores the problem of heterogeneity in the treatment effects. Having reasserted that RCTs are only “informative about the mean of the treatment effects” (*Ibid*:26), he explains that “the trial might reveal an average positive effect although nearly all of the population is hurt with a few receiving very large benefits” (*Ibid*:27). Banerjee and Duflo respond that this problem is not specific to RCTs. Indeed, counterfactuals cannot be observed with any empirical strategy and as a

consequence distributions of the treatment effects cannot be computed. This does not prevent econometricians from computing post-trial subgroup means of the treatment effects.[12] But as Deaton argues, "a sufficiently determined examination of any trial will eventually reveal some subgroup for whom the treatment yielded a significant effect of some sort." (*Ibid*:28) In other words, *data mining* may jeopardize the supposed impartiality of the results. Additional hypotheses would be required in order to avoid this and Banerjee and Duflo are not reluctant to adopt them (*Ibid*:170).

This debate raises several questions about the validity of RCTs. First, RCTs, while seemingly driven by empirical objectivity, in fact necessitate more assumptions than it might seem at first glance in order to produce meaningful results. In practice, so-called gold-standard evidence *and* debatable results coexist in a variety of empirical strategies. It might even be the case that this situation is not accidental. Indeed, according to Deaton, "RCTs, although frequently useful, are not exempt from the routine statistical and substantive scrutiny that should be routinely applied to any empirical investigation." (*Ibid*:40) Banerjee and Duflo do not seem to dispute this claim. Second, instead of calling into question the supposed superiority of this methodology, those compromises with the ideal protocol reassert it. Randomization is still considered the best way of dealing with the looming issue of endogeneity. As noticed by Deaton (*Ibid*:29), the results from an RCT are compelling. For example, a physician who, in his professional judgment, believes that a particular patient's condition will worsen with a certain RCT-proven treatment has the right to make the decision to refuse to prescribe that treatment. In practice, however, such a decision becomes harder to make when it contradicts results from a RCT. So, even when the internal validity of a result requires additional hypotheses that contrast with the strict evidential standards of randomization, the importance of the very method, rather than being played down, turns out to be unchallenged.

Furthmore, what if, as Deaton wonders, exogeneity of the treatment to the error term was not guaranteed? He explains that "the flagship study of the new movement in

12 Deaton comments on this econometric model:

$$Y_i = \beta_0 + \beta_1 T_i + \Sigma_j \theta_j X_{ij} + \Sigma_j \varphi_j X_{ij} T_i + u_i$$

The *X's* are various controls. This model allows the effect to vary according to individuals' characteristics.

development economics, Miguel and Kremer's (2004) study of intestinal worms in Kenya" did not actually employ RCTs. "Private communication with Michael Kremer has confirmed that, in fact, the local partners would not permit the use of random numbers for assignment, so that the assignment of schools to three groups was done in alphabetical order." (*Ibid*:38-39) It might be the case that projects implemented by other NGOs or the Kenyan government were conducted using the same method of organization, that is alphabetization. Orthogonality of the treatment could be then put in jeopardy. Deaton also argues that, more generally speaking, "subjects may fail to accept assignment, so that people who are assigned to the experimental group may refuse, and controls may find a way of getting the treatment, and either may drop out of the experiment altogether" (*Ibid*:36). Any kind of compromise with the ideal protocol is likely to endanger the most vaunted property of RCTs, namely the exogeneity of the treatment. Then, the very fact that the advocates of this methodology are willing, at least in certain circumstances, to jeopardize its scientificity is quite baffling. It is even more surprising when considering how fearful those economists are of the potential distrust compromised figures could cast on their endeavor. Duflo writes, for instance, that "a failure, when it occurs, is likely to call into question *all* [emphasis added] the efforts devoted [to combat poverty]."[13] Why would they then take such a risk? Probably because, at least in Kremer and Miguel's case, it was not truly hazardous. At the risk of overinterpreting this example, it seems that the authors were willing to sacrifice some of the impartiality of the protocol to make sure that the experiment would take place. In other words, the expected outcome of a likely discovery outweighed the cost induced by the potential introduction of some kind of bias. The "RCT" conducted in France before the implementation in 2009 of the Revenu de Solidarité Active shows that Kremer and Miguel's case is not isolated: neither regions nor recipients were randomly assigned to treatment and control groups. In practice, is it then the case that these deviations from the strict rules of RCT experimental design are not accidental? It is very probably true that those quasi-randomizations do not alter the results of the experiments in a fundamental way. But it cannot be neglected that they contrast forcefully with the alleged purity and objectivity of RCTs. If the superiority of RCTs consists precisely in their status as bias-free, *randomized* experimental models, then any compromises to RCT experimental objectivity go to the heart of the model's reputation

13 "L'échec, quand il survient, risque de discréditer l'ensemble des efforts fournis." (Duflo, 2010a:16)

and validity. Without the randomization that characterizes their method, RCTs become little more than a cousin to other research methods, opening the door to the same criticisms and concerns.

As argued previously, those issues are generally overlooked and critiques of RCTs primarily tend to focus on their lack of generality. Rodrik (2008:5) summarizes those views by saying that "the 'hard evidence' from the randomized evaluation has to be supplemented with lots of soft evidence before it becomes usable." Indeed, a program may not be generalizable to some areas given the existence of important heterogeneities. It might also be the case that its implementation on a large scale cancels the positive effects documented in a small region (see Banerjee and Duflo for a discussion:167-169). But general equilibrium effects[14] and environmental dependence do not seem to be that much of a hindrance, according to Banerjee and Duflo. Indeed, they write that to address those issues, "actual replication studies need to be carried out. Additional experiments have to be conducted in different locations, with different teams." (*Ibid*:160) Interestingly, this recommendation is at odds with one of the reasons why RCTs are implemented in developing countries. Financial resources of often-corrupt states and of NGOs are frequently scarce, and rigorous, small-scale evaluations make the most of meager resources.[15] While this logic might suggest that randomizations are necessarily occasional and should be used only when financial resources are limited, in practice, issues related to the generalizability of the results call for a widespread and frequent use of RCTs.

Another problem lies in the fact that this methodology is not always accepted by local populations, and the fear that there may be heterogeneities between those who refuse it and those who do not. But, according to Banerjee and Duflo, "this is becoming less of an issue as randomized evaluations gain wider acceptance." They add that "this situation will continue to improve if randomized evaluation comes to be recommended by most donors." (*Ibid*:163) As argued earlier, RCTs, when implemented, do require many assumptions. In this case, though, the additional hypotheses required to obtain interpretable results come

14 "This phenomenon of equilibrium effects poses a problem that has no perfect solution. Fortunately, in many instances, this phenomenon does not present itself." (Banerjee and Duflo:167)

15 "Location-level randomization is justified by budget and administrative capacity (…)." (Banerjee and Duflo:166)

from the practice itself: the more randomization will be resorted to, the more solid its outcomes will be. The same could be said about the fact that Banerjee and Duflo seriously suggest that compulsory participation could bridge the gap between the measures of the average treatment effect and the intention to treat.[16] All in all, performativity appears then to play an important role in improving the external validity of experiments. RCTs seemed at first to focus only on local mechanisms and truths situated in space and time. But their validity actually hinges on the degree to which they manage to mobilize financial resources (no matter how scarce), researchers, and populations in order to conduct ever more experiments.

The rising tide of RCTs appears to be sometimes at odds with the issues they investigate. For instance, the increasing amount of money devoted to this kind of evaluations makes one wonder why control groups are still created when programs explicitly target "the ultra poor." The Bandhan study is one of them.[17] This study tries to determine whether providing income-generating assets instead of microcredit is beneficial to the ultra poor and if it enables them to eventually become good microfinance clients. If "ultra-poverty" was as "ultra" as the name seems to indicate, allowing for the existence of a control group would be criminal. RCTs and poverty alleviation would then be ethically incompatible. But the J-PAL is obviously not an organization of mass murder, so the emergency status attached to ultra poverty does not prevent the ultra-poor from waiting until the experiment is done. Still, they are genuinely poor, which is the reason why an experiment targets them and why they are supposedly more likely to accept the protocol. As argued by Banerjee and Duflo, the poor "are often used to such arbitrariness" so "randomization appears both transparent and legitimate." (*Ibid*:166) By Banerjee and Duflo's logic, poverty may bring about the resigned acceptance of protocols in which assignment to one group or the other is arbitrary. As they explain, though, "the evaluation

16 "The IV estimate using the intention to treat as an instrument correctly estimates the average of the impact of this program on the people *who chose to participate* [emphasis added]. However, this fact does not provide information on the average impact of a training program that was made compulsory for all welfare recipients. To find this out, one would need to set up an experiment with compulsory participation." (Banerjee and Duflo:165)

17 Here is the link to the presentation of this ongoing experiment: http://www.povertyactionlab.org/evaluation/helping-ultra-poor-use-microcredit-murshidabad-india

design assumed that everyone who is offered the asset will grab it, which turned out not to be the case." Indeed, "a significant fraction of the clients (18%) refused the offer." (*Ibid*:166) According to them, this might have flown from the lack of information recipients had about the general goals of the program. But "Bandhan may not have put in the kind of public relations effort to inform the villagers about why the program was being conducted, precisely because they were not planning to serve the entire population of the very poor in each village." (*Ibid*:166) First, it seems that the J-PAL assumes that extreme poverty will ensure people's acceptance of the organization's help. In the Bandhan program, Banerjee and Duflo could then only discover to their surprise that "sometimes even a *gift* [emphasis added] may be refused (…)." (*Ibid*:166) Clearly, Banerjee and Duflo do not consider the possibility that program initiatives may not actually be "gifts." Second, the fact that the organization was purposefully unclear about the experiment questions the alleged transparency of the methodology as well as its legitimacy: why would they have hidden the fact that a RCT was being implemented if nobody was likely to dispute its fairness? Banerjee and Duflo also explain that ethics committees generally allow, if needed, that participants are not told they are part of a RCT. In other cases, experimenters sometimes say that the program under scrutiny will soon be generalized, whether it is true or not. If it is true, the very use of RCTs is puzzling: why would the measure not be implemented on a large scale directly? Is randomization not meant to evaluate policies *ex ante*, in other words, before they are made sure that they are going to be enacted? Conversely if experimenters do not tell the truth, then again the transparency of randomization is severely called into question. Furthermore, Banerjee and Duflo say that "when the control areas are given the explanation that the program has enough budget for a certain number of schools only, they typically agree that a lottery is a fair way to allocate those limited resources." (*Ibid*:166) What if those budgets were actually less scarce than what experimenters say? There is indeed an embarrassing contrast between the increasing amount of money and energy devoted to RCTs and the fact that this methodology is at least partly justified by the scarcity of financial resources. Of course, it is surely the case that – sometimes abundant – funds are allocated conditionally on the results of the experiments. At the risk of being refused the right to implement randomization, it would be more accurate to inform participants of this condition, rather than emphasizing the supposed scarcity of available resources.

All in all, RCTs, when put into practice, seem to leave room for debate, oddities and consequently skepticism. Indeed, the infallibility of this methodology as well as its results do not seem as self-evident as argued by their advocates. Moreover, its flaws actually encourage its increased use. Indeed, owing to the fact that RCTs produce only localized results, researchers simply do more and more RCTs, rather than recognizing RCTs as limited in scope.

3
The End of Poverty?

As just shown, the dialectics of poverty and abundance involved in RCTs is rather complex. First, they require poor populations, likely to accept the arbitrariness of randomization, but not so poor that they cannot wait until the results are established. Second, they necessitate scarce financial resources in order to justify the existence of a control group, but not so scarce that the program is unlikely to be implemented on a large scale. Third, they require the humble work of experimenters who study local mechanisms with a few assumptions and gather geographically and historically situated evidence, but not so humble that the very practice of RCTs as well as their results cannot be generalized to other places and periods of time. So, poverty is not the raw and unequivocal material that theories and methods strive to grasp – otherwise, such dialectics would not exist and the poor would be plainly poor. Instead, it seems that RCTs, when put in practice, assign poverty particular characteristics that enable experimenters to say something particular, and perhaps presupposed, about the poor. Poverty is then not a mere given, drawn from evidence and indisputable fact. It is decisively shaped by what can be called a specific set of power relations which articulate what can be done and what can be said in a particular place and period of time. Technocracy and positivism are two examples since both specify the way techniques – or what can be done – and science – or what can be said – are related to each other. Different set of power relations will produce different objects, according to the singular ways they articulate discourses and practices. Poverty is one of those objects. I would like to emphasize that I do not mean to imply that the poor do not exist, or that

adversity is unreal. The products of every single set of power relations are real and must be taken seriously into consideration. Instead, I want to suggest that poverty cannot be separated from such architecture of power relations from which it originated. In other words, poverty cannot be distinguished from what I called its dialectics. Assuming that power relations can be more easily described by focusing on their stress points, I have devoted the previous section of the paper to mapping some of the paradoxes, debates, and odd inconsistencies endemic to randomized experiments. I will now define the set of power relations to which they belong.

Interestingly, the New Jersey experiment, often presented as the first large-scale RCT applied to social policies, was implemented in 1968, at a time when the US had declared "War on Poverty." It is, at least at this point of my argument, far from being certain that there was a structural link in this confluence, and it would be interesting to know how the experts of that time conceived the idea of applying RCTs to the evaluation of social reforms. Still, this historical coincidence indicates that RCTs might be a distinctive feature of wars against poverty. Given how humble and good-hearted randomization is according to its advocates, it might seem surprising at first to associate it with war waging. Indeed, Duflo repeatedly explains that, contrary to Sachs (2005), the J-PAL is not promising the end of poverty and is not likely to discover any kind of miracle solution. Instead, it aims at promoting small but tangible advances which are likely to change the lives of the poor right here and now.[18] But nothing prevents the humility of this technique from occupying the leading role in a war waged without heroism against poverty. A second objection would be to say that this expression is not used very often by its supposed advocates.[19] They more often say that they are fighting or combating poverty.

18 "Mais, si l'on veut pérenniser la lutte contre la pauvreté, tâtonnements, créativité et patience sont indispensables non pour trouver la baguette magique qui n'existe pas, mais pour mettre en place une série de petites avancées qui, dès aujourd'hui, améliorent la vie des plus pauvres." (Duflo, 2010b:104)

19 Duflo gave recently a talk at the Simon Fraser University (SFU), as well as at RAND corporation and at the University of British Columbia, entitled "Experiments, Science and the War Against Poverty." As of yet, it is the only instance of this expression I can find directly associated with her name. Here is the link to this lecture:
http://www.sfu.ca/cstudies/mpprog/projects/public/bmo_lectures.php
http://www.rand.org/media/advisories/2010/03/01.html
http://globalencounters.ubc.ca/events/experiments_science_and_the_war_against_poverty/

But a fight does not require the same kind of logistics, does not imply the same kind of temporality as a war. It seems that this word lets then the paradoxes discussed in the previous section go unnoticed, whereas the expression "war on poverty" gives them a more specific role in the functioning of the set of power relations to which they belong.

First, and as argued previously, RCTs are compelling. If accepted and implemented, their results can hardly be debated. The discussion, if it takes place, only focuses on the disputable hypotheses which most of the time are required in order to use the results. The purity of the method is thus rarely called into question. And in any case, it appears that the importance of heterogeneities in the treatment effects is minimized. It is first assumed that the conclusions drawn from the results hold for all subjects in the sample. In other words, the distribution of the treatment effects is taken into account only subsequently. Conversely, if an NGO or an international organization does not systematically employ RCTs, it may be accused of "lazy thinking" and "resistance to knowledge." This is what Banerjee said of the World Bank, additionally bemoaning the "lack of distinction made between strategies founded on the hard evidence provided by randomized trials or natural experiments and the rest." (Banerjee, 2007, chapter 1, quoted by Deaton:24) All in all, even if, as argued by Deaton, RCTs are necessarily combined with standard econometric methods, implementing them suffices to stem usual debates over the validity of the results as well as the actual content of the evaluated program. Fighting against teachers' absenteeism with cameras with date and time stamps might sound odd at first.[20] If it is rigorously proved to work, however, which is to say, if it has been evaluated and proven effective with a RCT, political resistance to the generalization of the program might not carry a lot of weight as compared with the expected benefits of a more efficient mechanism for the accumulation of human capital.[21] Randomization has

20 Here is the link to this program conducted in Rural Udaipur, India: http://www.povertyactionlab.org/evaluation/encouraging-teacher-attendance-through-monitoring-cameras-rural-udaipur-india

21 The website reads:

"Teachers in government schools are often more politically powerful than teachers in informal or private schools. Thus, it may prove difficult to institute a system in which government teachers would be monitored daily and their pay linked to attendance. However, the above evidence suggests that if teacher attendance can be improved this should flow through into improved test scores."

indeed a strong pacifying power. When waged, the war on poverty makes allegedly unproductive disputes about figures and methods vanish and calls for a spirit of *Union Sacrée* or *Burgfrieden*: in other words, a wartime truce.

Second, this consensus is not jeopardized by some compromises with the ideal protocol. As argued previously, the advocates of RCTs are not reluctant to tamper once in a while with its central principle, namely randomization (Deaton, 2009). Interestingly, the generalization of the program sometimes occurs before clear-cut results are available. This happened in France when the RSA was implemented even though the experiment was still ongoing. As reminded by Gomel and Serverin (2009), the main promoter of this reform, Martin Hirsch, called for a swift generalization of the measure because the poor could not have been made to wait any longer for an improvement of their condition. In both cases, the economic emergency seems to allow core principles of randomized evaluations to be broken. In other words, the war on poverty has to be waged right here and now, no matter if weapons have to be sacrificed in order to win. If randomization cannot be used, alphabetization can do the trick. If an exhaustive assessment of a measure is too long to establish, and if the first results are positive, its generalization should be soon enacted. Does this mean that RCTs sometimes do not fit with the functioning of the war on poverty? Would it then be the case that this specific way of exerting power only incidentally requires this statistical technique? First, if by "incidentally" one understands that randomization could possibly be employed in different set of power relations, then it might be true, but all the more abstract. Indeed, nothing prevents RCTs from being used in a great variety of contexts, but then it is likely that the practices and discourses attached to this technique would be extremely different from the concrete situation under examination. However, it is false to suggest that, in the case that a war on poverty is really taking place, randomized experiments have nothing to do with it. The very fact that the scientificity of randomization is sometimes sacrificed for the sake of winning that war does not imply that RCTs are foreign to this set of power relations. There is indeed a deep homology between them both. The very existence of a treatment and a control group seems to indicate that, at least implicitly, some kind of exception to the normal decision-making process has been authorized, allowing both groups to benefit from a derogation to the measure implemented – or not – in the other. The structure of the experiment implies that one group will have to disappear at some point. Similarly, it appears then that the

experiment itself, in the war on poverty, can be seen as an exception to the usual implementation of policies on a large scale. As such, it has to be ended at some point, whether its results are reliable or not – which must be the case most of the time because, as argued by Labrousse (2010:15), all the effects of a treatment can be too long to manifest themselves during the evaluation. Finally, it is this same sense of economic emergency that allows researchers to compromise the purity of randomization when it is rejected by its likely recipients: as long as they preserve some degree of randomness, other methods can be employed. The urgency of poverty calls for exceptional measures thanks to which the war can be waged and RCTs can be subsequently employed.

Third, mass mobilization is required by this set of power relations in order to produce its effects. As indeed argued previously, the more RCTs are used, the more their external validity is ensured. Researchers have been successful in actively promoting the use of RCTs, as seen in the fact that experimental results now constitute the benchmark against which any econometric strategy is compared. Deaton reminds us, however, that the widespread popularity of RCTs has not always been the case, and that their role as methodological gold standard was recently occupied by structural models derived from well-accepted theories. Contrary to what is sometimes said, empirical economics and experimental protocols have only recently started to converge and RCTs hastened this process. For their part, poor populations are strongly invited to join the movement. Indeed, neutrality can hardly be an option. RCT results obtained from one population cannot be generalized to account for a population that refused to accept the RCT because the population that is not studied may have produced results different from those obtained with the RCT group. This is most problematic when one considers that potential differences between RCT results could be attributed to the same causes and conditions that would lead a population to refuse the RCT in the first place. Therefore, populations that resist or refuse RCTs may compromise the generalizability of results across the board, rendering policies derived from specific RCT results ineffective on a broad scale. Poor populations are therefore confronted with the choice of agreeing to the RCT, or to threatening the accuracy and applicability of their results through their non-compliance. Furthermore, some economists fear that this methodology is going to attract all available funds, leaving little left for other approaches (Labrousse, 2010). If financial resources are scarce and if RCTs are the best way to employ them, then other empirical strategies,

whether conducted by economists or other social scientists, are not only useless but potentially harmful, considering the waste of time, money and energy they may encourage. In this context, the expression “economic imperialism” – used to describe the way economics as a discipline has continued to absorb fields of research which once had nothing to do with the production and distribution of wealth – is given its full warlike meaning. The war on poverty can only be declared and waged by economists.

Fourth, there is no contradiction between mass mobilization and the exceptional derogations thanks to which this war can be waged. At first though, one may wonder why extraordinary measures require the participation of the greatest number of people. If they really were exceptional, were they not supposed to be geographically and temporally situated? Indeed, RCTs were devised for a precise analysis of local mechanisms, but only insofar as the results could possibly be generalized. In other words, the here and now quickly turns into an everywhere and forever. The war on poverty transforms exceptional interventions into a perennial fight (Duflo, 2010b:104) ; it makes the exceptions proliferate and gives them the temporal and spatial continuity they require to crystallize in an effective long-standing situation in which they become the rule. The J-PAL does not promise the eradication of poverty. But its economists seem to think that RCTs are going to put an end to the crass ignorance of poverty's mechanisms. According to Banerjee and Duflo, “if we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates.” (*Ibid*:162) In other words, external validity, which has been so far the most important caveat of RCTs, would *no longer* be an issue. One eschatology has thus been replaced by another. Instead of the discontinuity of the big push from poverty to affluence advocated by Sachs, the J-PAL promotes a continuous and *non-reversible* path out of material distress. In other words, each RCT from the list of the countless evaluations which have to be conducted heralds a small but tangible end of history: each result from a randomized experiment settles the debate forever, no matter how limited the problem was and how modest the advocated solution is. There is then no wonder why ethics committees are not reluctant to grant exemptions from full disclosure: the poor can be made blind to the fact that an experiment is ongoing because they may be less poor thereafter. They can be deprived of their autonomy because, already being less than autonomous in their state of poverty, they will become more

autonomous thereafter. The war on poverty is waged in the name of a gradual pacification of societies converging toward an everlasting peace, which is seen to be final.

4
Conclusion: the End of RCTS?

According to Duflo, the suffering of the poor does not have much to do with "systems of power."[22] In this formulation, poverty is a sometimes-deadly situation of anxiety and material distress: it has no direct link to any kind of social conflict. The close link between RCTs and the medical practice highlights the closeness of the poor and the sick in this rationale. The logic of the fight against poverty is not essentially different from the eradication of diseases.[23] In each case, a war has to be declared against an enemy in order to bring about a peaceful era of affluence and health. Sapir (1990) showed that in USSR, *shortages* of raw materials and warlike mobilization of the economy went hand in hand. Similarly in the case of the J-PAL, war can only be waged against *poverty* – and not against its potential causes, such as exploitation, neo-colonial practices... This would maybe mean that whenever a warlike architecture of power relations is functioning, scarcity becomes what has to be fiercely combated and conversely, peace becomes nothing more than material affluence. The existence of social conflicts and power relations is then overshadowed by the allegedly ongoing war. At any rate, in the case of the war on poverty, such a set of power relations produces the problem it addresses and the ways to solve it. In other words, material distress is not necessarily the natural enemy in any kind of warlike endeavor. In addition, the sense of emergency attached to this understanding of poverty may be quite monotonous: it is the same unequivocal call to fight against under-development which may arise from any war on poverty. Instead, a truly localized approach

22 http://web.mit.edu/newsoffice/2011/poor-economics-j-pal-0426.html

23 Critiques of RCTs in development economics often content themselves with the traditional arguments pointing to the weaknesses of clinical tests in medicine. But they unfortunately leave unchallenged the very analogy between these two fields.

would probably be confronted with situated and unexpected emergencies created by social conflicts, whether latent or not. Therefore, the importance the RCT methodology gives to empirical observation as opposed to *a priori* intuitions should not be overstated: all of these much-lauded, unexpected discoveries are made possible by the processes and functions of the war on poverty itself. Besides, none of them is likely to call into question the very use of RCTs, which is taken for granted. The success of randomized experiments tends toward establishing an epistemological monopoly, solidifying its own power and barring other kinds of empirical protocols.

Furthermore, waging war on poverty and the subsequent use of randomized experiments in developing countries could be perceived as being neo-colonial. First, it seems that the J-PAL and its economists are most of the time those who formulate the problems by determining which questions are asked. According to them, this is hardly an objection. Banerjee and Duflo explain indeed that RCTs "offered the possibility of moving from the role of the evaluator to the role of a coexperimenter, which included an important role in defining what gets evaluated. In other words, the researcher was now being offered the option of defining the question to be answered, thus drawing upon his knowledge of what else was known and the received theory." (*Ibid*:155) Given the fact that results from RCTs are very compelling, being a "coexperimenter" affords a considerable degree of power over the political agenda. Second, the statistics extracted during a randomized experiment cannot be used again by the populations under scrutiny. Indeed, RCTs cannot answer questions for which they were not conceived. Knowledge is produced by the experimenter, from the beginning to the end of the process. Third, coexperimentation sometimes does not suffice. In order to address issues raised by the presence of heterogeneities in the treatment effects, a new approach aims at integrating "the process of evaluation and learning into an explicit framework of program design. They therefore try to put themselves explicitly in the shoes of the policymaker who is trying to decide not only whether or not to implement a program, but also how to implement it (Should the program be compulsory? Should the administrator be given some leeway on who should participate?). (…) This literature tries to develop a theory of how the administrator should decide, taking into account both heterogeneity and uncertainty in program benefits (…)." (Banerjee and Duflo:171) At the risk of overinterpreting this new development in RCT-driven researches, economists would now be invited to engage in thought experiments

thanks to which existing political structures would be overlooked, if not purely dismissed as irrelevant. In other words, mental coups would be seriously taken into consideration to improve the results of RCTs. At any rate, the mere fact that the economists from the J-PAL do not seem to have elaborated on the potential neo-colonial undertone of their endeavor does not speak in their favor.

This methodology raises not only ethical questions but also political ones. Although it is advocated by people who think that poverty does not have a lot to do with "systems of power", it actually requires a specific way of exerting power to be put in practice. In addition, the poverty it deals with is not a mere given but is actually decisively shaped by such architecture of power relations. RCTs were welcomed as the end of autism in economics: for once, real economic phenomena were seriously taken into consideration. Unfortunately though, the mechanisms on which this methodology focuses are not as geographically and historically situated it might have seen at first. The real economic emergencies to which it tries to answer are translated into a perennial fight: the here becomes an everywhere, the now become a forever. This does not imply that rigorous evaluations of public policies are always harmful and most of the time useless. But methodologies of evaluations have to be evaluated from both a scientific and a political perspectives. RCTs, because of this strong link they have with the war on poverty, seem to structurally dodge such debates. The wartime truce thanks to which they are put in practice appears to stifle them. Most significantly, Duflo (2010b:101) concludes her book with this question: "Should we, do we have to give back the fight against poverty to the poor (…) ?"[24] Instead, it would be crucial to wonder why and how they have been deprived of the means to fight the outrageously unequal distribution of wealth. And rather than exerting once more the power of formulating problems, it would be better to start by hoping that this power will be one day, at least partially, torn from our hands.

24 "Pouvons-nous, devons-nous rendre aux pauvres la lutte contre la pauvreté, comme nous y appellent régulièrement des apôtres plus ou moins intentionnés ?"

II
THE HISTORY OF RANDOMIZED CONTROLLED TRIALS

5
The Origins of Randomization

I previously argued that there was probably some kind of connexion between the J-PAL's endeavor and the 1960s-1970s War on Poverty's reforms agenda. It is hardly enough, though, to prove that any of the RCTs run in both periods have something to do with actual war-waging, except if one is ready to dubiously give a considerable weight to all the war-related metaphors that Johnson's administration, and the reformers involved in this series of research and programs, used to describe their political action (Tobin, 1967; O'Connor, 2001:169, 178, 182; Gillette, 2010). Of course, nobody can deny the high level of international tensions with which the US had to cope at that time, and it is very likely that the launch of Spoutnik was not foreign to the refurbishment of policy-evaluation techniques which occurred in the 1960s (Monnier, 1992:12).[25] Still, the war-like architecture of power relations to which I referred earlier does not seem to frame with the way the Cold War was waged. More particularly, the mass mobilization phenomenon is likely to have more to do with the two World Wars than with the tense post-war era. Hence the need of an accurate historical investigation aiming at uncovering the origins of RCTs. Two important conclusions will be drawn: (i) this methodology was first used in psychology or in relation with themes generally associated with this discipline; (ii) American psychology bore witness to an extraordinary expansion of its prerogatives during World War II (WWII), and one might then expect that the destiny of RCTs was decided at that time. I will discuss the latter in the next section, and focus on the former now.

Interestingly, most of RCTs practitioners (or critiques), when dabbling in the history of this methodology, locate its origins either in the medical field (Bhatt, 2005), in the supposedly seminal work of Ronald Fisher (Levitt and List, 2009; Campbell and

25 Saying that the launch of the Soviet satellite into orbit was "the most crucial event for the whole field of policy evaluation" seems exaggerate though.

Stanley, 1963; Bloom, 2006), or both (Labrousse, 2010; Shadish, Cook and Campbell, 2002).[26] Before all, it is perhaps worth noting that RCTs now implemented on an international scale do not owe their existence to the postwar medical trend which gives an unprecedented role to such evaluations with, among other things, the proclamation in 1962 of the Kefauver-Harris Bill (Keel, 2011; Marks, 1999). With this new law, the Kennedy administration was ratifying what had been going on since the aborted RCTs of the 1940s and, more importantly, since the first rigorous trial that the Public Health Service ran in order to evaluate the respective effects of various treatments against tuberculosis: from then on, the Food and Drug Administration would have to approve any design setting aiming at measuring the therapeutic properties of new medications.[27] But such historical record was, at the very least, paralleled, if not prefigured, by another historical thread, a consequence of which was the implementation of the Perry Preschool Project experiment, initiated in 1962 and whose goal was to improve 3 and 4 year-old African American children's schooling conditions (see for instance Parks, 2000). In other words, social experimentations did not wait for the so-called evidence-based medicine to achieve full recognition of their potential power. Moreover and quite strikingly, while medical randomized trials were prevented from being carried on during WWII for various reasons – either because of the Army's half-heartedness in the case of gonorrhea (Marks, 1999:149) or because of clinicians' resistance to the idea of refusing treatments to those in need (*Ibid*:160) – RCTs were widely used in propaganda research (Hovland, Lumsdaine and Sheffield, 1949). All in all, if there is some kind of link between modern clinical trials and socially-concerned randomized experiments, it cannot be a linear one. This statement still applies to the interwar years during which, in a mutual ignorance of each other's activities, physicians[28] and psychologists were conducting, respectively, most of the time uncompleted and quite often rigorous evaluations. Furthermore, objecting that these experiments came after a series of medical research led in the 18th and 19th centuries and

26 This is the official historiographical stance of the J-PAL. (http://www.povertyactionlab.org/methodology/when/when-did-randomized-evaluations-begin)

27 Furthermore, a 1969 rule would make RCTs compulsory before any market authorization (Pocock, 1983). In fact, this issue is debated insofar as this rule was quite laconic as for the way randomization had to be concretely conducted. See Keel, 2011:123.

28 Keel (2011:37) mentions, among other things, the existence of a genuinely randomized experiment implemented in 1926 in the sanatorium of Northville (Michigan).

which bear some resemblance with the modern protocol of RCTs does not carry a lot of weight with respect to what this historical investigation tries to achieve. My goal here is not to document every single isolated randomized trial or their probable pioneering attempts throughout history but rather to find out when such methodology became some kind of a routine to which researchers had to appeal in order to gain scientific credibility. Of course, 19th century medicine bore witness to an increasing interest in quantification and experimentation, as argued by Jorland *et al*. (2005). But RCTs were far from being the main focus of that period. Similarly, James Lind, who is generally given credit for the invention of controlled experiments on human beings after completion of his 1747 evaluation of treatments against scurvy, published in 1753 *A Treatise of the Scurvy* (Lind, 1772) which did not call the attention of a lot of his peers (Jorland *et al*., 2005:27).[29] As a consequence, it is very likely that medical research and psychological investigations stand on two separate historical threads.[30]

A lot more attention is generally given to Ronald Fisher, his work at the Rothamsted Agricultural Station, and his two most widely cited studies, "The Arrangement of Field Experiments" (1926) and the textbook *The Design of Experiments* (1935). The bone of contention lies in the extent to which one considers or not this British statistician and biometrician as the founding father of RCTS. Oakley first held that "educational and psychology researchers", among who Thorndike and Woodworth with their 1901 study on mental functions, and Winch with his experiments on transferability of memory skills, "invented randomized assignment to experimental treatments (...) independently of, and considerably earlier than, R A Fisher's work (…)." (1998:1240) She

29 A lot less is usually said about the fact that Lind's experiment was conducted on a HMS Salisbury, patrolling home waters during the 1740-1748 War of Austrian Succession. Besides, according to the Scottish surgeon himself, this disease was the first cause of death among British soldiers, a disease more lethal then than French and Spanish weapons (Bown, 2005). Therefore, fighting scurvy was not only a medical goal, but also a military one.

30 There might be some links between the convergent interests of medicine, psychologists and social reformers in experimental settings, first on the verge of the 20th century and then at the turn of the postwar era, but those links have probably more to do with a reflexion about the virtues of controlled experiments in natural science. Such discussion is beyond the scope of this article. Furthermore, my argument does not intend to belittle the relevance of studies, like Labrousse's one, which explore criticisms formulated within the contemporary medical field against the hegemony of RCTs. It rather aims at highlighting the historical conditions of possibility of such hegemony.

then had to admit that nothing testifies in favor of the use of a rigorous randomized design in both works (Oakley, 2000b:166; quoted by Forsetlund, Chalmers and Bjørndal, 2007). Similarly, Hacking (1988) and then Dehue (1997; 2001) emphasized the decisive role played by psychical and psychophysical researches in the introduction of randomization in design settings. But, as argued by Forsetlund *et al*. (2007:372), one should be careful not to mistake "random allocation as an unbiased way of generating comparison groups when assessing the effects of interventions" for "random ordering to achieve blinding in investigations of perception" or telepathy, to which – and only to which – studies analyzed by Hacking and Dehue were crucial. Neither one nor the other make this mistake and furthermore, as argued by the latter, the former's argument implies that the transition from one kind of randomization to the other was made possible by Fisher. This is also the conclusion drawn by Forsetlund *et al*. who, thanks to a thorough analysis of articles and experiments conducted before Fisher's work and which had been falsely claimed to have used rigorous RCTs, can write that "Fisher's 1926 paper in the *Journal of the Ministry of Agriculture* is widely and correctly regarded as a landmark in the introduction of random allocation in experimental design." (*Ibid*:374)

Dehue's argument (1997) is actually more intricate and convincing than what those authors may let think. Since the 19th century, psychophysics had been concerned, among other things, with comparing the actual distance between two pins on the skin with the perceived one. Random ordering had been made necessary in order to prevent the subject from expecting, given all the already tested distances between the two points, how big the next one was going to be. But this was hardly enough because of the so-called progressive "habituation" of the subject to the exercise, and his correlative increased sensitivity. Hence the need for two distinct groups, one of which being the control, in order to get rid of this bias, but also to investigate for itself this "transfer of training phenomenon." In other words, "[w]hereas the transfer of training experiments were designed to establish the effects of an intervention, the pinprick and psychical trials had been conducted 'just' to prove or disprove the existence of particular phenomena such as the lawlike relationship of stimuli and sensations or the truth of telepathy. (…) With a slight anachronism, one could say that in the latter cases the result of a particular treatment was at issue." (*Ibid*:661) Therefore, Fisher was not the necessary link between random ordering and controlled experiment designs. Neither was he as to randomization itself. As argued by

Dehue, the Progressive Era bore witness to an increasing interest, especially among psychologists, in education, and more particularly, in how to make it more efficient. One of the examined questions of that time had to do with the importance of "formal disciplines", namely mathematics and Latin, in strengthening mental capacities. The "transfer of training phenomenon" naturally became one of the central themes researchers, among who the aforementioned psychologists Thorndike and Woodworth, started to carefully explore. The controlled group design was soon introduced in this field but without the certainty that the differences in outcomes between groups were not caused by some other uncontrolled variables. Thorndike's Ph.D. student McCall suggested, in his 1923 *How to Experiment in Education*,[31] several methods to solve this problem, one of them being chance. And even though, as recalled by Forsetlund *et al*., he clearly gave his preference to matching (*i.e.* making sure that individuals from one group are paired with the ones from the other group with respect to some key variables), he also gave such a detailed account of the way randomization can be rigorously achieved and emphasized so repeatedly the relative cheapness of this methodology that it would probably be a mistake to ignore his contribution to the history of RCTs. Finally, Dehue lays stress on the most of time ignored acquaintance of Fisher with psychological theories of his time. Indeed, the first chapter of *The Design of Experiments* is entitled "The Principles of Experimentation Illustrated by a Psycho-Physical Experiment."[32] All those arguments clearly point to psychology as the cradle of RCTs.

Furthermore, it is worth noting that Forsetlund *et al*.'s historical investigation does not refute this assertion. Using key words in various databases, they tried to give a thorough account of all the appearances of RCTs in research articles within the field of psychology from 1867 to 1948. Interestingly, the ten studies they collected (*Ibid*:377-378) deal with research themes closely related to a sub-field of psychology, analyzed by Dehue, highly interested in education, as well as to the more general issue of motivation (either encouraged by active counseling, or favored by group patterns). Indeed, Remmers (1928;

31 Interestingly, Dehue mentions that, in his introduction to his treatise, McCall assessed that better teaching methods would save the cost of $134,680,000,000,000 over the next 100 generations of Americans, which would amount to "790 times the costs of the first World War" and "390 times the costs of all wars in recorded history." (Dehue, 1997:668).

32 Nevertheless, Dehue admits that "[t]here is no evidence, however, that Fisher derived his random group design directly from psychology." (669)

1933) worked on failing students and final examinations in College, Walters (1931; 1932) on counseling, Miller and Dollard (1941) on children learning, Simon and Divine (1941) on motivation in administrations, and so on. In addition, all those studies fit remarkably well with Dorwin Cartwright's retrospective account of what interwar psychology had been concerned with.[33] He wrote in 1948 that "[i]n the last half of the 1930's there appeared in social psychology a vigorous development in the use of experimental techniques and of mathematical and statistical procedures, a development that captured the interest of many who had previously viewed *perception*, *learning*, and *motivation* [emphasis added] as the only rigorously scientific branches of psychology."[34] (Cartwright, 1948:334). Such statement casts light on the lag between the actual historical trends that can be reconstructed now and the way they were perceived in the immediate postwar period, letting one think that experimentation and quantification in psychology gathered momentum in the 1930s. Most interstingly, none of those articles mentioned Fisher's name. Once again, the central role Forsetlund *et al*. give him in the history of RCTs is called into question.[35] Simultaneously, Dehue's view according to which the increasing pervasive place of randomization in research designs was "the unplanned outcome of a lengthy historical process rather than the instantaneous creation of a single genius" (1997:655) is confirmed.

Of course, one could ask why psychology was the birthplace of RCTs. Formulating a satisfying answer is beyond the scope of this article. Nevertheless, and before moving on to WWII and its decisive role in the history of RCTs, it is worth noting that the initial success of such methodology in the sub-field of psychology interested in education might not have been coincidental. As argued by Dehue (2001:290), the emphasis of early 20th century liberalism on "creating self-supporting individuals" contributed to put the schooling system and the question of how to make it more efficient to the fore. And psychology, which was rapidly considered as the best-geared discipline to address such

33 Cartwright himself contributed to the war effort within the Division of Programs and Surveys in the U.S. Department of Agriculture. He worked more specifically on incentives to buy war bonds and the impact on German morale of bombing.

34 The ongoing changes in focus in psychological research Cartwright witnessed to will be discussed in the next section.

35 The question of the late rediscovery of Fisher's works might then be asked. Dehue (1997:670 n. 4) offers some insights.

issues, could rely on schoolchildren and female teachers' compliance with its methodological requirements. “According to Boring (1954:588), children (and rats) were the standard subjects of early controlled experiments because children, like rats, were 'inexpensive and plentiful.' […] And, most importantly, the school population – again like rats – was easy to handle. It was feasible to assign subjects to experimental or control groups and make them adhere to research protocols. The children's compliance was enforced by the teachers and the teachers' acquiescence by the powerful school management.” (*Ibid*) Dehue goes even on to hypothesize that the time-lag in the more and more systematic use of RCTs between schools (early 20th century) and the medical field (postwar era) owes a lot to the respective sociology of both professions: on the one hand, a quite feminized sector; on the other hand, a masculine one which could count on a strong tradition in favor of clinicians' discretion.[36] All in all, the political economy of RCTs seems to point to the poor, in the J-PAL's case, and more generally to the low-status individuals as their most likely subjects.

6
Waging World War II with RCTs

If RCTs were first used in psychology, and if, as argued by Herman (1995), psychologists bore witness to a considerable expansion of their discipline during WWII, then one should expect a subsequent increased importance of randomization in research designs. Now that it is certain that the first premise is true, and before examining the latter assertion, let's examine the second one. Even before Pearl Harbor, psychologists were starting to get braced for what seemed inevitable, especially at the instigation of Robert Yerkes, well known for his works on comparative psychology and primatology, and who had already served his country during WWI. This resulted in the creation in 1939 of the Emergency Committee in Psychology, within the National Research Council – founded in

36 Another likely reason for the success of psychology and its methods might lie in the turn taken by philanthropic practices in favor of a more efficient use of donations. See Dehue, 2001:291.

1916 and for which Yerkes worked as a chairman –, in order "to prepare the profession for a great national crisis." (Dallenbach, 1946:497; quoted in Herman, 1995:17) And as early as November 1940, this Committee was calling for "the mobilization of psychological knowledge having to do with problems of human engineering in times of national crisis and defense" which led to the publication of a *Psychological Bulletin* six months later entirely devoted to military psychology and addressing a vast array of issues like fatigue, propaganda or war neuroses (Sperling, 1968:98). This call for mass mobilization was widely attended. Indeed, as recalled by Herman (1995:18), "[b]efore the United States had been in the war for a year, a full 25 percent of all Americans holding graduate degrees in psychology were at work on various aspects of the military crisis, most employed full-time by the federal government." Throughout the war, the number of members affiliated to the American Psychological Association raised from 2937 in 1941 to 4173 in 1945, and this growth would not slow down in the immediate postwar period.[37] All those efforts were far from having being vain, as testified at the end of the war by Captain Lybrand Palmer Smith, navy representative to the National Defense Resource Council, who stated that "[t]he application of psychology in selecting and training men, and in guiding the design of weapons so they would fit men, did more to help win this war than any other single intellectual activity." (quoted in Herman, 1995:19) On the whole, commonsense had it, by the end of the war, that psychology had been immensely successful in guiding and assisting the war effort.

Throughout this entire period, psychologists were involved in as many research themes as the so-called civilian and military "morale", human management in internment camps, racial interrelations, the psychological foundations of democracy and authoritarianism, public opinion, smooth demobilization and so on (see the two reports respectively written by Allport and Veltfort in 1943, and Allport and Schmeidler in 1944 for an overview). According to Cartwright though, "[i]t is clear that the major scientific contributions during the war were methodological rather than theoretical." (1948:348) Indeed, WWII was a period of intense creativity as far as research designs were concerned. Opinion polls, measurement scales, intelligence testing, attitude surveys, interviews, personality analysis proliferated in order to meet the wide array of war

37 See the website of the association:
(http://www.apa.org/about/archives/membership/index.aspx)

objectives. Among them, a major role was to be played by controlled field experiments, especially in the researches conducted by the Experimental Section of the Research Branch of the Army's Information and Education Division (Lumsdaine, 1984:198). According to Lumsdaine, personally involved in this research project, "[t]he World War II experiments developed originally as extensions of, and in the milieu of, cross-section surveys of soldier opinions, directed by Samuel Stouffer. The experimental studies, under Carl Hovland's direction, extended the survey's concern with the *status* of opinion, attitude, and information, to engage the question of causative factors in producing *changes* in opinion, attitude, etc." He adds that "[s]upplementing static correlational techniques with the introduction of controlled experiments was generally regarded at the time as a significant departure in the social sciences." (*Ibid*:198) The same Stouffer, who had been Charles E. Merriam and Louis L. Thurstone's student – both committed to the promotion of behavioristic methods in the interwar period and involved in various experimental research projects (Dehue, 2001) –, explicitly made a case for rigorous evaluation in a June 1942 memorandum to the head of the Army's Information and Education Division, saying that "[t]he only certain way to demonstrate that A has the effect B is by controlled experiment." (quoted in Lumsdaine:198) In other words, and in conformity with the research area of the Experimental Section, messages delivered on radio, in newspapers as well as in films were treatments one had to carefully evaluate. However, such emphasis on controlled experiment design was not that new, unlike what Lumsdaine may let think, and is actually quite reminiscent of what Edwin G. Boring, involved himself in the war effort, was advocating in 1933: "In the simplest experiment there are always at least two terms, an independent variable and a dependent variable. The experimenter varies *a* and notes how *b* changes, or he removes *a* and see if *b* disappears. He repeats until he is satisfied that he has the generalization that *b* depends upon *a*. The independent variable, *a*, can now properly be spoken of as a *cause* of the dependent variable, *b*."[38] (Stouffer, 1933/1963:8) But never this experimental framework, had it not be invented by WWII psychologists, had been given beforehand such an unprecedented extension. And interestingly, Stouffer's own experience of the war did not make him change his mind, but on the contrary strengthened even more his convictions about the centrality of this basic principle in

38 Dehue (1997:664-665) traces such design back to Coover and Angell's 1907 study on "formal discipline".

research protocols (Stouffer, 1950). Similarly, as argued by Lumsdaine, it was not either the first time that wartime psychologists dabbled in film-based propaganda: John B. Watson – founding father of American behaviorism with his key lecture *Psychology as the Behaviorist Views it* delivered at Columbia University in 1913 – and Karl S. Lashley were the coauthors of a little-known study addressing the impact of motion pictures on WWI soldiers' morale.[39] Nor was it then the first time that psychologists considered the war as a major opportunity for the advancement of their discipline. As a consequence, the Experimental Section's main achievement was not the realization of the fact that a war would make possible the exploration of new research areas with new methodological tools, but rather, thanks to a higher level of preparedness, the promotion of a lot more systematic approach in dealing with them.

Therefore, were RCTs as central as expected? To answer this question, one may want to have a look at the overall appearances of such design in research articles and books. Given the available lexicographical tools, I devised two strategies: 1) a superficial quantification of the occurrences of RCTs in American books throughout the 20th century thanks to Google Books N-gram Viewer; 2) a closer look at the same occurrences in psychological research article databases from 1918 to 1968, date of the landmark New Jersey Income Maintenance Experiment. Unfortunately for the first strategy, among all the potential key words fitted to this kind of search (randomized controlled trial, social experimentation, field controlled experiment, random allocation, alternation, randomized controlled group and so on), only a few allow to be perfectly sure that any spotted experiment was rigorously conducted in strict conformity with the core principles of RCTs[40]: (i) "randomized controlled trial" is the generic and unambiguous term, but potentially used quite late in the 20th century; (ii) "randomized experiment", though a bit more ambiguous, captures the same idea, but without specifying if the experiment was conducted in a laboratory or in real life; (iii) "randomization" is even more ambiguous and may refer to random ordering, random sampling, or random allocation to various treatment

39 In the same vein, Sperling (1968:98) recalls that the aforementioned Thorndike and Woodworth contributed to the WWI effort. See Thorndike, 1919 for an overview.

40 The search gets even more arduous given the impossibility of using boolean operators and the case-sensitivity of the search engine.

and control groups, but is more likely to target the latter concept rather than the two other ones. In addition, note than none of those key words or expressions allow to disregard experiments conducted in the medical field. Results and general comments about N-gram Viewer are presented in Appendix A.

Not surprisingly, the expression "randomized controlled trial" started to be used in the late 1970s and the number of its appearances increased exponentially in the second half of the 1980s. Knowing that the actual technique was used earlier than that, this graph points both to a change in vocabulary and, subsequently, a growing interest in such standardized methodology. More interestingly, the occurrences of "randomization" are virtually inexistent prior to 1930, not many until the mid-1940s, and then raise continuously until the 1980s after a drop from the mid-1960s to the early 1970s. More precisely, in the mid-1960s, this term is six time more used every year, relatively to all other single words in the database, than it was in the late 1940s.[41] Again, since "randomization" has different meanings, and since it has been widely used in the medical literature at least since the 1950s, such figures are insufficient to prove that WWII and its immediate aftermath were decisive in the history of RCTs. Finally, the expression "randomized experiment", slightly more reliable than the previous term, obeys a similar pattern, with a sharper drop in the late 1960s until the early 1970s: in the mid 1960s, this expression, virtually not referred to before 1940, was five times more used than in the 1940s. Once more though, it is impossible to disentangle what is due to the medical field from what belongs to the psychological one. However, this raw overview seems to point to WWII as the origin of the more systematic use of RCTs.

The second strategy, greatly inspired from Forsetlung *et al.*, and allowing for more flexibility given the possibility of using boolean operators, yields a quite different general impression. I searched the entire PsycInfo database, from 1918 to 1968, with a variety of key words and expressions, and tried to assess, only by reading the 472 abstracts[42] I found,

41 One could be tempted to add that, for instance, in 2000, "randomization" is approximately two and half times more referred to than the expression "randomized controlled trial" is as compared to all the three word units mentioned in at least 40 books of the same chosen sub-sample published that year. *Stricto sensu*, though, those figures are not perfectly comparable. Indeed, if there are n words in the sub-sample of books published in a given year, there are necessarily $n(n-1)(n-2)$ three word units. Consequently, the denominators are not the same, making comparisons arduous.

42 Hence the relative imprecision of the method.

if the article had employed a randomized controlled trial, and if it had done so in a real life setting. Appendix B presents, in addition to more specific remarks regarding the methodology and the dataset, two sets of results. First, raw figures show a clear surge in the amount of research designs using RCTs around the early 1960s: their number is multiplied by nine in eight years. However, such increase should not eclipse the fact that, in the 1950s, every year saw the publication of at least two articles whose results rested on a RCT, as compared to the previous decades during which their use was less consistent. The number of real life settings follow a similar pattern, albeit with a smaller amplitude. Second, and in order to explore the possibility that the soaring figures of the 1960s were only driven by a parallel increase in the number of new academic journals, some filters were applied to the raw data: (i) only journals which published at least three articles in the whole dataset and (ii) whose years of publication have a standard error strictly superior to three were conserved, in order to get rid of the reviews which potentially start out in the latest dates of the sample. Interestingly, there is still a surge, naturally smaller in amplitude, but also slightly earlier in time: from 1957 to 1968, the number of articles appealing to RCTs is multiplied by three. Moreover, a close look at the six journals kept in this sub-sample did not reveal any change in the number of articles published per year, which could have accounted for this increase. Finally, the 1950s show the same kind of pattern, as compared with earlier decades, as previously. On the whole, those figures contrast with what a rawer lexicographical research could let think: the centrality of WWII in the history of RCTs turns out to be less clear, even though the 1950s seem to have been a period of a more consistent appeal to this methodology, as compared with earlier years.

However, it is very likely that RCT-based researches, if any was undertaken during the war, were not published in academic journals. Given the high mobilization of psychologists, those years were probably not the most intense period of academic publications, or at the very least, publications of experiments conducted for the sake of the war – not to mention the fact that some of them might have been classified. Indeed, it is only as late as in 1949 that the *Studies in Social Psychology in World War II: The American Soldier*, probably the most exhaustive account of what psychologists had been focusing on during the war, were published under the supervision of Stouffer. Interestingly, the collection of researches presented in the third volume, *Experiments on Mass Communication*, for the most part gathered by Lumsdaine, Hovland and Sheffield,

include a lot, if not a wide majority of controlled experiments. Out of the six main research themes presented in this book, four involved the systematic use of this methodology.

Topic	Chapters	Use of controlled experimentation (and page number where mentioned)
Effectiveness of the "Why We Fight" series on soldiers' understanding of the war objectives and their motivation	2 & 3	Yes (29-30)
Audience's subjective evaluation of films	4	No
Experimental comparison of alternative presentations (motion picture v. film-strip presentation; "commentator" radio program v. "documentary" or dramatic radio program; introductory discussion v. review)	5	Yes (123)
Effects of films on men of different intellectual ability	6	No
Short-time and long-time effects of film presentations	7	Yes (182-184)
Presenting "one side" versus "both sides" of an argument on controversial subjects	8	Yes (206)
Effectiveness of audience active participation	9	Yes (228-229)

Dehue (1997:21) is right when she recalls the methodological compromises to which psychologists had to consent when conducting their experiments. Most notably, none of the experimental design presented there randomly assigned soldiers to either the test or the control group, as bemoaned by Hovland *et al*. (1949:29). But such technique would probably have made it obvious that an experiment was going on, could then have aroused suspicions among its subjects, and consequently biased its results. Therefore, randomization was still used, but only at the level of the Army's units, and matching procedures were then applied so as to ensure that the experimental groups were similar as regards a wide array of observable characteristics. However, this hardly challenges the idea that WWII played a central role in the history of RCTs. First, and as recalled by Hovland *et al.* (1949:*v*), the studies presented in this volume "are the ones thought to be of general interest to persons concerned with the use of mass-communication methods and those engaged in research on the effectiveness of these media." For example, the Experimental Section also dealt with issues related to soldiers' physical conditioning, only alluded to in the preface (*Ibid*:*vi*). It is then very likely that among all the unpublished

studies, some employed rigorous RCTs. Second, the methodological superiority of RCTs over matching seemed well accepted by psychologists of the Experimental Section, as testified by their reluctance to employ any other experimental design. In other words, randomized trials were no longer, during the war, a statistical technique like any other. Third, researchers soon realized that "the methods used by the Army in assigning men to companies were in most cases essentially random, that is, company units were found to differ in most cases no more than would be expected in random samples of about 200 men (the typical size of a company)." (*Ibid*:251) Therefore, even if artificial randomization was not always easy to perform, psychologists could rely on the intrinsically random nature of the allocation of men to units, within the Army. Besides, the allocation of men to the Army itself, thanks to the draft procedures, was also, in some sense, random. As recalled by Hovland *et al*. (*Ibid*:15), "the studies were carried out under advantageous conditions not usually possible in film research with civilian subjects during peacetime. Although the audiences were restricted to the male population and to the age range of those eligible for military service, they had a wide range with respect to intellectual ability and various regional and socio-economic factors." In other words, the draft had created the condition of a somewhat satisfying random sampling of the American population. All in all, controlled experiments conducted by the Experimental Section would *de facto* be, at least to a certain extent, randomized. Finally, the mere use of such protocol on this unprecedented scale is in itself interesting, even if it very improbably turned out that none of the research designs employed during the war managed to impose the use of rigorous randomization. On the one hand, randomization cannot be performed if the whole sample has not been first and foremost split into test and control. If WWII psychology accustomed researchers to such practice, it can be rightfully considered as having paved the way to RCTs. On the other hand, randomization is important for my argument only insofar as it both improves the inferential properties of the experiment and consequently makes any compromises with the purity of its principles enigmatic (see first section).[43] Its exceptional dimension as well as the mass mobilization for which it calls still holds for controlled

43 However, it is very likely that controlled experiments, in which matching is substituted for randomization, provide with far more robust results than other statistical techniques, arguing in favor of its methodological superiority.

experimentation.[44] Therefore, if it happened that randomization was not performed until the end of the war, it would still be made possible by the repeated use of controlled experiments during WWII, and in that case, randomization could be interpreted as a mere artifact mimicking the natural random assignation inherent to war-waging.

All in all, never before had psychologists had the opportunity to reflect to that extent upon those experimental methodologies and the best ways to promote them. Besides, the fact that *Experiments on Mass Communication* was published in 1949, and that, as observed previously, the 1950s bore witness to a more consistent use of such design, might not be coincidental: a lot of the articles in the dataset analyzed earlier deal with issues closely related to themes central to the wartime studies like motivation (Fitch, 1951; Saltzman, 1951), or attitude change (Levine, 1952; Eriksen, 1955; Feshbach, 1957; Buss, 1958). In addition, some institutional relays might have been crucial in the diffusion of Hovland *et al.*'s results, like the Yale Attitude Change Program, founded in the early 1950s and to which the latter highly contributed. On the whole then, the two lexicographical studies and the close look at *Experiments on Mass Communication* clearly show the decisive importance of WWII in the history of RCTs: strongly recommended by the Experimental Section, they then spread out in academic research with an approximate ten year lag certainly due to the time it took, for intensively mobilized psychologists, to adapt to the new postwar conditions.[45]

Paul Lazarsfeld, who had actively collaborated for the Research Branch of the army's Morale Division, and, as recalled by Herman, "a great admirer of the [aforementioned] *American Soldier*" (76-77) asked, in 1949: "Why was a war necessary to give us the first systematic analysis of life as it really is experienced by a large sector of the population?" (Lazarsfeld, 1949:404; quoted in Herman, 1995:77). One of the answers is given by Herman, according to who: "Conveniently, soldiers' attitudes were more accessible than civilians' to both measurement and manipulation. The fact that military

44 At any rate, a close look at the War Department reports mentioned by Hovland *et al*. (1949:*viii*) would be necessary in order to gauge the relative importance of randomized experiments conducted throughout the war as compared with the studies employing different techniques.

45 Note that this overall statement is supported by various researchers, including Oakley (2000:322), Lumsdaine (1984:199), Stam, Radtk and Lubek (2000), Stouffer (1950:356), Danziger (2000:342), Insko (1967:1).

institutions exerted much more direct control over individual behavior, and therefore offered much greater support too (at least in theory), led many morale specialists to design civilian morale programs on the basis of the military model. During wartime, exerting too much control was not the biggest mistake that could be made, after all. The availability, albeit temporary, of the military total institutions was yet another benefit of war, much appreciated by researchers eager to prove the scientific validity of their experimental methods and procedures." (*Ibid*:66-67) Those two arguments, however close, must be firmly distinguished. On the one hand, it is probably true that the army turned out to be quite accommodating as for the methodological requirements psychologists had to impose in order to run their experiments. As recalled by Lumsdaine, involved in the motion picture RCTs, "it was possible to conduct the WWII studies in a nonartificial, 'real-life' atmosphere in which the film showing was perceived as a part of normal military training." (Lumsdaine, 1984:201 n. 7) But soldiers and their hierarchy also manifested some resistances to the idea of being reduced to mere guinea pigs. Indeed, Stouffer kept complaining, throughout the war, against the army's half-heartedness as to the rigorous use of RCTs (Lazarsfeld, 1949:385; Stouffer, 1950:356). On the other hand then, war provided with advantageous conditions at another level. Herman (*Ibid*:22) points to the fact that "Figures including Eli Ginzberg, Daniel Lerner, Alexander Leighton, and Samuel Stouffer referred to the military as a 'laboratory' and observed that the war presented unmatched opportunities for scientific experimentation into the mysteries of human motivation, attitudes, and behavior. They were usually careful, however, to keep such language to themselves, understandably nervous that their 'subjects' would resist being cast as rats and guinea pigs." Therefore, it is not so much because the war made possible a close collaboration between psychologists and the army as such, but rather thanks to the war itself, that the experimentalist paradigm came to the fore as the inescapable trustworthy research design. In other words, it is not only the sociological structure of the army, with which psychologists had the chance to acquaint themselves, that made possible, thanks to the high degree of control exerted over its soldiers, the rise of RCTs. *Controlled* experiments needed another kind of control, more diffuse, and very likely inherent to waging a world war. Indeed, its identification with an ideal laboratory by many psychologists shows that they were fighting on a parallel battleground, with an overall approach most of the time – but not always, as already noticed – convergent with the one

the army advocated.

Finally, it would be simplistic to mistake this parallel battleground with the realm of knowledge, or with the academic world as a sociological entity. Of course, psychologists thought that they were fighting for the advancement of evidence-based decision-making, as well as for the promotion of their discipline. And it is true, as already argued, that nothing would be quite the same for American psychology in the postwar era. But as early as 1941, Gordon Allport told his peers: "*Don't confuse lobbying for psychology with national service*:– Working for the introduction of psychologists into national and local services may be helpful to the profession, but it is not necessarily beneficial to the nation." (Allport, 1941:238; quoted by Herman, 1995:18). In other words, psychologists' main focus was and had to be the war itself. Interestingly, Stouffer made eloquent parallels between rigorous methodological principles and weapons,[46] or between potential biases and enemies.[47] To be sure, the best way to fight the latter was the former, that is, RCTs. Hence the need of a mass mobilization of psychologists in order to promote them. Hence also the need for holding sway over the military hierarchy so as to make sure that no resistance would prevent controlled experiments from being conducted, but also to select the problems whose solutions required such protocol.[48] Furthermore, those goals had to be achieved as quickly as possible. Indeed, war had created a high sense of emergency among psychologists, as still testified in Stouffer's same article.[49] This is probably why

46 "In the army no one would think of adopting a new type of weapon without trying it out exhaustively on the firing range. But a new idea about handling personnel fared very differently. The last thing anybody ever thought about was trying out the idea experimentally." (Stouffer, 1950:356)

47 "(...) there is all too often a wide-open gate through which other uncontrolled variables can march." (*Ibid*:357)

48 "Can anything be said about guides for selecting problems? I certainly thinks so. (…) Now in view of the tremendous cost in time and money of the ideal kind of strict empirical research operations, it is obvious that we cannot afford the luxury of conducting them as isolated fact-finding enterprises. Each should seek to be some sort of *experimentum crucis*, and, with rare exceptions, that will happen if we see its place *beforehand* in a more general scheme of things." (*Ibid*:359)

49 "The public expects us to deal with great problems like international peace, full employment, maximization of industrial efficiency. As pundits we can pronounce on such matters; as citizens we have a duty to be concerned with them; but as social scientists our greatest achievement now will be to provide a few small dramatic examples that hypotheses in our field can be stated operationally and tested crucially. *And we will not accomplish that by spending most of our time writing or reading papers like*

compromises with the ideal protocol, often bemoaned in *Experiments on Mass Communication*, were quite usual.

The link between the set of power relations in which the J-PAL's endeavor is made possible and the one thanks to which RCTs became a routinized methodology is now obvious. First, waging WWII required some kind of pacification of existing social conflicts within the US for which psychology would prove highly efficient. As recalled by Herman (1995:72-74), psychologists of the Research Branch for example explored some of the determinants of one of the most worrying issues with which the US had to deal with at that time, that is, interracial tensions. Those initiatives culminated with the redaction of a landmark report, *An American Dilemma*, written under the supervision of Gunnar Myrdal and with Stouffer's close support (*Ibid*:177), which would pave the way of postwar policies. In another vein, as documented by Goldin and Margo (1992), the US income distribution had never been and was never going to be again as egalitarian as it was with the advent of WWII, which suggests that a society less likely to condemn outrageous inequalities would be more willing to accept some of the war-induced sacrifices. In this context, given the unchallengeable nature of the war objectives that controlled experiments, whether randomized or not, would serve to meet, such proof production technique would contribute to make unfertile discussions vanish by basing decision-making on indisputable facts. Second, the exceptional nature of WWII would suffice to justify the use of exceptional measures, like controlled experimentation which had never been employed to that extent before. The state of emergency would allow the implementation of derogatory measures for some units of the Army, especially if those measures were suspected to significantly improve their likelihood of success. Third, and as already argued, psychologists had been strongly invited to join the fight, since the beginning of the war. Put differently, mass mobilization was the norm, including in academic circles. In the case of RCTs, this phenomenon would translate into soldiers' strongly encouraged compliance with the core principles of scientific methodologies which would be most of the time ensured by hiding them the very fact that experiments were actually being carried on in their units. The only arguable dissonance between both sets of power relations has to do with the specific nature of the eschatology attached to them: in the case of the J-PAL, gradual but irreversible ends of history; in the case of

this one [emphasis added]." (*Ibid*:361)

WWII, the widely shared hope that this war would be the last one. The beginning of the next section will actually show that such a statement is incorrect, given the fact that the clear-cut distinction between peace and war was already being blurred. As a consequence, the sets of power relations underlying the J-PAL's endeavor and the use of controlled experiments during WWII are so alike that the former can rightfully be considered as the heir of the latter.

7
Peaceful Knowledge?

If WWII was so central to the history of RCTs, how did this methodology survive the postwar era? Similarly, how did it become the cornerstone of the War on Poverty? A first answer lies in the fact that, as previously argued, the use of RCTs did not depend so much on the army, as an institution, but on the war itself, and the architecture of power relationships it implied. Hence the relative ubiquitousness of this methodology. But this does not explain how RCTs could be still used in the aftermath of what had been a crucial condition of possibility. A close look at psychologists' articles and declarations right at the beginning of the postwar era is then needed. And it appears, first of all, that, paradoxically, the restoration of peace did not seem to necessarily imply the end of war-waging. Most eloquently, Yerkes's "Recommendations Concerning Post-War Psychological Services in the Armed Services", among which emphasis was laid on the "importance of preparedness" with respect to likely conflicts to come, argued in favor of a permanent mobilization of psychologists in order to defuse potential threats against peace (quoted by Herman:79). In other words, the same psychological weapons which had proved highly effective in winning the war would now be used to win the peace. To be sure, such recommendation was not only the reflection, within the realm of psychology, of the new international tensions to the rise of which the immediate postwar era was bearing witness. Indeed, as argued by Herman (*Ibid*:30), "the new emphasis on nonmaterial determinants of military outcomes", typical of the way WWII had been waged, "blurred the distinction

between war and peace, a confusing state of affairs that would come to feel entirely normal during the Cold War." Indeed, if the newly advocated "psychological warfare", substituted to the old "propaganda", had pointed to behaviors in general as a decisive battleground, the silence of weapons would no longer be enough to ensure the conditions of a genuine peace, especially in the already tense international situation of the beginning Cold War. More generally, if psychology and its proliferating methodologies had been so importantly conceived as the most effective way of pacifying international, as well as racial and interpersonal relationships, psychologists could not be demobilized, at the end of the war, without jeopardizing the peace treaties. A much illustrative idea which gained wide circulation during WWII was to establish some kind of behavioral "weather stations", as coined by Leighton, in order to keep a constant eye on the level of tensions likely to cause future wars (*Ibid*:78). Similarly, Gardner Murphy predicted that: "Social and political psychology [would] become a psychology of social order and social control." And he added that "[t]rough the agony of these years we have learned something about the problems which confront an *international social psychology*. (…) Social psychology will have to become as international as physics. (…) The internationalization of social psychology means the internationalization of the search task of war prevention." (1945:271; quoted by Herman:80) More specifically now, if RCTs had been the least prone to infertile discussions method, in other words the most pacifying way, of waging war against wars, their abandon would have meant a regression to potential conflicts and aggressions. In the already quoted conclusion of his articles, did not Stouffer point to "international peace, full employment, [and] maximization of industrial efficiency" (Stouffer:361) as the paramount missions of experimental psychologists? All in all, the clearcut eschatology usually – or at least since WWI – attached to wars and according to which a final decisive fight was necessary to make possible the conditions of an everlasting peace was being replaced by a gradual one aiming at the progressive and, if possible, irreversible pacification of behaviors.

By saying that WWII psychology blunted the distinction between war and peace, I do not intend to play down the actual high level of international tensions the US had do deal with immediately after the signature of the peace treaties, nor do I want to indict psychologists for having created such tensions. Instead, I am arguing that psychology gave a theoretical and methodological content to this new state of affairs in which apparent

diplomatic peace could actually be threatened by latent warlike behaviors and social conflicts requiring then the use of some pacifying remedies. Interestingly, no expression summarizes better this general endeavor than the widely praised "social engineering," of which the psychologists' 1945 Peace Manifesto made an inspirational case.[50] A lot of researchers welcomed with satisfaction the fact that WWII had greatly contributed to call into question the stereotypical distinction between theory and its practical applications (Cartwright, 1948:333-334; Allport and Schmeidler, 1944:172). Yet, Stouffer, even though enthralled by the perspectives opened by this vast call for social engineering, was not as enthusiastic about the fact that social scientists had sometimes compromised their results for having mistaken some key theoretical and methodological principles with their immediate applications (Herman:75). For him, the main goal was still the lengthy and painstaking construction, thanks to systematic experimentation, of a cumulative science (Stouffer, 1950:361), a goal whose achievement WWII had sometimes had to postpone. In other words, social engineering had probably less to do with the supposedly structuring distinction between theory and practice than with a new reorganization of their relationship thanks to which experimental techniques, either confined in the laboratory or increasingly conducted in real life, would become central (Cartwright:1948). As noticed by Cartwright, an important evolution was indeed going on: "At one time many believed that groups would not permit significant experimentation upon themselves by some outsider, but experience in recent years suggests that, as social scientists demonstrate their ability to help solve the urgent problems confronted by groups, these groups will request experimental analysis of their problems and will cooperate closely in genuinely scientific experiments."[51] (*Ibid*:349) Lumsdaine retrospectively corroborated such statement, by

50 Its first principle stated that:

"**War can be avoided: War is not born in man** (sic)**; it is built into men** (sic).

"No race, nation, or social group is inevitably warlike. The frustrations and conflicting interests which lie at the root of aggressive wars can be reduced and redirected by social engineering. Men can realize their ambitions within the framework of human cooperation and can direct their aggressions against those natural obstacles that thwart them in the attainment of their goals."

Note its impregnation with the so-called "cultural lag" rhetoric, quite pervasive at that time, and according to which the maladjustment of human culture to ever-accelerating technological progress entailed overwhelming incomprehension and potential frustrations. For a similar link between social engineering and that rhetoric, see Allport, quoted by Herman:79.

51 Danziger (2000:345) interestingly drew a parallel between the increasingly use of field experiments and

arguing that WWII paved the way of "the philosophy and technology of conducting *field experiments* for assessing the effects of educational and other societal programs or innovations." (Lumsdaine, 1984:199) As a consequence, social engineering is less defined by social scientists' sudden awareness of issues they had overlooked for too long and to which they would now devote their time at the expense of a careful reading of their old theoretical treatises, but rather the attempt to pacify sometimes menacing behaviors thanks to the systematic use of experimental techniques.

Understood in those terms, it is then not surprising that such paradigm did not remain confined to psychology and rapidly gained a high level of disciplinary ubiquitousness. Since WWII itself, "Social psychologists, anthropologists, sociologists, and economists worked together in governmental agencies, and it was frequently difficult to distinguish the work of one from that of another." (Cartwright, 1948:335) Such collaboration was very likely made easy by the more general truce required to wage war and which probably made endless controversies between competing disciplines vanish. But a deeper explanation lies in the fact that this alleged interdisciplinarity, especially in the federal agencies and departments of the army whose researches rested on experimental techniques, did not revolve around much more than often standardized methodologies and theories of behavior mostly drawn from the psychological corpus. In perhaps no other case the methodological distance was the biggest between what Paul Lazarsfeld had written in the 1920s, including *Marienthal: The Sociography of an Unemployed Community*, and what he would devote himself to within the Experimental Branch to which he largely contributed. Interestingly too, Cartwright (*Ibid*:345) as well as Ernest R. Hilgard (1946) rapidly saw how potentially effective psychology could be in economic matters. The latter was indeed assuming that "(...) the problems are psychological", and he noticed that "economists make many psychological assertions in talking about them" (*Ibid*:15) but bemoaned that most of the time, they lacked the appropriate methodological guides that psychology would certainly provide.[52] On the whole, the social engineering rhetoric would

the idea that psychologists as well as politicians' main focus was now *populations*. The individual is no longer examined *per se* but only insofar as he is defined by a set of well defined variables whose values are shared by entire groups. Therefore, behavioral change is now studied not only at the individual scale but also at the level of populations. Cartwright's emphasis on groups, and their likely resistance or compliance to the use of experimental techniques, appears consistent with such a diagnosis.

52 In the same vein, Herbert Simon, who would renew, in the second half of the 20th century the economic

prove particularly pervasive in a lot of allegedly interdisciplinary studies conducted in the postwar era (Rossi and Wright, 1984:333).

As a consequence, it is now clear that the project of an "Experimenting Society", of which Donald T. Campbell made a strong case in the 1960s and early 1970s (Campbell, 1966/2002; 1969; 1973), owes for the most part its existence to WWII and its aftermath, as testified by some of its prominent characteristics:

(a) An indisputable link with psychology theory: as recalled by Dehue (2001:294), Campbell himself, whose 1966 textbook has been described as the "Bible of evaluation" (Smith, 1980:251), "started his career as an army attitude and propaganda researcher" and was later elected President of the American Psychological Association in 1975; moreover, some of the key institutions promoting a similar view as the yet to come project of an "Experimenting Society", among which the Harvard Department of Social Relations, the Center for Advanced Study in the Behavioral Sciences at Stanford, the Research Center for Group Dynamics (RCGD) at the Massachusetts Institute of Technology and the Institute of Social Research (ISR) at the University of Michigan[53] laid a strong emphasis on psychology (Herman:68).

(b) The haunting experience of WWII and the subsequent worries about likely future wars: Campbell's early career and the fact that all those institutions were created right at the end of WWII are, in themselves, quite eloquent; interesting too is the fact that James Miller's 1958 report,[54] *National Support for Behavioral Science*,

thought with the bounded rationality hypothesis, was far from being reluctant to use RCTs, as testified by his 1941 study, coauthored by Divine, on organizational improvements of administration.

53 The ISR is actually the common creation of the RCGD and the Survey Research Center (SRC), to whose foundation Angus Campbell and Rensis Likert (inventor of the Likert Scale thanks to which attitudes and motivation could efficiently be measured) greatly contributed. Interestingly, Cartwright, who took part in the establishment of the RCGD and who later worked for the ISR, referred to the SRC as a good example of psychologists' increasing interest in solving, with their own methodologies, economic problems. For the decisive influence of behaviorism on researches conducted by the SRC, see Huret, 2008:82-87. Furthermore, this same agency would pave the way of the experimental turn taken by the War on Poverty – and to which the New Jersey experiment would give momentum – with a small experiment aiming at evaluating the positive effects of family allowances (*Ibid*: 171).

54 James Miller had a position at the University of Michigan in the Mental Health Research Institute. I was

commissioned by Vice-President Nixon in order to assess the overall state of affairs of American social sciences, strongly supported behaviorism as the best theoretical weapon to counteract the communist scientific model and its potential for totalitarian thinking (Huret, 2008:29);[55] finally, it is in this context that the expression "War on Poverty" appears as more than a mere catchy metaphor.

(c) A strong commitment to experimental techniques among which RCTs stand as the irreplaceable gold standard: widely cited Campbell's textbook clearly stated that social engineering culminated in the enlightened systematic use of controlled field experiments;[56] besides, the lexicographical investigation undertaken in the previous section clearly shows the increasing role such a methodology was to play as early as in the 1960s.

(d) A gradual eschatology and its correlative narrative of a continuous pacification of behaviors: Campbell did not conceive the "Experimenting Society" as the one in which "experimental administrators" (Campbell, 1969:428) would put a definite end to all social issues but would rather pave the irreversible way toward small but tangible improvements (Fine and Saxe, 1981:15); more importantly, according to Daniel P. Moynihan (1965:12), the so-called "professionalization of reform", direct heir of the econometric revolution, would base decision-making on consensus rather than on conflict thanks to a process in which "mile-long petitions and mass rallies" (*Ibid*:15), including the ongoing Civil Rights movement, would turn out to be out-of-date.

Interestingly, such political thought would have very specific consequences in the way the War on Poverty was actually waged under Johnson and Nixon's administrations.[57]

---

not able to determine if he had any connections with ISR or SRC scholars.

55 As recalled by Huret, this report is one the offspring of the growing suspicion to which postwar social sciences bore witness. They were indeed suspected for their potential political radicalness and their repeated intrusions in people's private lives. As a consequence, debates preliminary to the creation of the National Science Foundation converged toward refocusing social sciences on their alleged scientific core, that is, behaviorism (Huret, 2008:25-30).

56 For the convergence of social engineering and the field experiment protocol, see also Campbell (1969).

57 A thorough account of the War on Poverty is beyond the scope of this article. For some insights, see Huret (2008) and O'Connor (2002).

Since its beginning, two conceptions about the tools one should use to fight poverty competed for leadership within the Office of Economic Opportunity (OEO) created in 1964: on the one hand, the advocates of Community Action Programs (CAP) and their correlative strategy of *empowerment* of the poor; on the other hand, the Research, Programming, Planning and Evaluation (RPP&E) division, favoring a more top-down approach, based on analytic methods directly inspired from the kind of decision-making the RAND corporation theorized right at the end of WWII. Interestingly though, as early as 1965, "the CAP demonstration program", thanks to which envisioned measures would be first put into practice on a small scale, "had been permanently labeled a hotbed of radicalism by critics on the outside and as a lighting rod for unwelcome controversy within the [OEO] agency." (O'Connor, 2002:171) Indeed, a previous program conducted in Syracuse had had to be ended because of the rent strikes, various sits-in and city-hall protests it had aroused, with Mayor William F. Walsh indicting the organizers, as well as the OEO, for what he described as "class warfare." (*Ibid*:171) And such cases were not isolated. As a consequence, "[i]n 1966, the legislative allocation for experimental demonstration programs was slashed to a fraction of what it had been, as a clear punishment for its most confrontational projects." (*Ibid*:172) The outbreak of the Watts riots in 1965, that some people related to CAP activism, and the escalating costs of the Vietnam War, almost buried what was left of this subdivision of the OEO, and gave to RPP&E a renewed importance.[58] "By the late 1960s, it had become standard legislative practice to require and in some instances to earmark funds for impact evaluation in authorizing legislation for educational, manpower, antipoverty, health, and other social welfare programs, and by 1972, one study reported, expenditures had reached nearly $100 million for evaluation in only four of the largest social welfare agencies. A decade later, a General Accounting Office survey reported 228 non-Defense evaluation operations within the federal government."[59] (*Ibid*:189) In less than your years, so-called analytic methods, including the increasingly used RCTs, had taken over the lead of the War on Poverty,

58 O'Connor (*Ibid*:186) mentions one staff member's description of the level of tensions between the CAP and the RPP&E branches of the OEO as being reminiscent of "East/West situations." Given the pronounced ideological overtone of the accusations one branch would make of the other's initiatives, such description cannot be only metaphorical.

59 It is then incidentally made clear that impact evaluation was still widely used in the military field, which argues in favor of a remarkable continuity with WWII methodologies.

making Campbell's (1969:2) hope that policy-making would only deal with experimentable measures come true.

However, was not this outcome in some sense predictable? As acknowledged by Sanford Kravitz, one of CAP advocates, "[t]here was a gnawing question about the capacity for a structure based on 'consensus' to work effectively for broad social change but none of us, in our euphoria over the opportunity to mount the program at a nationwide level, were really prepared to raise openly that question." (quoted by O'Connor:169) Indeed, empowerment would not be more than wishful thinking if activists were not ready to modify the intricate architecture of power relations which allowed for the existence of poverty. As early as 1965, some CAP promoters had already drawn this conclusion and were leaving the OEO. Among them was Richard Boone who then contributed to the creation of the Citizens Crusade Against Poverty which joined the Civil Rights movement in 1968 in order to organize the Poor People's March. But a deeper reason probably lies in the fact that, after all, the War on Poverty, as a war, was not meant to cast light on the conflicting substance of social relations which allowed for the existence of poverty, but rather to ensure the conditions of an effective pacification for which social engineering would turn out to be the best ally. As a consequence, the shared reference to war-waging within the OEO would turn out to be Achille's hill of CAP advocates, creating favorable conditions for disputes-free experimental methods, that is, RCTs, at the expense of the more controversial demonstration programs.[60] In other words, the warlike architecture of power relationships that the War on Poverty would try to recreate would not frame with the latter as well as with the former.

60 This is not to say that such programs did not have any drawbacks. Indeed, their focus on communities would tend, first, to isolate geographical areas from each other, overshadowing the existence of trans-regional variables affecting the reproduction of poverty; second, to restrict the logic of *empowerment* to communities themselves, making impossible a more general redistribution of power and its exercise.

8
An Everlasting Golden Age?

A lot of researchers who dabbed in the history of RCTs (Oakley, 1998:1240-1242; 2000:322-325; Monnier, 1992:12-13; Rossi and Wright, 1984) point to the 1960s-early 1980s as the Golden Age of evaluation, after which this methodology would have entered a period of relative decline and heated debates about its alleged effectiveness. The subsequent explanation which is usually given is two fold. First, RCTs turned out to be quite disappointing as a political tool. Indeed the large-scale field experiments of that period turned out to yield non statistically significant outcomes, on the basis of which no clear-cut decisions could be made. In addition, the lag between the long term temporality of research and the ever-changing political agenda made realize that time-consuming field experiments were probably not the best method to deal with urgent issues. Finally, RCTs practitioners themselves started to discover the inherent methodological issues that field experiments were unable to solve, such as the external validity of results. Second, the more and more conservative administrations, from Nixon to Reagan, proved half-hearted as for putting into practice a methodology potentially quite constraining which could force them to implement measures poles apart from their ideological stances. But a closer look at the history of RCTs does not support this overall statement and gives a significantly contrasted picture.

Nixon's election did not bring to an end the first large-scale experiments. First, as recalled by O'Connor, his administration turned out to be in favor of the philosophy behind the negative income tax, on which the New Jersey experiment launched in 1968 rested. Indeed, the Family Assistance Plan, which was to be soon implemented even before the first results of this landmark experiment were made available, explicitly obeyed this rationale. But such support given by a Republican administration to one of the measures the OEO had come to put to the fore is however not extremely surprising, knowing the fact that the idea of a negative tax originated in Milton Friedman's *Capitalism and Freedom* (1962/2002), whose rejection of a lot of measures promoted by Democrats is not disputable.[61] Moreover, no one, either in the field of research or in the agencies created

61 Did not Friedman (1953) himself bemoan the fact that economics could only rely, for its predictions, on uncontrolled experiments?

during Johnson's era, seems to have complained about the fact that Nixon administration was implementing a measure whose effects could have been first studied by the ongoing New Jersey experiment. Therefore, the Family Assistance Plan paved the way to a specific articulation of policy-making and field experiments in which the lag between their respective temporality was not a drawback but a strength: the latter would then provide with the scientific framework thanks to which the former could hastily deal with urgent issues. Second, and quite surprisingly, the prerogatives of the OEO were considerably extended. Indeed, its R&D mission would no longer be restricted to the field of poverty alleviation but would now encompass domestic policy in general. Policy-making would then be closely linked with systematic social experiments. Most famously, the New Jersey experiment was followed by a series of other large-scale experiments, aiming all at evaluating the effects of the same policy, but in different socio-economic contexts: the Rural negative income tax experiment tried to put the idea into practice in non-urban areas, the Gary income maintenance experiment focused on African Americans living in deprived neighborhoods, and finally the Denver-Seattle income maintenance experiments attempted to systematize some of the conclusions of these already conducted experiments.[62] But a lot more themes could now be systematically explored with RCTs, as testified by the thorough list of experiments collected by Boruch, McSweeny and Soderstrom (1978). All in all, Nixon's Republican administration did not constitute, at least at first, a hindrance to what the advocates of the “analytic revolution” had envisioned.

Crucial to the traditional historiography of RCTs is the alleged end of the War on Poverty in 1973 with the closure of the OEO, as well as the publication of the disheartening results of the New Jersey experiments. Near zero effects were indeed legion and their advocates were starting to wonder why so much money had been devoted to this unfertile endeavor. The mid-1970s would have then made possible the more accentuated decline of RCTs at the beginning of Reagan's era. This narrative is however severely called into question by two important facts. First, as argued by Greenberg, Shroder and Onstott (1999:161), the number of RCTs did not decrease from 1975 onwards. On the contrary, while 1,75 randomized experiments were conducted each year from 1962 to 1974, there were 7,4 from 1974 to 1982, and 5,4 from 1983 to 1996. However smaller in

62 See Allègre, 2008 for an overview.

scale, their results were more systematically used in order to design new policies (Greenberg and Robins, 1986). One of the reasons lies in the fact that 1974-1982 experiments tended to focus on specific measures rather than on the measurement of behavioral parameters relevant to a wider range of potential policies: the lesser degree of complexity of both the measures of interest and subsequently of protocols entailed a better integration of RCTs in the political process. In addition, the average length of experiments decreased from 3 to 10 years in 1962-1974 to 1 to 3 years in 1975-1982, making their results more rapidly available to policy-makers. Second, 1973 can only be a major turning point for those who only focus on federal scale driven policymaking. Indeed, the closure of the OEO, which at that time resembled more an "in-house 'think tank' for domestic policymaking" (O'Connor:192) than a decision-making federal agency, should not overshadow the existence of what O'Connor describes as a real poverty research industry, composed of an intricate series of actual think tanks and private evaluators. If 1973 knelled the death of a federally defined War on Poverty, it also gave a renewed importance to state level policies, in close relation with a more and more restricted number of RCTs practitioners, namely Abt Associates, the Manpower Demonstration Research Corporation (MDRC), and Mathematica Policy Research (MRP). In other words, the kind of extraordinary federalism to which the US had been bearing witness during Kennedy and Johnson administrations was a sufficient but not necessary condition of the widespread use of RCTs. Strongly supportive of this assertion is the fact that, until 1996, state-administered welfare reforms were required to use random allocation protocols so as to receive a federal waiver.

Finally, most historians of RCTs lay stress on Reagan administration as the burier of social experiments (Monnier, 1992:12-13; Rossi and Wright, 1984:336). Therefore, the link between progressive social reform and randomized experiments would not be questionable. But such statement is severely challenged by Greenberg and Robins' findings. Indeed, from 1983 on, the average scale of experiments, as well as their length increased. Similarly, as recalled by Levitt and List (2009:6), the 1988 Family Support Act, for which RCTs played a decisive role, made some recommendations about evaluation methods: "a demonstration project conducted (...) shall use experimental and control groups that are composed of a random sample of participants in the program." Interestingly too, forty percent of the experiments started after 1982 were mandatory, as

opposed to previous periods in which none was, according to Greenberg *et al*. (1999:161-162). Therefore, probably as many of them were aimed at evaluating "obligations on public assistance or unemployment compensation recipients in exchange for benefits." More generally speaking, RCTs have been widely employed, during Reagan's era in order to evaluate incremental changes in existing welfare provisions making the access to aids and allowances harder. Objecting that this methodology has been altered or that it is neutral in itself would not carry a lot of weight, especially if one does so in order to pit against each other 1980s experiments and the first large-scale programs, which all focused on labor incentives rather than on the actual mechanisms allowing for the existence of poverty: the New Jersey experiment already pointed to the philosophy of reforms which brought to the fore, throughout the 1980s until the 1990s, the idea of a "Workfare State". Besides, the link between RCTs and alleged "poverty alleviation" does not seem to be coincidental: "Of the 143 social experiments completed by 1996, 35 percent targeted public assistance recipients, another 14 percent looked at low-income families, and yet another 13 percent looked at the unemployed. About 12 percent of the experiments were focused on youth, and almost all of those targeted young people from low-income families. Clearly, without a strong policy interest in the poor, far fewer social experiments would be conducted." (Greenberg *et al*., 1999:159) But one could ad that the kind of interest in poverty which supposedly made possible the widespread use of RCTs is actually narrow in scope and quite dependent on the implicit approach to social issues that this methodology conveys. Furthermore, if Greenberg *et al*.'s statement is right, this would mean that the poor are more often than others subjected to experimental techniques on the basis of which – often universal – reforms are implemented. If now one reasonably assumes that poverty does not only depend on the poor's behaviors, it is then quite surprising that the "Experimenting society" made them bear the blunt of its experiments, instead of having every segment of the entire population evenly assigned to research protocols.

Therefore, the acquaintance of RCTs with neo-liberal policies might not only be coincidental. From a close analysis of the J-PAL practice, Labrousse however draws a different conclusion, suggesting a surprising parallel between Duflo's political project and German cameralism (Labrousse, 2010:20-23). However fertile, such interpretation, arguably then incomplete, is shaped by the fact that the conception of neoliberalism on

which it lies is reduced to the extreme form that it takes in American libertarianism and its correlative emphasis on self-entrepreneurship. It consequently ignores one of the most important assumptions of neoliberal thinkers and politicians according to which efficient behaviors and their not less efficient regulation by the market are intrinsically fragile, hence the need for some kind of indisputable knowledge procedures thanks to which a harmonious adjustment is made possible. Indeed, most of poverty alleviation measures tested thanks to RCTs aimed at ensuring the good functioning of labor markets in presence of some corrective welfare provisions. But more importantly, such conception of behaviors, and their potential ineffectiveness, necessarily calls for societal transformations likely to preserve, if not foster, this conflict-free harmony. Interestingly, German neoliberalism imposed itself as self-evident at the turn of WWII, when the failure of the German state in ensuring pacified social relations was made obvious. Similarly then, the rise of RCTs took place in the United States during that same war as the most effective way of, both, putting an end to these fights, and preventing any likely future disruptions of the yet to come peace. This is probably the reason why the 20th century American epitome of neoliberal government, namely Reagan's administration, did not stall on RCTs as a political tool. This is also probably the reason why the Golden Age of such methodology does not seem to have been seriously called into question.

CONCLUSION

In light of this historical survey, the J-PAL's endeavor, even though presented as a radical breakthrough in economic thinking by its most enthusiastic advocates, is not as outstandingly innovative as one could have first thought. As argued earlier, RCTs have a century-long history in which their use in developing countries is nothing more than one of the most recent stages of its evolution. For example, one of the research themes for which this methodology first proved useful is extremely close to some of the main questions explored by the J-PAL, namely education and how to make it more efficient.[63] Interesting too is the remarkable continuity of RCT practitioners' interest in behaviors and attitudinal changes, especially enhanced by WWII studies in psychology.[64] Even the institutional configuration of the J-PAL fits well with the progressive constitution, throughout the second half of the 20th century, of a poverty research industry in close relation with universities that O'Connor exhaustively documented (2002:213-241). However, qualifying J-PAL advocates' repeatedly emphasis on the novelty of their method does not suffice to criticize it and may even let some of the oddities of its historical roots go unobserved. In some sense, such emphasis must be seriously taken into consideration insofar as it seems to point to the relative young age of both RCTs and the historical foundations on which their use was made possible – as well as to the direction to follow in order to work a strategic critical argumentation out. And indeed, the historical investigation conducted in this paper clearly shows that such a technique became a routinized proof production technique only seventy years ago, in the immediate aftermath of WWII.

Consequently, there is no wonder why the J-PAL's epistemology tends to reduce

63 Which suggests that RCTs played and are still playing a crucial role in what Goldin (2001) described as the human-capital century. Exploring the links between the "high school movement," which took place in the U.S. in the early 20th century, and the first studies in educational psychology would then be interesting.

64 However beyond the scope of this article, a systematic examination of the links between behaviorism and the use of RCTs would help determine the extent to which that psychological doctrine was meant to develop this methodological tool.

politics to mere war-waging. Decisively made possible by WWII researches in psychology, RCTs necessarily convey a conservative conception of politics for which the highest priority is the gradual pacification of societies, without calling into question by any means the structure of the power relations on which they rest. Instead of encouraging the expression of conflicting views about which decisions should be collectively made, they contribute to make allegedly unfertile discussions vanish; instead of promoting political actions whose principles might be of universal value, they content themselves with forceful generalization of exceptional measures; instead of wondering how people can and, first of all, do take an active part in shaping their destiny, they strongly invite everyone to follow an already chosen journey; instead of attempting to put a definite end to poverty, they call for an indefinite fight. Interestingly, Duflo and Banerjee (2011) recently addressed an issue about which they had remained relatively silent until then – apart from some studies devoted to corruption and its harmful consequences –, that is, democracy and the best ways to ensure its existence. In light of these conclusions, and if, as expected, RCTs are appealed to play an important role in attaining such an objective, democracy would not then be fostered with genuinely democratic methods.

Throughout this paper, I implicitly defended the idea that only *political epistemology* is likely to give an overall picture of the scientific and political implications of RCT-driven researches. Given the paradoxical way the theory of randomized experiments and their actual implementation are intertwined, I drew general consequences about the conception of politics on which such paradoxes must rest. To conclude, I would like now to give some insights about the symmetrical conception of knowledge which underlies this methodology. When justifying his views, Stouffer (1950:355) described the main obstacle he had to confront in those unambiguous terms: "A basic problem-perhaps the basic problem – lies deeply imbedded in the thought – ways of our culture. This is the implicit assumption that *anybody with a little common sense and a few facts can come up at once with the correct answer on any subject* [emphasis added]. Thus the newspaper editor or columnist, faced with a column of empty space to fill with readable English in an hour, can speak with finality and authority on any social topic, however complex. He might not attempt to diagnose what is wrong with his sick cat; he would call a veterinarian. But he knows precisely what is wrong with any social institution and the remedies." What is highly important here is not so much the fact that by saying this,

Stouffer was making sure that experts would hold sway over the political agenda, than the extremely skeptical tone of his assertion, which also impregnates most of J-PAL advocates' public declarations.[65] At first, this might seem contradictory with the extremely high degree of confidence WWII psychologists and today's economists have in RCTs and their results. One may be then tempted to describe such a confidence as the manifestation of crass scientism. But if this blind faith in alleged scientific methods prevails, it is because other kinds of knowledge, especially knowledge acquired through political action and discussions, have been first and foremost ruthlessly dismissed as potentially deceptive. Therefore, it is not surprising that RCT promoters lay stress to that extent on the need for humble initiatives. And to the person who would bemoan the fortunes spent in experiments aiming at evaluating measures which are obviously beneficial to their subjects, J-PAL advocates would respond, first, that a quantitative assessment cannot be achieved thanks to mere commonsense or abstract reasoning, and second, that some of their results are sometimes counter-intuitive, such as those obtained with the deworming experiment. Nevertheless, were not simple interviews with schoolchildren and their mothers enough to discover that one of their most urgent needs was the use of treatments against parasites? And is quantification really necessary *when it is beyond question that some basic needs are to be uppermost met*?

Furthermore, the fact that such an extreme skepticism identifies the fight against poverty as its privileged outlet may not be coincidental. Can even the most skeptical person on the Earth deny the validity the bleak list of figures with which Duflo started his inaugural lecture at the Collège de France in 2009?[66] However imprecisely defined is the

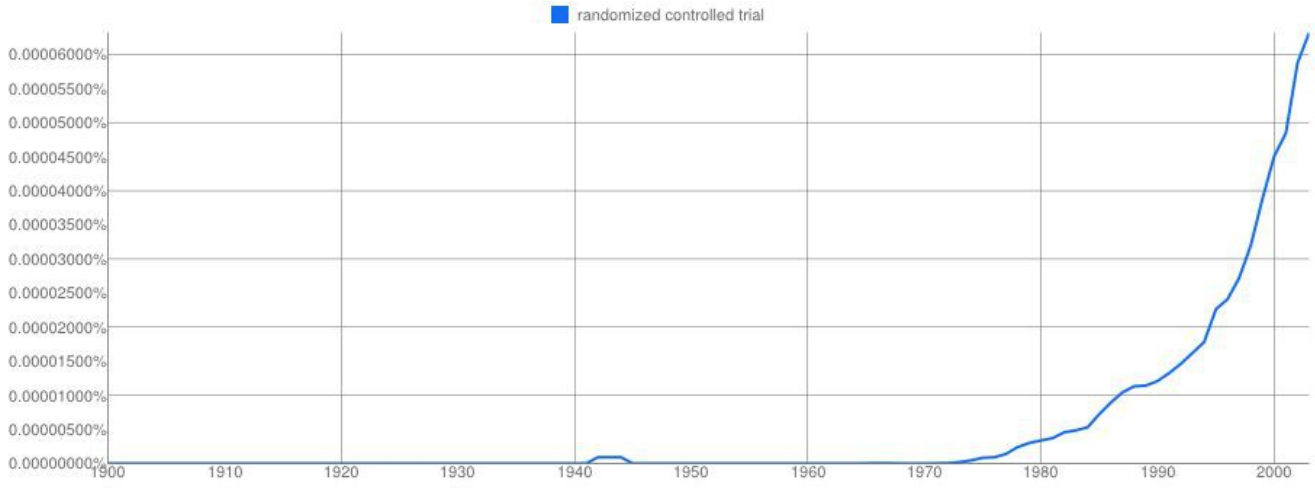
65 Does not Esther Duflo assert, at the very beginning of her inaugural lecture at the Collège de France in 2009, that "we do not have the keys to putting an end to poverty"? ("Nous ne détenons pas les clés de la fin de la pauvreté.") See this link for the recorded version of this lecture: http://www.college-de-france.fr/site/esther-duflo/experience_science_et_lutte__1.htm

66 "In 2005, 1,4 billion people were living with less than a dollar per day. Each year, at least 27 million children are not provided with basic vaccinations. 536 000 women die when pregnant, and more than 6,5 million children die before having reached their first birthday. More than half of schoolchildren in India today don't know how to read a paragraph." ("En 2005, 1,4 milliards de gens vivaient avec moins de un dollar par jour. Chaque année, au moins 27 millions d'enfants ne reçoivent pas les vaccinations essentielles. 536 000 femmes meurent en couche et plus de 6,5 millions d'enfants meurent avant leur premier anniversaire. Plus de la moitié des enfants qui sont scolarisés en Inde aujourd'hui ne savent pas lire un paragraphe.")

concept of poverty on which J-PAL practitioners work, the very first impetus of their researches is then, in itself, incredibly indisputable. Put differently, RCT advocates' extreme skepticism is meant to encounter the extreme certainty of the existence of poverty. And since such a high degree of certainty can only be supported by a gloom series of stylized facts, each singular experience of the actual ways the poor live, whose interpretation is necessarily less certain, gives rise to under-determined definitions of poverty, which are made consistent with the use of RCTs. Now, if J-PAL advocates were not as skeptical, if then other forms of knowledge (RCT subjects' opinions and theories, sociology, anthropology, political thought, philosophy and so on) were seriously taken into consideration, would poverty still be problem number one? Would not we focus on other and probably more important issues such as worldwide inequalities in wealth distribution, or land and natural resources preemption? Would not poverty appear as nothing more than the consequences of various mechanisms which can be precisely studied with a wide array of different methods, and to which an end can be collectively put?

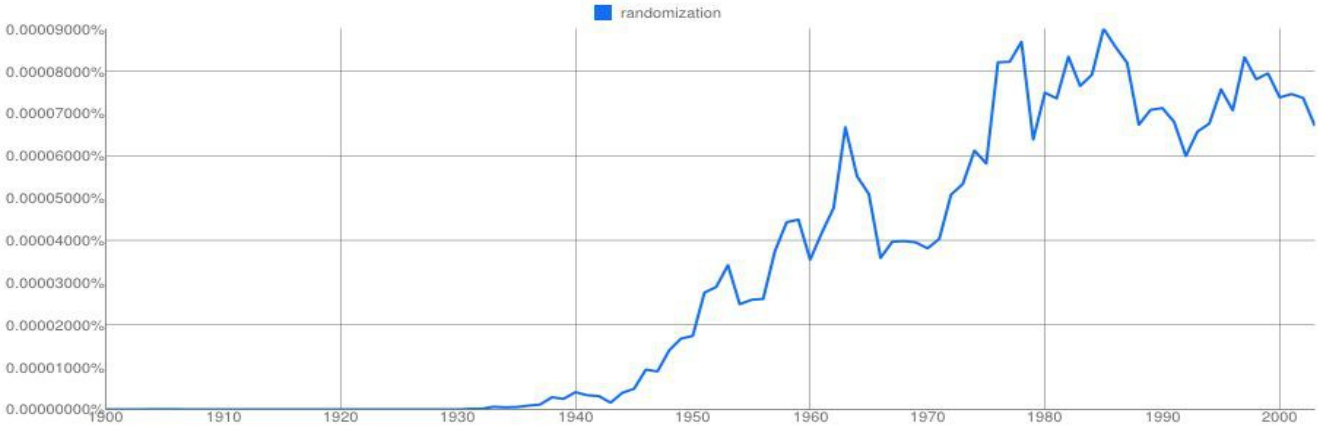
APPENDIX A – N-GRAM VIEWER RESULTS

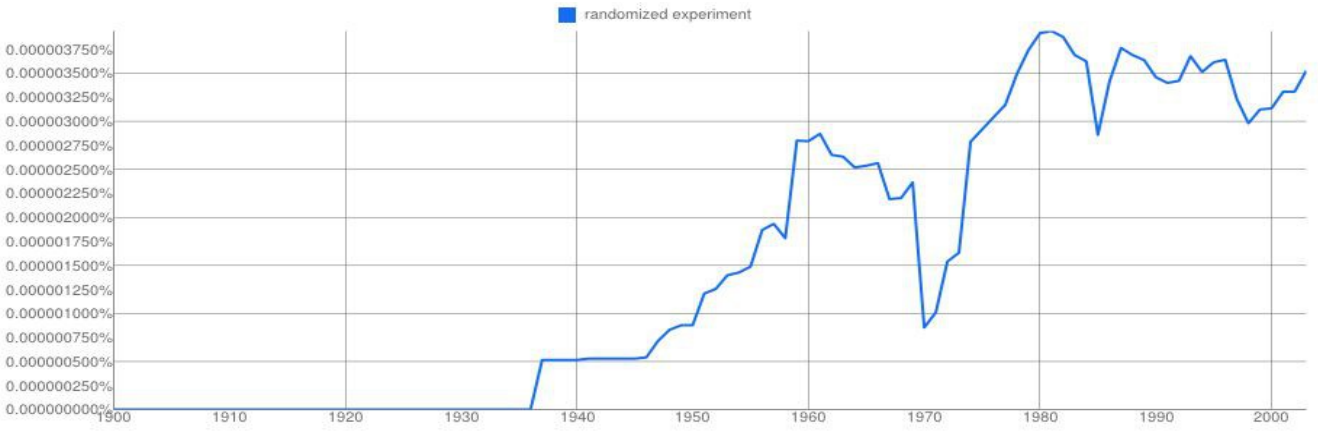
The Google Labs N-gram Viewer is a newly developed lexicographical tool which enables the search of "grams" (a single word would be counted as a "1-gram", an expression of two words, as a "2-gram", and so on) in a database of more than 5 million books (more or less 4% of all books ever published). More precisely, and in order to take into account the fact that the number of published books may vary from a year to another, N-gram Viewer computes the number of occurrences of the n-gram of interest out of all the other n-grams present in the database for each year.[67] As argued by Michel *et al*. (2011) who designed this tool, searches performed thanks to the English corpus (English and American books) from 1800 on are more likely to provide with relevant results, considering the substantial size of the database and its reliability (less typos, lesser risks of spurious digitalization...). I nevertheless restricted the present search to the American corpus, considering the fact that RCTs had first been used in American psychology. The period of interest is 1900-2003 (date of the creation of the J-PAL). A smoothing of 1 is applied to the data for the first two graphs, of 5 for the last one. Note that the search engine is unfortunately sensitive to case and plural.



67 This tool will soon perform more specific searches such as the quantification of n-grams on distinct pages and in distinct books. It might indeed be the case that the great bulk of the occurrences of a given n-gram for a given year are all located in a few books and on a few pages within them, qualifying then the raw impression of a widespread use of that word or expression.

APPENDIX A


randomization
0.00009000%
0.00008000%
0.00007000%
0.00006000%
0.00005000%
0.00004000%
0.00003000%
0.00002000%
0.00001000%
0.00000000%
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000


randomized experiment
0.000003750%
0.000003500%
0.000003250%
0.000003000%
0.000002750%
0.000002500%
0.000002250%
0.000002000%
0.000001750%
0.000001500%
0.000001250%
0.000001000%
0.000000750%
0.000000500%
0.000000250%
0.000000000%
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

APPENDIX B – PSYCINFO RESULTS

Drawing from Forsetlund *et al.*'s approach (2007), I searched the PsycInfo database, outstandingly comprehensive as far as research articles in psychology are concerned, from 1918 (end of WWI) to 1968 (launch of the New Jersey Experiment). Here is the algorithm I designed:

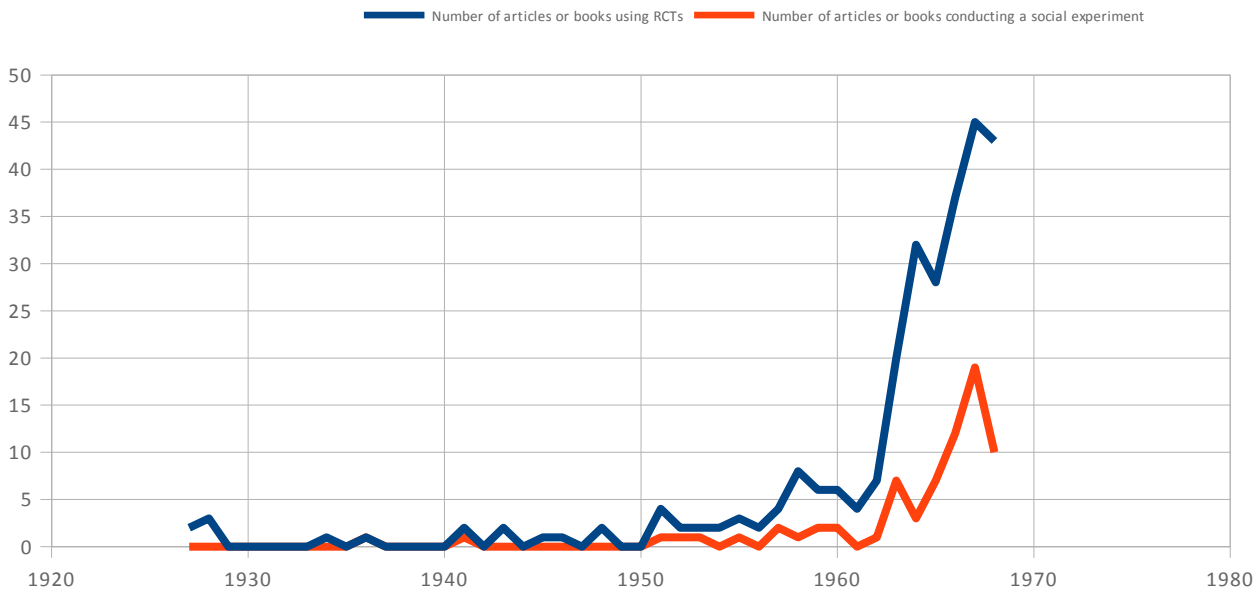
#1 random* N8 group*
#2 alternate N8 group*
#3 alternation N8 group*
#4 chance N10 group*
#5 medic*
#6 rat*
#7 mouse
#8 mice
#9 monkey*

#10 #1 OR #2 OR #3 OR #4
#11 NOT(#5 AND #6 AND #7 AND #8 AND #9)
#12 limit #10 AND #11 to
(year = 1918-1968 AND English AND population group = human)
#13 also search within the full text of articles

Where: '*' is the truncature symbol; 'A N# B' requires words A and B not be separated by more than # words.

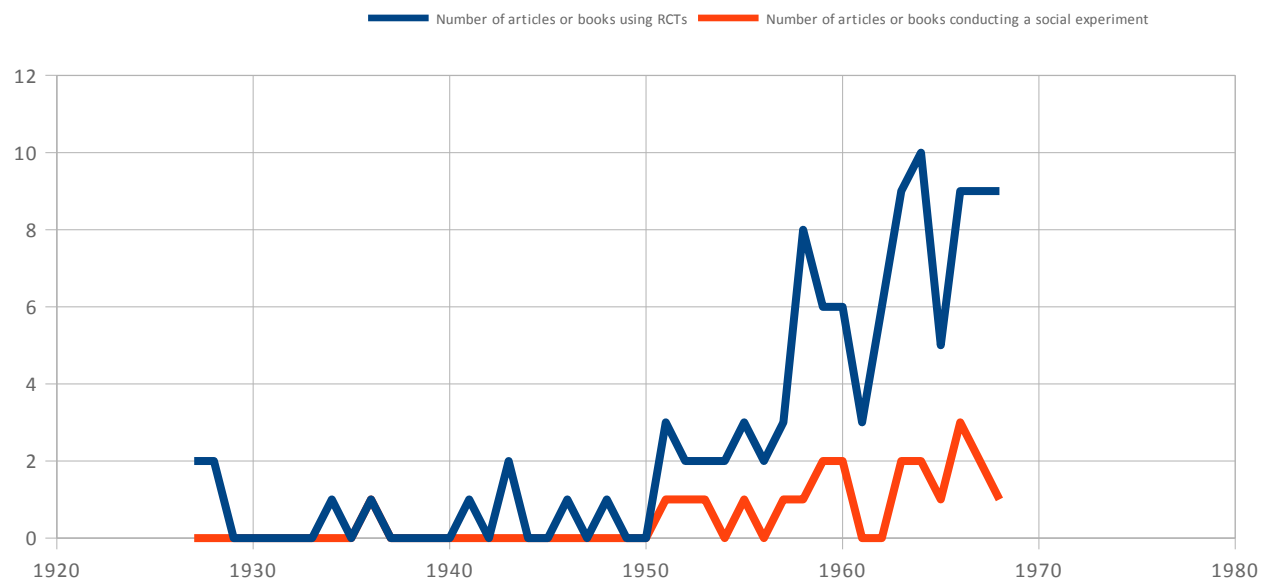
This search came up with 472 books and research articles, out of which I only retained the ones published in American journals and/or written by American researchers, and whose research subject had nothing to do with the medical field or experimentations on animals (final sample size: 406). Reviewing their abstracts, I identified the ones which explicitly randomly split their whole sample into at least two experimental groups, and then tried to distinguish experiments conducted in real life setting as opposed to laboratory

ones.[68] The first graph shows the raw results, the second one applies some aforementioned filters.



68 Unfortunately, the criteria on which such distinction rests are quite weak. Does an experiment conducted in a classroom for a short period of time fall into the first or second category? What if a similar experiment is conducted in a laboratory which mimics what would happen in the classroom? The very use of controlled experimentation on humans tends indeed to blur the usual distinction between the laboratory and real life. Therefore, I proceeded as follows: 1) an experimental design would be considered as real-life if conducted in a social context for the most part familiar to its subjects; 2) if not, it would still be considered so if the experiment may have long-lasting consequences on the subjects' lives (for instance, if the program aimed at evaluating new ways to teach how to read).

APPENDIX B



Note that the application of filters created a sub-sample in which only six journals were retained:

Journal name	First year of publication and time period of indexing in PsycInfo
Journal of Applied Psychology	1917
Journal of Consulting Psychology	1937
Journal of Counseling Psychology	1954
Journal of Educational Psychology	1910
Journal of Experimental Psychology	1916
Journal of Abnormal and Social Psychology	1906

Apart from the Journal of Counseling Psychology, none of those journals was created in the postwar era during which research publications boomed. They are then very likely to give an accurate account of research trends and fashion throughout the entire period. In addition, no break in publication volumes was observed in any of those journals between 1957 and 1958, when the surge of RCT-based researches is observed.

BIBLIOGRAPHY

Allport, F. H. & Lepkin, M. 1943, 'Building War Morale with News-Headlines', *Public Opinion Quarterly*, vol. 7, no. 2, pp. 211-221.

Allport, G. W. 1941, 'Psychological service for civilian morale', *Journal of Consulting Psychology*, vol. 5, no. 5, p. 235-235.

Allport, G. W. & Veltfort, H. R. 1943, 'Social Psychology and the Civilian War Effort', *The Journal of Social Psychology*, vol. 18, no. 1, pp. 165-233.

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*, vol. 24, no. 2, pp. 3-30.

Armatte, M. 2010, *La science économique comme ingénierie : Quantification et Modélisation*, Presses de l'Ecole des mines, Paris.

Banerjee, A. V. 2007, *Making aid work*, The MIT Press, Cambridge, MA.

Banerjee, A. V. & Duflo, E. 2009, 'The Experimental Approach to Development Economics', *Annual Review of Economics*, vol. 1, no. 1, pp. 151-178.

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

Betts, G. L. 1947, 'The Detection of Incipient Army Criminals', *Science*, vol. 106, no. 2744, pp. 93-96.

Bhatt, J. 2005, 'Causality and the Experimental Turn in Development Economics', *New School Economic Review*, p. 50-50.

Bloom, H. S. 2006, 'The core analytics of randomized experiments for social research', *New York: Manpower Demonstration Research Corporation*.

Boring, E. G. 1954, 'The nature and history of experimental control', *The American Journal of Psychology*, vol. 67, no. 4, pp. 573-589.

Boruch, R. F., McSweeny, A. J. & Soderstrom, E. J. 1978, 'Randomized field experiments for program planning, development, and evaluation', *Evaluation Review*, vol. 2, no.

4, pp. 655-695.

Bown, S. R. 2005, *Scurvy: How a Surgeon, a Mariner, and a Gentlemen Solved the Greatest Medical Mystery of the Age of Sail*, St. Martin's Press, New York City, NY.

Brearley, H. C. 1931, 'Experimental sociology in the United States', *Social Forces*, vol. 10, no. 2, pp. 196-199.

Burtless, G. 1995, 'The case for randomized field trials in economic and policy research', *The Journal of Economic Perspectives*, vol. 9, no. 2, pp. 63-84.

Buss, A. H. & Gerjuoy, I. R. 1958, 'Verbal conditioning and anxiety', *The Journal of Abnormal and Social Psychology*, vol. 57, no. 2, p. 249-249.

Campbell, D. T. 1969, 'Reforms as experiments', *American psychologist*, vol. 24, no. 4, p. 409-409.

Campbell, D. T. 1973, 'The Social Scientist as Methodological Servant of the Experimenting Society', *Policy Studies Journal*, vol. 2, no. 1, pp. 72-75.

Campbell, D. T., Cook, T. D. & Shadish, W. R. 2002, *Experimental and quasi-experimental designs for generalized causal inference*, Houghton Mifflin, Boston, MA.

Cartwright, D. 1948, 'Social psychology in the United States during the second world war', *Human Relations*, vol. 1, no. 3, pp. 333-352.

Couture, P. 2008, 'Informed Consent in Social Science', *Science Science*, vol. 322, no. 5902, p. 672-672.

Dallenbach, K. M. 1946, 'The emergency committee in psychology, National Research Council', *The American Journal of Psychology*, vol. 59, no. 4, pp. 496-582.

Moynihan, D. P. 1965, 'The Professionalization of Reform', in *National Affairs*.

Danziger, K. 2000, 'Making social psychology experimental: A conceptual history, 1920-1970', *Journal of the History of the Behavioral Sciences*, vol. 36, no. 4, pp. 329-347.

Robins, D. H. 1986, 'The Changing Role of Social Experiments in Policy Analysis',

Journal of Policy Analysis and Management, vol. 5, no. 2, pp. 340-362.

Deaton, A. S. 2009, 'Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development', *National Bureau of Economic Research*.

Dehue, T. 2001, 'Establishing the experimenting society: the historical origin of social experimentation according to the randomized controlled design', *The American journal of psychology*, pp. 283-302.

Dehue, T. 1997, 'Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design', *Isis*, pp. 653-673.

Desrosières, A. 2008, *L'argument statistique t.1, pour une sociologie historique de la quantification*, Presses de l'école des mines, Paris.

Duflo, E., Kremer, M. & Robinson, J. 2009, 'Nudging farmers to use fertilizer: evidence from Kenya', *NBER Working Paper*, vol. 15131.

Duflo, E. 2010a, *Lutter contre la pauvreté : Tome 1, Le développement humain*, Seuil, Paris.

Duflo, E. 2010b, *Lutter contre la pauvreté : Tome 2, La politique de l'autonomie*, Seuil, Paris.

Duflo, E. 2009, *Experience, science et lutte contre la pauvreté*, Fayard, Paris.

Eriksen, C. W. & Wechsler, H. 1955, 'Some effects of experimentally induced anxiety upon discrimination behavior', *The Journal of Abnormal and Social Psychology*, vol. 51, no. 3, p. 458-458.

Feshbach, S. & Singer, R. D. 1957, 'The effects of fear arousal and suppression of fear upon social perception', *The Journal of Abnormal and Social Psychology*, vol. 55, no. 3, p. 283-283.

Fisher, R. A. 1926, 'The arrangement of field experiments', *Journal of the Ministry of Agriculture of Great Britain*, vol. 33, no. 33, pp. 503-513.

Fisher, R. A. 1935, *The design of experiments*, Oliver & Boyd, Oxford, England.

Fitch, M. L., Drucker, A. J. & Norton Jr, J. A. 1951, 'Frequent testing as a motivating

factor in large lecture classes', *Journal of Educational Psychology*, vol. 42, no. 1, p. 1-1.

Forsetlund, L., Chalmers, I. & Bjorndal, A. 2007, 'When was random allocation first used to generate comparison groups in experiments to assess the effects of social interventions?', *Economics of Innovation and New Technology*, vol. 16, no. 5, pp. 371-384.

French, J. R. P. 1944, 'Organized and unorganized groups under fear and frustration', *University of Iowa Studies: Child Welfare*, vol. 20, 409, pp. 229-308.

Friedman, M. 1953, *Essays in positive economics*, University of Chicago Press, Chicago, IL.

Friedman, M. & Friedman, R. D. 2002, *Capitalism and freedom*, University of Chicago press, Chicago, IL.

Gage, N. L. 1963, *Handbook of Research on Teaching: A Project of the American Educational Research Association*, Rand McNally, Skokie, IL.

Gillette, M. L. 2010, *Launching the War on Poverty: An oral history*, Oxford University Press, Oxford.

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*, vol. 2, no. 4, pp. 213-235.

Goldin, C. 2001, 'The human-capital century and American leadership: Virtues of the past', *The Journal of Economic History*, vol. 61, no. 02, pp. 263-292.

Goldin, C. & Margo, R. A. 1992, 'The Great Compression: The Wage Structure in the United States at Mid- Century', *The Quarterly Journal of Economics*, vol. 107, no. 1, pp. 1-34.

Gomel, B. & Serverin, E. 2009, 'Expérimenter pour décider? le RSA en débat', *Document de travail du CEE*, vol. 119.

Gould, R. & Lewis, H. B. 1940, 'An experimental investigation of changes in the meaning of level of aspiration', *Journal of experimental psychology*, vol. 27, no. 4, p. 422-422.

Greenberg, D. H. & Shroder, M. 2004, *The digest of social experiments*, Urban Institute Press, Washington, DC.

Greenberg, D., Shroder, M. & Onstott, M. 1999, 'The social experiment market', *The Journal of Economic Perspectives*, vol. 13, no. 3, pp. 157-172.

Hacking, I. 1988, 'Telepathy: origins of randomization in experimental design', *Isis*, vol. 79, no. 3, pp. 427-451.

Herman, E. 1995, *The romance of American psychology. Political culture in the age of experts*, University of California Press, Berkeley, CA.

Hertzman, M. & Festinger, L. 1940, 'Shifts in explicit goals in a level of aspiration experiment', *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 439-452.

Hilgard, E. R. 1946, 'Psychological factors in the restoration of the civilian economy', *Journal of Consulting Psychology*, vol. 10, no. 1, pp. 15-22.

Hilgard, E. R., Sait, E. M. & Margaret, G. A. 1940, 'Level of aspiration as affected by relative standing in an experimental social group', *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 411-421.

Hovland, C. I., Lumsdaine, A. A. & Sheffield, F. D. 1949, *Experiments on mass communication*, Princeton University Press, Princeton, NJ.

Huret, R. 2008, *La fin de la pauvreté?: Les experts sociaux en guerre contre la pauvreté aux Etats-Unis, 1945-1974*, Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

Insko, C. A. 1967, *Theories of attitude change*, Appleton-Century-Crofts, New York City, NY.

Jorland, G., Weisz, G., Opinel, A. & Mérieux, F. M. 2005, *Body Counts: Medical Quantification in Historical and Sociological Perspective*, McGill-Queens, Canada.

Labrousse, A. 2010, 'Nouvelle économie du développement et essais cliniques randomisés: une mise en perspective d'un outil de preuve et de gouvernement', *Revue de la régulation. Capitalisme, institutions, pouvoirs*, no. 7.

Lazarsfeld, P. F. 1949, 'The American Solidier. An Expository Review', *Public Opinion Quarterly*, vol. 13, no. 3, pp. 377-404.

Levine, J. & Butler, J. 1952, 'Lecture vs. group decision in changing behavior', *Journal of Applied Psychology*, vol. 36, no. 1, p. 29-29.

Levitt, S. D. & List, J. A. 2009, 'Field experiments in economics: The past, the present, and the future', *European Economic Review*, vol. 53, no. 1, pp. 1-18.

Lewin, K., Lippitt, R. & White, R. K. 1939, 'Patterns of aggressive behavior in experimentally created "social climates"', *The Journal of Social Psychology*, vol. 10, no. 2, pp. 269-299.

Lind, J. 1772, *A treatise on the scurvy: in three parts. Containing an inquiry into the nature, causes, and cure, of that disease. Together with a critical and chronological view of what has been published on the subject*, Crowder, London.

Lippitt, R. 1940, 'An experimental study of the effect of democratic and authoritarian group atmospheres', *University of Iowa Studies: Child Welfare*.

List, J. A. 2008, 'Homo Experimentalis Evolves', *Science Science*, vol. 321, no. 5886, pp. 207-208.

Lucas, R. E. 1988, 'On the mechanics of economic development', *Journal of monetary economics*, vol. 22, no. 1, pp. 3-42.

Lumsdaine, A. A. 1984, 'Mass communication experiments in wartime and thereafter', *Social Psychology Quarterly*, vol. 47, no. 2, pp. 198-206.

Macarthur, D. M. 1968, 'Current Emphasis on the Department of Defense's Social and Behavioral Sciences Program', *American Psychologist*, vol. 23, no. 2, pp. 104-107.

MacMartin, C. & Winston, A. S. 2000, 'The rhetoric of experimental social psychology, 1930-1960: From caution to enthusiasm', *Journal of the History of the Behavioral Sciences*, vol. 36, no. 4, pp. 349-364.

Marks, H. 1999, *La Médecine des preuves: histoire et anthropologie des essais cliniques (1900-1990)*, Institut Synthélabo pour le progrès de la connaissance, Le Plessis-Robinson.

Merton, R. K. & Lazarsfeld, P. F. 1950, *Continuities in social research: studies in the scope and method of "The American soldier."*, Free Press, New York City, NY.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. & Aiden, E. L. 2011, 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science Science*, vol. 331, no. 6014, pp. 176-182.

Miguel, E. & Kremer, M. 2004, 'Worms: identifying impacts on education and health in the presence of treatment externalities', *Econometrica*, vol. 72, no. 1, pp. 159-217.

Monnier, E. 1992, *Evaluation de l'action des pouvoirs publics*, Economica, Paris.

Murphy, G. 1945, 'Human nature and enduring peace: Third yearbook for the Society for the Psychological Study of Social Issues', .

Myrdal, G. & Bok, S. 1944, *An American dilemma: The Negro problem and modern democracy*, Transaction Publishers, Piscataway, NJ.

O'Connor, A. 2002, *Poverty knowledge: Social science, social policy, and the poor in twentieth-century US history*, Princeton University Press, Princeton, NJ.

Oakley, A. 2000, 'A Historical Perspective on the Use of Randomized Trials in Social Science Settings', *Crime & Delinquency*, vol. 46, no. 3, pp. 315-329.

Oakley, A. 2000, *Experiments in Knowing: Gender and Method in the Social Sciences*, Polity Press, Cambridge.

Orcutt, G. H. & Orcutt, A. G. 1968, 'Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes', *The American Economic Review*, vol. 58, no. 4, pp. 754-772.

Orshansky, M. 1968, 'The Shape of Poverty in 1966', *Social Security Bulletin*, vol. 31, p. 3-3.

Parks, G. 2000, 'The High/Scope Perry Preschool Project', *Juvenile Justice Bulletin*, pp. 1-7.

Pocock, S. J. 1983, *Clinical trials: a practical approach*, Wiley, Hoboken, NJ.

Rodrik, D. 2008, 'The New Development Economics: We Shall Experiment, but How

Shall We Learn?', *SSRN eLibrary*.

Rosenthal, D. & Cofer, C. N. 1948, 'The effect on group performance of an indifferent and neglectful attitude shown by one group member', *Journal of Experimental Psychology*, vol. 38, no. 5, pp. 568-577.

Ross, H. 1966, *A proposal for a demonstration of new techniques in income maintenance*.

Ross, H. L. 1970, 'An experimental study of the negative income tax', *Child Welfare*.

Rossi, P. H. & Wright, J. D. 1984, 'Evaluation research: An assessment', *Annual review of sociology*, pp. 331-352.

Sachs, J. 2008, 'The end of poverty: economic possibilities for our time', *European Journal of Dental Education*, vol. 12, pp. 17-21.

Saltzman, I. J. 1951, 'Delay of reward and human verbal learning', *Journal of experimental psychology*, vol. 41, no. 6, p. 437-437.

Sapir, J. 1990, 'Economie socialiste ou économie mobilisée? Le rôle de l'économie de guerre dans la genèse et l'interprétation de l'économie soviétique', *Revista CIDOB d'afers internacionals*, no. 19, pp. 17-40.

Schmeidler, G. R. & Allport, G. W. 1944, 'Social Psychology and the Civilian war Effort: May 1943-May 1944', *The Journal of Social Psychology*, vol. 20, no. 1, pp. 145-180.

Simon, H. A. & Divine, W. R. 1941, 'Controlling Human Factors in an Administrative Experiment', *Public Administration Review*, vol. 1, no. 5, pp. 485-492.

Smith, N. L. 1980, 'The Feasibility and Desirability of Experimental Methods in Evaluation', *Evaluation and Program Planning: An International Journal*, vol. 3, no. 4, pp. 251-255.

Sperling, P. I. 1968, 'A new direction for military psychology: political psychology', *American Psychologist*, vol. 23, no. 2, p. 97-97.

Stam, H. J., Radtke, H. L. & Lubek, I. 2000, 'Strains in experimental social psychology: A textual anaylsis of the development of experimentation in social psychology', *Journal of the History of the Behavioral Sciences*, vol. 36, no. 4, pp. 365-382.

Stephan, F. F. 1948, 'History of the uses of modern sampling procedures', *Journal of the American Statistical Association*, vol. 43, no. 241, pp. 12-39.

Stouffer, S. A. 1950, 'Some Observations on Study Design', *American Journal of Sociology*, vol. 55, no. 4, pp. 355-361.

Stouffer, S. A., Lumsdaine, A. A., Lumsdaine, M. H., Williams Jr., R. M., Smith, M. B., Janis, I. L., Star, S. A. & Cottrell Jr., L. S. 1949, *The American soldier: combat and its aftermath (Studies in social psychology in World War II, Vol. 2.)*, Princeton University Press, Princeton, NJ.

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. & Williams Jr., R. M. 1949, *The American soldier: adjustment during army life (Studies in social psychology in World War II, Vol. 1.)*, Princeton University Press, Princeton, NJ.

Thorndike, E. L. 1919, 'Scientific Personnel Work in the Army', *Science Science*, vol. 49, no. 1255, pp. 53-61.

Torrance, E. P. 1959, 'An experimental evaluation of "no-pressure" influence', *Journal of Applied Psychology*, vol. 43, no. 2, p. 109-109.

Tribe, L. H. 1972, 'Policy science: Analysis or ideology?', *Philosophy & Public Affairs*, vol. 2, no. 1, pp. 66-110.

Walters, J. E. 1931, 'Seniors as Counselors', *The Journal of Higher Education*, vol. 2, no. 8, pp. 446-448.

Walters, J. E. 1932, 'Measuring effectiveness of personnel counseling', *Personnel Journal*, vol. 11, no. 000004, p. 227-227.

Williams, R. M. 1989, 'The American Soldier: An Assessment, Several Wars Later', *The Public Opinion Quarterly*, vol. 53, no. 2, pp. 155-174.

Williams, W. & Evans, J. W. 1969, 'The politics of evaluation: The case of Head Start', *The Annals of the American Academy of Political and Social Science*, vol. 385, no. 1, pp. 118-132.

Winston, A. S. & Blais, D. J. 1996, 'What Counts as an Experiment?: A Transdisciplinary Analysis of Textbooks, 1930-1970', *The American Journal of Psychology*, vol. 109, no. 4, pp. 599-616.

Woodworth, R. S. & Thorndike, E. L. 1901, 'The influence of improvement in one mental function upon the efficiency of other functions', *Psychological review*, vol. 8, no. 3, p. 247-247.

Worrall, J. 2007, 'Evidence in medicine and evidence-based medicine', *Philosophy Compass*, vol. 2, no. 6, pp. 981-1022.

Wright, M. E. 1942, 'Constructiveness of Play as Affected by Group Organization and Frustration', *Journal of Personality*, vol. 11, no. 1, pp. 40-49.

Yates, F. 1964, 'Sir Ronald Fisher and the Design of Experiments', *Biometrics*, vol. 20, no. 2, pp. 307-321.

Yerkes, R. M. 1941, 'Psychology and defense', *Proceedings of the American Philosophical Society*, vol. 84, no. 4, pp. 527-542.

Yerkes, R. M. 1946, 'Psychology in world reconstruction', *Journal of Consulting Psychology*, vol. 10, no. 1, pp. 1-7.