SciencesPo
SCHOOL OF PUBLIC AFFAIRS

PUBLIC POLICY MASTER THESIS

May 2021

The rise and fall of a gold standard

The case of Randomized Controlled Trials within the Experimentation for Youth Fund

Anne-Pauline de Cler

Master's Thesis supervised by Dominique Boullier

Second member of the Jury: Emmanuel Didier

Master in Public Policy
Economics and Public Policy

Abstract: In the economics, development and policy evaluation fields, randomized controlled trials (RCTs) have been put forward as a "gold standard" for measuring the impact of public policies. The French Experimentation for Youth Fund (FEJ) is a public-private institution that funds experimental projects and their evaluations, in view of producing knowledge about youth policy and eventually generalizing the projects. In its beginnings, the FEJ encouraged the use of RCTs and promoted them as an evaluation method to be privileged. However, their use gradually disappeared from the FEJ and they are no longer advocated as the best method to evaluate the projects it funds. Our research aims to explain the disappearance of the so-called "gold standard" that are RCTs within the FEJ. Using an approach inspired by the Science and technology studies, we study the RCTs undertaken within the FEJ as they were constructed "in the field", via the screen of their evaluation reports. From our analysis of 19 cases of RCTs, we address the various practical conditions, arrangements and contingencies that characterize them. This allows us to bring out their limitations in terms of methodology and lack of policy relevance, which partly explains their abandonment by the FEJ. In light of our findings, we recommend the use of pluralistic methods for policy evaluation, as well as greater transparency expressed by the economist community regarding the practical limitations of RCTs.

Key words: Randomized controlled trials, policy evaluation, Fonds d'Expérimentation pour la Jeunesse, Science and technology studies.

Table of Contents

1. Why should I read this research?

The Covid-19 crisis has revealed and is continuing to reveal many of our societies' inner contradictions. One of them is exposed by the omnipresence of numbers in the public discourse, who by their apparent objectivity and scientific aura provide legitimacy to the political decisions that are determined by them. An illustration of this case is the uplifting of restrictive measures made conditional upon the number of Covid cases counted at a given time. Simultaneously, these numbers are the product of other decisions and material constraints, such as the existence of PCR tests and administrative files in which to compile their results, which themselves partly depend on the arbitrary threshold upon which the PCR tests are constructed to detect a certain presence of the virus. Provided that these thresholds, themselves determined by conventions as well as material constraints, ultimately determine whether a person is going to be tested positive or negative, their nature and construction may be put into question. As one can observe, such quantified measures, which then guide political measures, are indeed the objects of virulent debates. Retaking Alain Desrosières' expression, these quantified measures are at the same time the references and objects of debate (Desrosières 1993, 7). These numbers, at the same time real and constructed, are entangled in a web of social institutions, scientific practices and political decisions. The motivation of our research lies in shedding some light on the chain of the social construction of numbers, from their production to their use by public decision-makers.

The socially constructed quantified measure that we decided to investigate in this research is that which is produced through randomized controlled trials (RCTs), used in policy evaluation. In the academic field of economics, such as in the Economics Stream of the Sciences Po School of Public Affairs, randomized controlled trials are usually tackled in the abstract, and most usually in their purely, apparently neutral, scientific dimension. Even when limits of RCTs are addressed, they remain set in the methodological realm, with causal identification always looming as an undebated value to pursue. Although RCTs essentially are *field* experiments, meaning experiments that are supposed to happen in the real world with its individual actors and social structures, they are ultimately crowned, both with the "gold standard" and the "Nobel" titles, for enabling the measure of a narrowly convened quantified impact. As such, RCTs seem to forget that they belong to a *social* science, in the double meaning referring to both their social nature and their social objects. Our aim is thus to tackle RCTs as they are realized or rather constructed in the field, beyond the unidimensional scientific screen upon which they are standardly presented. Our analysis of the Experimentation for Youth Fund, an institution that attempted to seize the acclaimed scientific and rigorous causal knowledge produced by RCTs to improve our world, but ultimately abandoned the idea, will provide us with some insights about the real, material and social resistances that the method faces.

2. Introduction

"*Creating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century, just as randomized trials revolutionized medicine during the 20th."*

Esther Duflo (2004), in *The Lancet* editorial "The World Bank is finally embracing science".

2.1. Randomized Controlled Trials or the "gold standard" for policy evaluation

For anyone who stepped into the fields of economics, development or policy evaluation, the reference of randomized controlled trials (hereafter referred to as RCTs[1]) as a "gold standard" for measuring the impact of policies has become common. Originally formalized by the statistician and geneticist Ronald Fisher, the method of RCTs was first used in the field of agriculture in the United States in the 1920s (Jatteau 2016, Desrosières 2013, Labrousse 2010). A decade later, it was largely developed in the field of clinical medicine, giving rise to the movement of Evidence Based Medicine which later influenced that of Evidence Based Policy, which characterizes the mainstream policy practices of today. In its most basic form, this experimental method consists in providing a treatment (a medicine or a policy[2]) to a randomly drawn group of people, whose outcomes are then compared to those of a control group who did not receive the treatment, in order to measure its effects. The evidence that RCTs produce thus concerns the impact of policies on individuals, or the effectiveness of policies, hence the use of the terms "measuring what works" to describe their purpose. Promoters of RCTs in the fields of economics, development and policy evaluation have, as per the above citation, a claimed affiliation to this "gold standard" for evidence in medicine (Jatteau 2019, Abdelghafour 2017, Favereau 2016, Labrousse 2010). Although large discrepancies exist in the use of RCTs between these and the medical fields, especially in terms of methodological robustness, concern about their ethical implications and reflexive positioning with regards to their limits (Abramowicz & Szafarz 2020, Jatteau 2019), such rhetoric provides scientific legitimacy to their proponents and played a role in their propagation.

Before addressing the success and criticism that RCTs encountered, let us turn to what might qualify them as a "gold standard". RCTs are most commonly known as being internally valid, in the sense that they are able to solve the problem of causal inference or of identifying a causal impact without bias. Hence their qualification as an "experimental ideal" (Angrist and Pischke 2009, 2010). Particularly dear to econometricians among social researchers (Bédécarrats et al 2019a, 1), the problem of causal inference concerns the attribution of an observed impact to a particular intervention in a given environment. As per the philosopher Nancy Cartwright (2007, 2010), who is also one of the pioneering critics of RCTs, one of the purported advantages of RCTs is also that they do not require prior theoretical knowledge to make a causal inference.

[1] For an extensive assessment of the various terms used to refer to this method, cf. Jatteau (2016, 26-30). By convention in the English language, we will use that of randomized controlled trials and its abbreviation RCTs or RCT in the singular, without excluding the use of other terms when relevant.

[2] It may be inappropriate to talk about *policies* in the context of RCTs since they usually apply to smaller-scale programs or projects that are typically not of the same scope of policies (Jatteau 2016, 46). However, we decide to stick to this general label since it is commonly used, especially in academic curricula.

They assume a probabilistic theory of causation, according to which an event A causes another event B if and only if A raises the probability of B when all other potential causes are held fixed. Ideally, RCTs do hold all such potential causes other than the policy intervention of interest fixed, because through randomization they will be distributed identically between the control and treatment groups. This last assumption rests on the probabilistic law of large numbers, according to which the characteristics of the control and the treatment group will be identical thanks to randomization *on average* and provided that the sample is large enough. Thus, on this understanding of causality, the difference in the outcomes of the control and of the treatment groups should be "caused" or attributed to the policy intervention. A similar demonstration was also developed by Rubin (2005), in his "potential outcomes" framework. Overall, this is where RCTs are known to have high internal validity, as in their ideal form they imply causality deductively.

2.2. The successful reception of RCTs in different fields and parts of the world

2.2.1. An initial entanglement with policy evaluation

Historically, RCTs made their first large-scale appearance in the contexts of policy evaluation and social experimentation that emerged in the 1960s in the United States with the Great Society programs launched by Johnson to fight poverty and racial injustice (Jatteau 2016, Fougère 2000). With the economic crises that characterized the following decade, the need for justifying public spending grew stronger. In a country particularly averse to state intervention such as the United States, the institutional practice of policy evaluation was particularly welcomed. It grew in importance as it was a means to provide proof of the efficiency of public spending (Perez 2000, 1), thereby reinforcing its legitimacy and public accountability. On a more global scale, the neoliberal turn of the 1980s anchored the movement of policy evaluation and thereby evidence-based policymaking in most industrialized countries (Devaux-Spatarakis 2014a). This reinforced the logic of rationalization of public action or that of maximization of the performance of public actors and institutions through the use of evidence and quantification (Bruno & Didier 2013). In line with the international bureaucratic movement of New public management (cf. Bezès 2020), in which public administrations imitate the performance logic of the private sector, the use of policy evaluation is favored. Besides, let us note that RCTs are a method for quantitative impact evaluation, which is only one of the various methodological forms that the practice of policy evaluation may take. Policy evaluation is thus in general addressed by a variety of disciplines of the social sciences. Anyhow, while the use of RCTs did not follow a linear progression that coincides with that of policy evaluation, the method (re)gained a significant global interest in the fields of policy evaluation and development since the beginning of the year 2000 (cf. Jatteau 2016, Devaux-Spatarakis 2014a).

2.2.2. The success of RCTs in development

In 2019, the Swedish Bank's Prize in Economic Sciences in Memory of Alfred Nobel was bestowed upon RCTs, being attributed to three of its major proponents: Esther Duflo, Abhijit Banerjee and Michael Kremer. In a speech[3] following the reception of the prize, Esther Duflo

[3] URL to the speech of Duflo in her reception of the "Nobel prize" reported by France 24: https://www.youtube.com/watch?v=Rzz0hQ8ztk8 [Accessed on April 28th, 2021].

stated that the prize was no proof of their method being the right one, but of their colleagues of the Swedish Bank *thinking* that it is the right one. She added that if there were a proof of their method being the right one, it would rest in the improvement of the lives of those affected by the policies that they study. Humility and humanity aside, this reward of RCTs comes after almost two decades of efforts from "*randomistas*" to institutionalize their experimental method on a wide-scale, and this especially in the field of development. From the end of the 1990s, the decline of the Washington consensus, whose rationale was mostly based on highly theoretical macroeconomics, left room for a new wave of development economics to develop, based on a more empirical and pragmatic approach (Jatteau 2016, 147; Pénissat 2011, 234, Labrousse 2010). Besides, while "new", this method, which was later fortified with behavioral insights, remains anchored in the neoclassical economics paradigm of the rational and autonomous individual responsible of his own choices, which coincides not only with the neoliberal ideal but with the desire to identify scientific, causal relations in the social world (Bezès 2020, Supiot 2015, Pénissat 2011, Bourgois 2010, Gautié 2007, Lordon 1997). The associated practices, based on scientific-based concrete solutions to improve the efficiency of State intervention (Labrousse 2010), were encouraged to be undertaken by international organizations (Duflo & Kremer 2008). The first to adhere to and foster this movement was the World Bank (Jatteau 2016), which hosted economists such as Banerjee and Duflo and promoted the use of their approach in developing countries to "solve" poverty. The use of RCTs in the field really plummeted since the 2003 creation of the Jamul Lateef Poverty Action Lab (J-PAL) by Duflo, Banerjee and Mullainathan, based in the Massachusetts Institute of Technology. Proof of its international dominance and success in the promotion of RCTs, the J-PAL has conducted, to date, more than a thousand RCTs in ten different policy sectors in more than eighty countries (J-PAL 2021). In comparison, estimations of the number of RCTs undertaken between the 1960s and the end of the millennium fluctuate between approximately one and three hundred on a global level (Jatteau 2016, 158-159). The countries concerned by the RCTs implemented by the J-PAL are not only developing ones, and some of the most important policy thematic they tackle are employment and education. These thematic are found in institutions similar to the J-PAL such as the What Works Centres in the United Kingdom or the Experimentation for Youth Fund in France. Overall, while RCTs are still majorly used in the field of development, they represent, at least until 2013, up to 60% of the different impact evaluations methods used around the globe (Jatteau 2016, 175).

2.2.3. The slow but sure reception of RCTs in France

Before the turn of the millennium, there were practically no RCTs undertaken in France. This goes hand in hand with the slower development of the French evaluation culture, relative to the above-mentioned case of the United States for instance. In France, contestations and distrust of State intervention are in general more limited due to the centralized nature of the State. Hence the limited calls for justifying it with tools such as policy evaluation (Jatteau 2016, 178). According to Barbier and Matyjasik (2010, 125), evaluation originated at the end of the 1970s in France, and progressively installed itself in the political field in the 1990s. To explain such rhythm in the development of policy evaluation in France, Nioche (1982 43, 46) talks about three types of obstacles: sociopolitical, administrative and methodological. The first refers to the fact that there were few evaluators in France at the time, and that national administrative

institutions such as the INSEE played a monopolistic role in the production of national statistics, thereby dissuading a similar activity by other actors such as evaluators. Administratively, evaluation may have been thought of as unnecessary or repetitive in light of the already existing administrative, political and legal controls in France such as those exerted by the national audit court (*Cour des Comptes*). Lastly, there were some obstacles to the use of evaluation methods such as randomized ones in particular due to ethical problems, which we will address further. Overall, while the country was initially quite reticent to the use of policy evaluation, these obstacles were eventually surmounted and the context of the early 2000s nevertheless became favorable to the implementation of RCTs on the territory.

Indeed, while RCTs started gaining interest in France at the very end of the 20th and at the very beginning of 21st centuries, notably through the works of Denis Fougère (2000) and Éric Maurin (2002) in the fields of social aid and employment, their use really plumetted as soon as a J-PAL branch was implanted at the Paris School of Economics in 2008 (Jatteau 2016, 177). According to Jatteau (*ibid.*), this factor played a significant role in developing the use of RCTs in France, as it provided them with an institutional framework as well as an academic community anchorage. Another factor that is said to have influenced the prominence of RCTs in France is Martin Hirsch's accession to government, from respectively 2007 and 2009 to 2010 as high-commissioner in the fields of poverty and youth ("*haut-commissaire aux Solidarités actives contre la pauvreté"* and "*haut-commissaire à la Jeunesse"*). He played an important role in the financing of RCTs, in particular through the institution of the Experimentation for Youth Fund that he presided, and in general held a favorable discourse towards experimentation and privileging the use of RCTs (Jatteau 2016, 178, 182; Pénissat 2011, 239). For instance, he is at the origin of the Active solidarity income (*Revenu de solidarité active* or RSA). It was one of the first French policy programs to be subject to an experimentation, between 2007 and 2009, in order to evaluate its effectiveness in making the unemployed return to employment (Penissat 2011, 223), and ultimately opening the field of social experimentation in France. Additionally, the French context was also shaped by the general neoliberal and New public management movement starting in the 1980s. For instance, the increasing importance of the benchmarking practice in both the public and the private sectors (Bruno & Didier 2013) and the 2006 Law of public finance (*Loi Organique Relative aux Lois de Finances*) (Jatteau 2016, 180), both show a growing requirement not only to conduct quantitative evaluations, but to make any form of public action depend on numbers.

2.3. The Experimentation for Youth Fund, a determining institution in the use of RCTs in France

2.3.1. The institutional structure of a public policy lab that puts experimentation at the service of youth policy

The Experimentation for Youth Fund (in French the *Fonds d'Expérimentation pour la Jeunesse*, hereafter referred to as the FEJ) is self-referred to as a public policy lab that puts social experimentation at the service of youth policy (FEJ, 2019). It was created by the 25th article of

the law of December 1st, 2008[4], in order to "fund experimental programs aimed at promoting student success, contributing to equal opportunities and improving the sustainable social and professional integration of young people below the age of twenty-five.[5]". As such, the FEJ is the first large institution dedicated to experimentation in France (Kerivel 2017, 3).

At its start, the FEJ's budget was of 230 million euros (FEJ 2010, 5), composed of a 150 million euros subsidy from the State, a 50 million euros funding by Total, and more modest contributions from other private firms (Devaux-Spatarakis 2014a, 294). Between 2009 and 2013, 24 million euros of this budget were dedicated to evaluation (Valdenaire 2013). As of 2019, the FEJ has made 27 project calls, received 6 717 applications, supported 959 projects that have benefitted three million young beneficiaries, and has overall received 260 million euros of funding (FEJ 2019). To put it simply, the FEJ is a public-private institution that funds experimental projects and their evaluations with the objective to produce evidence about the success of such projects before eventually generalizing them on a greater scale. The FEJ does so by launching national project calls directed to any public or private structure that is willing to propose an innovative action or to reform an existing policy device to make it more efficient. It thereby pairs a project-holding structure with an evaluation team, or in parallel an innovative project with an evaluation protocol, which respectively constitute, as per the FEJ's terminology, "the experimenters" and the "experimentation" (FEJ 2019). Among the eligible project holders are any structure that may receive a State subsidy, such as public educational institutions, territorial collectivities, NGOs, or local missions. As for the evaluation team, it can be any public or private structure that demonstrates some expertise in terms of qualitative, quantitative or mixed evaluation, such as research centers, teams affiliated to universities, or private firms. One of the essential characteristics of the FEJ is that it promotes independent evaluations, that are to be implemented from the start of projects and throughout their course in order to eventually generalize them.

In terms of selection of the experimentations to be funded, the process is two-fold. First, the interested project holders are invited to apply to the FEJ by presenting a pre-project. The pre-selected project holders then have to present a more definitive experimentation project. In parallel, the FEJ directs a project call to potential evaluators, who are selected based on their evaluating competencies. The FEJ aims for the evaluation structure to be truly independent and external from the project holders, in order to ensure the objectivity of the evaluation team with regards to the project. In the second phase, the selected project holders have to partner with an evaluation team, to elaborate together a detailed experimentation project, including the content, partners and implementation of the experimented device, as well as the development of an evaluation protocol. The final selection is then made upon this combined application file (FEJ 2010b).

The FEJ's governance is ensured by two main entities, a Management Board and a Scientific Board:

---

[4] Article 25 of the law of December 1st, 2008 URL: https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000019860428#LEGIARTI000023371495 [Accessed on April 23rd, 2021].

[5] As defined by the decree 2011-1603 of November 21st, 2011. URL: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000024831776/ [Accessed on April 23rd, 2021].

- **The Management Board** is the decision-making organ of the FEJ. Its missions are to define the priority thematic and decide of the allocation of funds for the projects it selects. It can designate a jury composed of actors in the experimentation field or in the Scientific Board to advise it in terms of project selection. It is presided by the minister in charge of youth, so currently, in April 2021, Jean-Michel Blanquer. It is composed of the central administrations concerned by the FEJ, such as those of employment, national education, higher education, and so on, as well as of the private partners involved in the Fund, namely the TOTAL foundation and the Union of Industries and Metallurgy Professions (*Union des Industries et des Métiers de la Métallurgie*) (FEJ 2018).
- **The Scientific Board** defines the modalities of evaluation of the different thematic programs. Its missions are to judge and issue methodological recommendations regarding the evaluations undertaken and to annually provide its opinion on their scientific quality and relevance. Since April 2017, this mission is ensured by the scientific and orientation board of the National Institute for Youth and Popular Education (*Institution national de la jeunesse et de l'éducation populaire*, INJEP). As of April 2021, it is co-presided by Philippe Da Costa, deputy general director in charge of external relations (head of the "Partner College") and doctor in education sciences, and by Yaël Brinbaum, head of the "Scientific College" and doctor and lecturer (*maître de conférence*) in sociology. It is composed of experts in evaluation and/or in youth policy (FEJ 2018).

The **animation of the FEJ** is ensured by a team of the INJEP, the Mission "Experimentation and Evaluation of Public Policies" (MEEP), affiliated to the National Education ministry. It is composed of two poles: one for the monitoring and animation of projects, and another for evaluation and capitalization. The INJEP, as a knowledge-producing observatory, is a resource and expertise center on questions relating to the youth and youth policies. Its mission is to contribute to improving the knowledge on such domains through the production of statistics and analyses, observation, experimentation and evaluation (FEJ 2018). Through its animation of the FEJ, the INJEP accompanies the initiatives of actors in the field by following the FEJ's experimentations, giving them visibility and capitalizing good practices from its evaluations. Its objective is thereby to study the effects of projects and to orient public policies towards the most efficient policy devices. Overall, it aims to contribute to the generalization and spread of good practices by working close to youth and popular education professionals (FEJ 2018).

2.3.2. Terminological specification of the notion of "experimentation": its meaning within the FEJ and its relation to randomization

The term "*expérimentation*" in French has several meanings, and is usually translated in English as "experiment", "experimentation" or "testing" (Kerivel 2017, 2). In its most recent evaluation-related methodological guide, the FEJ states that it abides to a definition of experimentation that is large and that includes a political dimension rather than an exclusively scientific one. Accordingly, the notion of experimentation includes that of experiment or experience (in French "*expérience*"), or the "small-scale implementation of a device and the opportunity to conduct its evaluation." (Kerivel 2017, 2. [Own translation]). The guide adds

that "on a political perspective, the objective of an experimentation is to measure the effects of a device and its conditions of implementation in order to spread it ["*essaimer*"] in whole or in part." (*Ibid.*). In a methodological guide that lasts from the beginnings of the Fund, the notion of experimentation is also constituted by that of evaluation. Precisely, it defines a social experimentation as "a social policy innovation initiated at first on a small scale, which takes into account the existing uncertainties about its effects and is implemented in conditions that allow evaluating its results, with a view of generalizing them if they prove convincing.[6]" (Conseil scientifique du FEJ 2009, 2). [Own translation]). As such, the notion of experimentation, as understood by the FEJ, refers not only to innovative social projects and their evaluation, but to the integration of these two elements. The FEJ accordingly emphasizes that one of its originalities is the promotion of social experimentations in which evaluation is a constitutive part of the implementation of small-scale innovative social projects, in view of eventually spreading or generalizing them (FEJ, 2019).

When expressions such as "scientific evaluation protocol" (Kerivel 2017, 2) or "meeting demanding scientific requirements" (FEJ, 2019) are used to qualify the experimentations fostered by the FEJ, one may wonder what their scientific dimension may be. As it turns out, it is closely tied to the notion of randomization. While the 2009 methodological guide mentions it first when describing evaluation methods (Conseil scientifique du FEJ 2009, 5), the latest one mentions it directly when defining that of experimentation. Indeed, it provides a dictionary definition of experimentations according to which they literally are a "scientific method based on experiment and controlled observation to verify hypotheses" (Kerivel 2017, 2. [Own translation]). More than that, it adds that "in a scientific acceptance", they consist in constructing a counterfactual allowing for the measurement of a causal relation (Kerivel 2017, 2; Conseil scientifique du FEJ 2009, 5). Understood in this way, it is said that an experimentation goes through the random draw of a population that will benefit from a device and another that will not (Kerivel 2017, 2), which precisely corresponds to the implementation of an RCT. As concrete examples that illustrate this "scientific definition", the methodological guide mentions the development of social experimentations under the Reagan government in the United States in 1980, as well as the introduction in France of such experimentations through the J-PAL (*ibid.*). As stated above, the FEJ prefers a definition of experimentation that has a broader and political scope, namely one that "accounts for temporalities and promotes the emergence of devices" (*ibid.*). However, the scientific definition is said to be regularly mobilized when the method of random or quasi-random experimentation is used (*ibid.*). As such, the FEJ has no intention whatsoever to reduce its experimentations to those characterized by randomization. Nevertheless, these considerations hint two elements that will receive further inquiry: first, that the random experimentation method, or in short that of RCTs, has a certain valued scientific aspect; and second, that randomization, insofar as it is involved in what the FEJ calls an experimentation, would form a constituent part of the device or socially innovative project that is to be also implemented as part of the experimentation.

[6] The terms "social policy innovation" and "device" can therefore be said to have the same meaning in that context. Here, the term device can be understood as small-scale policy interventions.

2.3.3. The FEJ's use of RCTs and positioning towards the method

As suggested above, the FEJ played an essential role in initiating the use of RCTs in France, via the funding that it provided to them as well as through the promoting discourse of its first president Martin Hirsch, strong proponent of the method (Jatteau 2016, 182). In its beginnings, the FEJ's Scientific Board was presided by Marc Gurgand, economist and currently scientific director of the J-PAL branch at the Paris School of Economics, and was composed of other evaluation experts known to have promoted the use RCTs (Devaux-Spatarakis 2014a). The FEJ was therefore initially composed of actors that both promoted the use of RCTs and played an essential role in influencing the evaluations that the FEJ would fund, in virtue of their positions in the governance structure of the FEJ. In this line, the FEJ's first methodological guide, drafted by the Scientific Board and addressed to evaluators, does reflect a clear intention to privilege the use of RCTs as an evaluation method. Indeed, RCTs are put forward in it, among quantitative evaluation methods, as the most reliable one, allowing the most precise and robust measurement of the impact of a program on variables of interest (Conseil scientifique du FEJ 2009, 5). It also states that RCTs are the best means to "constitute identical groups in all points, before the experimentation, which allows measuring its effect 'all else equal'." (*Ibid.* [Own translation]). In light of the description of the method given above, this seems to be an over-statement, if not simply an erroneous one. Even though this same guide previously mentions the complementarity of quantitative and qualitative evaluation methods, describing their respective purposes, it nevertheless ultimately stresses the superiority of RCTs as an evaluation method in general. Indeed, the intention is quite clear: "It is always desirable to favor an evaluation based on a randomized experiment protocol whenever the experimentation's conditions allow for it." (*Ibid.* [Own translation]). Moreover, several of the first activity reports of the FEJ, notably the 2010 report of the Scientific Board, confirm that the FEJ insisted on the use of RCTs as an evaluation method, noting its political support and putting forward its ability to produce "precise and transparent knowledge" (FEJ 2010a, 8). It also notes that it contributed to the development of the method at a time when it was "under-used" in France (*ibid*.). Interestingly, the criteria of selection of the experimentations, as they appear on the website still un-updated since 2010, also seem to favor the selection of projects that may be evaluated through a randomized protocol. Indeed, according to them the quality of the evaluation protocol is majorly judged upon its ability to precisely measure the effects or the effectiveness of the experimentation in question, with a rigorous determination of a control group to be compared to a treatment one (FEJ 2010b).

As such, the FEJ did initially promote RCTs as a privileged method for the evaluation of the experimentations that it funds. However, this does not mean that the FEJ exclusively financed RCTs. Indeed, since the very start, RCTs only appear to be a minority among the evaluations used in the experimentations undertaken within the FEJ. As presented in the 2010 activity report of the Fund, these only sum up to approximately 37 among the more than 400 projects initially selected (FEJ 2010a, 15). This report even talks about the fact that some of the randomized evaluation protocols were abandoned along the way, suspecting that these were only presented in the application file of the experimenters to receive funding, in spite of the "high relevance" that the method was claimed to have in these cases (FEJ 2010a, 15). Actually, purely quantitative evaluation methods always seem to be used in minority, summing up to 6% among

the different types of evaluation methods used in the experimentations funded by the FEJ until 2012 (FEJ 2012, 29). When counting both purely quantitative evaluation methods and mixed methods with a quantitative dominance, as per the typology defined by the FEJ, these sum up to 48% of the overall evaluations conducted (FEJ 2012, 29). In spite of the lack of consistency and transparency regarding this typology among the various activity reports, the FEJ notes in 2012 that 37% of the evaluated projects used an RCT, and the financing of such evaluations and their projects amount to 49% of the FEJ's funds, given the often "above-average size of the method" (FEJ 2012, 29). As such, we can note that the RCTs undertaken within the FEJ, in spite of being a minority among the different evaluation methods used, consisted in a major proportion of its budget.

While signs of the failure of RCTs among the FEJ are already apparent since its beginnings, these are only reinforced by the fact that most recently, RCTs seem to have completely disappeared from both the FEJ's methodological guidance and its activity reports. Indeed, in the activity reports following 2017, no mention of RCTs is made. Moreover, in the most recent methodological guide drafted by Aude Kérivel, doctor in sociology and current head of the INJEP team that animates the FEJ, RCTs are simply not put forward as the evaluation method to use whenever possible, as opposed to the previous guidance. Rather, the position of the FEJ seems to now truly favor a use of evaluation methods that is conditioned upon the nature of the project and the questions they ask, which all but prescribes the exclusive use of RCTs. Indeed, the most recent methodological guidance goes in the direction of a diversification and combination of methods, in order for them to be truly adapted to their context as well as to avoid being disconnected from reality. Actually, even if this does not coincide with the information provided in most activity reports and on the FEJ's website, this 2017 methodological guidance mentions that the generalization of experimentations is not a priority (Kérivel 2017, 4).

2.4. The disappearance of RCTs among the FEJ: a puzzle

As such, while RCTs were initially promoted by the FEJ as a favored evaluation method to use in its experimentations, we notice that first, they never were successfully used, and second, that the FEJ eventually abandoned promoting them as such. In light of the general attribution of RCTs as a "gold standard" for policy evaluation and the broader success that the method seems to have met, at least in its use, we may wonder why RCTs have failed within the FEJ.

The puzzle that we are facing is therefore two-fold: first, the FEJ seems to ultimately resist the general uptake of the "gold standard" for policy evaluation, and second, it seems to have gone against its own initial promotion of RCTs. Our problematic thus puts into question the status of RCTs as a "gold standard", not only per se but insofar as they are carried out in a particular institutional framework.

2.5. Research question, motivation, hypotheses and scope of research

2.5.1. Research question

Our main research question is therefore the following: why has the FEJ abandoned the so-called "gold standard" for policy evaluation, RCTs, when it itself initially advertised them as such and encouraged their use?

2.5.2. Motivation

To answer this question, it would be inappropriate to situate ourselves at a purely theoretical level, namely one that would only assess the theoretical shortfalls of the method and suppose that the FEJ directly adjusts its advocacy and use of RCTs according to them. On the contrary, as our research focuses on a "real-world" institution, the FEJ, and in particular the way that RCTs are carried out within it, we will attempt to answer this question by looking at the practical obstacles that RCTs faced in their deployment within this institution. Our motivation is thus to contribute to the critical literature about RCTs by focusing on a singular institution within which they were funded and carried out. While methodological issues will be raised, they will be so not on a purely theoretical level but insofar as they emerge from the practical application of the method, in its social and institutional environment. As we will develop further, our approach is majorly inspired from the Science and technology studies (STS), whose main line of research is driven by the hypothesis that scientific knowledge, practices and technologies do not happen in a vacuum but in a network of social, institutional, technological, political and economic arrangements. As such, we deem an explanation of the source of the "failure" of RCTs within the FEJ based solely on abstract, methodological limitations of the method insufficient. We are interested in how this evaluation method comes to life, or rather *is put* to life in a "real", social environment, as opposed to the abstract, ideal context in which it is usually presented in economics curricula.

2.5.3. Hypotheses

The puzzle and its associated research question raise two major interrelated hypotheses concerning the potential explanation of the abandonment of RCTs by the FEJ. First, that the social and institutional framework of the FEJ within which the RCTs were undertaken prevented their successful use. Second, that the RCTs themselves, as they were carried out in the field and its social and institutional environment, revealed their own limitations and lack of relevance in terms of their policy and evaluation purposes.

Our hypotheses and associated scope of research are thus broader than that suggested by Jatteau (2016), which mainly concerns the fact that Martin Hirsch, who presided the FEJ in its beginnings, was the main supporter of RCTs and that thereby his departure from the Fund coincides with their abandonment. While not incorrect, we decide to follow a different research track, namely one that focuses on the RCTs themselves and the ways in which they were carried out in practice in order to understand why, ultimately, they failed or were so little used, and eventually abandoned.

2.5.4. Scope of research

The scope of our research concerns the RCTs undertaken within the FEJ and the practical conditions in which they are made. The social and institutional framework of the FEJ will be considered insofar as it is entangled in the field in which RCTs are carried out, but our focus will be primarily held on the practical conditions of realization of the RCTs. Overall, even if our two hypotheses are related, our analysis will thus focus more on the second one. In terms of timeframe, our research's scope ranges from the foundation of the FEJ, in 2009, to today. As such, we account for the RCTs that were carried out throughout the evolution of the FEJ, such as the change in its methodological guidance and administrative composition.

2.6. Main conclusion and recommendations

Our analysis provides evidence in favor of our second hypothesis, namely that the practical or material conditions under which the RCTs undertaken within the FEJ were carried out revealed their limitations and ultimate lack of relevance in terms of policy purposes. In other words, we explain the failure of RCTs within the FEJ by the fact that in practice or as they are made, they are far from being the "gold standard" that they were hoped to be, both in terms of enacted methodology and policy relevance. In summary, our findings suggest that one of the most important limitations of RCTs may not be the difficulty to attain internal validity in practice, their external validity, or the ethical challenges that they raise, as the standard critical literature suggests, but their scope and actual relevance in producing policy-relevant knowledge. These findings come from assessing RCTs as they were made, or rather constructed, and noticing the various arrangements and contingencies that characterize them in the field. As such, our main recommendations are two-fold. Our first recommendation is addressed to the policy evaluation community, and consists in defending a methodological pluralism instead of an exclusive use of RCTs to evaluate policy. Second, and this recommendation is rather addressed to the scientific community, more transparence should be established regarding the actual limitations of RCTs, not only in terms of their practical application but in particular about their narrow scope and ultimate incapacity of producing, on their own, truly meaningful knowledge about policy interventions in the social world.

3. Interdisciplinary state of knowledge

3.1. Standard critical literature on RCTs in the field of economics: a focus on methodology

The philosopher Nancy Cartwright and economist Angus Deaton, 2015 "Nobel" prize, can be considered as the pioneering critics of RCTs. They contributed to bringing out their methodological limits as early as the beginning of the 2000s, when the method was only emerging as a dominant one in the field of development economics. In the literature set around their work and in the field of economics, we can distinguish three interrelated streams of criticism on RCTs: those concerning the internal validity, external validity and policy relevance of the method. While these limits have also been addressed by the proponents of the method[7], we decide here to stick to the literature written by non-"*randomistas*[8]", namely not by those identified as being part of the tight academic network defending and promoting the method.

3.1.1. Internal validity and discrepancy from ideal lab conditions in the field

Internal validity relates to the ability of an RCT to actually deliver an unbiased result, which in the case of an RCT amounts to identifying a causal impact. While internal validity is usually brought forward as the main strength of RCTs, it has been argued to face certain limitations.

[7] Cf. the J-PAL website for a list of resources on randomization [URL: https://www.povertyactionlab.org/fr/node/26?view=toc. Accessed on April 30th, 2021] or Banerjee & Duflo 2009, among the large set of methodological works written by proponents of the method.
[8] Cf. Chapter 4 of Jatteau's thesis (2016) for an extensive definition and sociological analysis of the "*randomistas*".

For instance, Deaton has argued that RCTs have no ability to produce more credible knowledge than other methods insofar as they ultimately face the same problems related to exogeneity and the handling of heterogeneity as other quasi-experimental methods (Deaton 2009, 44-45). In particular, it is their under-reliance on theory and on robust statistical methods that is problematic in terms of actually being internally valid. In an influential and more recent (2018) article called "Understanding and misunderstanding randomized controlled trials", Deaton and Cartwright stress that RCTs cannot automatically deliver precise estimates of average treatment effects. Moreover, Deaton (2009) brings out the various practical problems that RCTs may face and that undermine their internal validity, such as spillover, attrition, Henry, Hawthorne and other experimental biases. Spillover effects relate to the fact that control groups end up being affected by the treatment, and attrition bias relates to the possibility of participants leaving the experimental sample. Hawthorne and Henry biases relate to a change in the behavior by participants, respectively the treatment and the control group, due to their knowledge of being in an experiment. In medicine, these problems are dealt with by making the RCTs single or double blind (Abramowicz and Szafarz, 2020), but such solutions are rare in RCTs carried out in the social field. In a systematic review of the research using RCTs published in top economic journals, Peters et al. (2018) identify that a majority fail to discuss these threats and potential solutions to them.

Quite importantly too, given small sizes of samples, randomization does not guarantee that all characteristics are balanced between the control and the treatment group. This is also a reminder that the ideal RCT set-up relies on the probability law of large numbers, which states that individual characteristics are equally distributed between two different groups by randomization only *on average* and given large enough samples. While rules exist to compute the minimum size of treatment which is "large enough" given the statistical power searched for (Duflo & Kremer 2008, Banerjee 2007), they are not systematically met or even pursued in practice, as the case of the FEJ will show.

In order to address the shortfalls of real, practical RCTs in terms of internal validity, some solutions or "tinkering" can be enacted. Some common solutions to "fixing" internal validity in practice are econometric ones that involve statistical analysis of the data, such as using the random allocation of the treatment as an instrumental variable to deal with selective compliance or controlling for unequally distributed confounders. These are ultimately the same econometric methods that are used in regression analysis using observational, or non- or quasi-experimental data. As Bédécarrats et. al. (2019a, 8) show, assessing the internal validity of RCTs calls for analyzing their "making" and "tinkering" in the field. The term "*bricolage*" or "do-it-yourself" (DIY) is also used by Jatteau (2016) to describe the crafts or adjustments that have to be made in practice to attain internal validity. These include for instance the trimming of the sample, as Bédécarrats et. al. (2019a) notice in their case study, upon which the results of the evaluation strongly depend. As they note, even though RCTs are designed with the purpose of being carried out in the field, outside of the "artificial world of laboratories" (Bédécarrats et. al. 2019a, 9), their ideal protocols rarely apply exactly as intended. To reach the ideal conditions upon which RCTs are internally valid, interventions in the field are thus often carried out, such as communication campaigns to increase the take-up of treatment, or again modifications in the sample to obtain statistical power and balance between treatment and control groups.

3.1.2. External validity

External validity relates to the possibility of generalizing the results of the RCTs, usually to a larger context than that of the initial experiment. First of all, let us remind that in the context of the FEJ, external validity is key to its main objective of eventually generalizing the results from local or small-scale experimentations to the national level. In the literature, external validity is probably the most discussed because most evident limit of RCTs. Ultimately, establishing causality at the internal level does nothing for achieving external validity (Deaton & Cartwright 2018). In other words, Deaton and Cartwright stress that RCTs yield "at most" an unbiased estimate but are of highly limited practical relevance, insofar as they have highly limited external validity. Indeed, the causal relation that RCTs supposedly identifies only holds for the very specific context in which the experiment is carried out. Consequently, alternative methods and especially theory, as Deaton (2009) argues, will be needed for being able to generalize the results of an RCT.

Moreover, as internal validity is a necessary condition for external validity, some of the main threats to external validity identified in the literature are the same as those identified for internal validity, such as Hawthorne and Henry effects (Bédécarrats et al. 2020). Some other experimental biases that are particularly detrimental to external validity are general equilibrium effects or implementer's bias. The latter relates to the fact that different outcomes obtain when an intervention is carried out at a large scale by a government compared to when carried out at a small scale in the context of the RCT, for instance by an NGO. General equilibrium effects relate to the fact that the observed effects in a small sample may cancel out when carried out on a bigger scale.

3.1.3. Policy relevance

The above-mentioned limit of external validity directly relates to that of policy relevance, in that the results of an RCT that only apply to a very limited spatial and temporal context (namely that of the experiment) are relatively uninteresting from a policy perspective. Müller's work[9] (2014, 2015), and in particular his economics thesis, addresses the problem of external validity in light of the policy relevance of RCTs result, framing it in terms of causal interaction. Also, his work is focused on the field of education policy, in which RCTs have been prominent and with particular interest to us given the FEJ's policy field. Moreover, Deaton (2009) clearly voices attention to be drawn away from this method and towards theoretical mechanisms if any relevant knowledge is to be produced on "what works" in development. The fact that RCTs rely on few assumptions and prior theoretical knowledge prevents them from contributing to cumulative scientific progress, hence the need to combine them to other methods to understand not only "what" but "why" things "work" (Deaton & Cartwright, 2018). Moreover, given the institutional context within which RCTs are carried out, including for instance the staff that should carry out the evaluation, the decision-makers actually deciding upon conducting an experiment and so on, the set of projects that are tested by RCTs is not random. According to Müller (2020a), this provides a normative dimension to RCTs, which is often unacknowledged.

[9] Seán Müller posted on his blog (where he voices rather vehement criticism of RCTs and development economics in general) an in-process literature review on external validity and RCTs that was last updated in May 2016: http://www.seanmuller.co.za/Reference_List_EV_May2016.pdf [Accessed on April 30th, 2021].

Such program placement effects do not only pose problems in terms of external validity. They also, as per Ravallion (2009), put into question the ability to close knowledge gaps systematically. Overall, and even if Duflo stressed in her "Nobel Prize" speech that the aim for RCTs is not to answer any type of question but to be driven by specific ones and provide answer to these, it might not be useless to emphasize that RCTs are not able to answer *all* policy questions.

3.2. Ethical concerns raised by RCTs: insights from the legal, medical and philosophical perspectives

RCTs, through their essential process of randomization, imply important ethical challenges which are also widely raised in the literature of various disciplines. To name just a few of the most recent contributions on the ethical implications of RCTs, we can refer to Abramowicz and Szafarz (2020), Bédécarrats et al. (2020), Picciotto (2020) and Ravallion (2019). These have evidently also been addressed by proponents of the method (cf. for instance Banerjee & Duflo 2009). The argument is straightforward: the fundamental protocol upon which RCTs are based, namely that of randomized allocation of a policy treatment, is discriminatory. In other words, RCTs rely on voluntarily discriminating some individuals to the detriment of others in the access to a certain policy or any smaller, supposedly beneficial, public intervention. Such discriminatory practice, even if based on randomization, may be considered unfair. In the case of France for instance, the allocation of a policy benefit based on randomization goes against the republican principle that every citizen should be treated equally in front of the law. To address such issue, or rather surmount this obstacle to conducting RCTs, reforms were made to the Constitution in 2003 (articles 37-1 and 72) to allow for unequal treatment in the exceptional case of experimentations carried out by the State or by local collectivities (Pénissat 2011, 239).

While solutions exist in light of the ethical concerns raised by RCTs, they often pose a threat to internal validity. The first is that of requesting and eventually obtaining the informed consent of the participants in an experimentation, or simply asking them whether they consent to being potentially randomly selected to receive the treatment that the experimentation intends to study. Paradoxically, such procedures of informed consent request will introduce the very bias that RCTs are meant to avoid, such as the Hawthorne and Henry biases named above (Picciotto 2020, 267). We thereby face a vicious circle insofar as these biases may only be limited by using what is closest to a blind treatment in a social experimentation, namely not informing the participants that they are in an experimentation. Another solution is then to request the informed consent of participants but only by giving them limited information. For instance, they might be told that they are participating in a simple survey instead of an experiment. Of course, this leaves the proper issue of informed consent open, since this would only count as partially informed consent. Another solution, less common, is that of equipoise. While widespread in the medical field, equipoise is almost unknown to economists conducting RCTs (Abramowicz and Szafarz 2020, 280). The concept, first introduced by Freedman (1987) in the context of clinical trials, implies that there should be equal ignorance of the benefits and disadvantages of treatment options before trials, in line with the 1964 World Medical Association's Declaration of Helsinki's statement that control groups "must receive the best existing treatment." (Abramowicz and Szafarz 2020, 280). In practice, in the case of social as opposed to clinical

RCTs, the principle of equipoise can be closely attained by forbidding experimentations that may have potential negative side effects (Picciotto 2020, 267). As we will see, these problems of unfairness due to random allocation of treatment and informed consent are dealt with, in certain cases but not in all of the RCTs conducted within the FEJ.

Moreover, while different from the problematic discriminatory nature of RCTs, another ethical challenge that they face lies in their paternalistic dimension, especially when used in development (cf. Favereau 2020, Favereau & Brisset 2016, Labrousse 2010). RCTs may be accused of paternalism, or rather libertarian or "democratic" paternalism (Favereau & Brisset 2016) insofar as broadly, they consist in intervening in the lives of people under the hypothesis that such intervention contributes to their best interest. This logic is close to that advanced in behavioral economics by Thaler and Sunstein (2008) through nudges, whereby an intervention is made in individual's environment in order to delicately or unknowingly incite them to make better choices for themselves. Relatedly, RCTs have been argued to be a new form of behaviorism (Servet, 2018, Jatteau 2018b), as they give primacy to the individual and generally ignore their context, which also provides them with an important ideological dimension, in spite of being presented as "a-ideological" both politically and theoretically (Banerjee & Duflo 2011).

3.3. Interdisciplinary approaches to RCTs in the field

3.3.1. Jatteau's socioeconomic and historical approach to RCTs and the production of evidence by numbers

Arthur Jatteau's foundational work on random experiments and the production of evidence by numbers has provided a fundamental empirical and analytical basis to this research. His 2016 thesis, officially presented in the disciplines of economics and sociology, has provided us with essential elements on the history and sociology of RCTs in a global but also in the particular French context, including reflexive insights on the determining role of the FEJ. Jatteau particularly stresses the role of its initial president, Martin Hirsch, in developing the method in France. Of particular interest to our research, he argues that the "golden age" of RCTs in France appears to have passed after attaining its acme in 2012, associating it to the decline or "quasi-disappearance" of the FEJ and with it its funding of the method (Jatteau 2016, 185). Based on a socioeconomic and historical approach that uses both qualitative and quantitative methods, his research is situated at the crossroads of numerous fields: "bottom-up epistemology", socio-history of sciences, public policy evaluation, development economics, and sociology of economists and of quantification. As an original contribution of his work, he conducts a sociography of the J-PAL by using prosopography and network analysis. His research intends to answer the questions of *how* random experiments prove, or provide evidence, and *what* they are the proof or evidence of, which respectively concern the validity and the scope of the proofs or evidence that random experiments intend to produce. To answer such questions, he does not only focus on the theoretical functioning of random experiments, but on the way that they actually, practically function as they are carried out. His focus on the way that this evaluation method produces, or intends to produce evidence in practice, resembles the approach adopted in this research.

In general, Jatteau's research brings out several limits of RCTs identified above, such as internal and external validity as well as practical problems, for instance the existence of changes that may occur in a program throughout its evaluation and cost-related ones. In terms of solutions offered to solve the problem of external validity, he observes the use of alternative methods such as replication, structural models and accounting for context and underlying mechanisms. He notes that the question of causality is really the one that poses the problem of external validity, and that evidence of effectiveness (the "what" that works) should not be confused with evidence of causality (the mechanisms explaining what works). He argues that RCTs also face political problems, such as misalignment between the objectives and the temporality of researchers and public decision-makers. Overall, he advocates a conditional and non-exclusive use of RCTs, warning against their domination and calling for the coexistence of different methods.

3.3.2. A political economy, statistical and development anthropology approach by Bédécarrats, Guérin and Roubaud

Florent Bédécarrats, Isabelle Guérin and François Roubaud recently published a book called "Randomized Control Trials in the Field of Development: A Critical Perspective" (2020) which gathers, as its name suggests, a wide range of pluridisciplinary criticism on RCTs. These concern the epistemology, ethics as well as politics of RCTs, with a particular focus on the field of development. Of more specific interest to our research is their case study (2019a) on the implementation of six RCTs that were conducted to evaluate the impact of microcredit provision in several developing world regions, and which were advertised by a leading economics journal as the first rigorous and definitive study on the topic. Using tools from statistics, political economy and development anthropology, they analyze the deviation between ideal and theoretical RCT principles and the actual implementation of the entire RCT chains, from sampling, data collection, data entry and recoding, estimation and interpretation, to publication and dissemination of results. They notice significant limitations in terms of internal and external validity, ethics and interpretation of results. Another important aspect of their research is that it reveals the gap that may exist between the academic and the political success granted to RCTs.

In another article, called "All that Glitters is not Gold" (2019b), they use a political economy approach to resolve the paradox of the academic, media and political success of RCTs in spite of their numerous theoretical and practical limitations. Retaking Ravallion's (2019) expression, they argue that RCTs, like any other "real industry" have a market in which supply meets demand. According to them the demand is driven by the academic world and the donor community, which we understand as those funding RCTs. The supply is driven by scientific businesses and entrepreneurs who created a new business model to build a "monopoly and rent position" on the RCTs market. The "domination strategy" of the supply rests on sweeping research from other methods on their topics of interest, disengaging from a "data culture", ignoring criticism to an extent, and "sidestepping certain rules of scientific ethics" (Bédécarrats et al. 2019a, 30). They argue that the so-called *randomistas* often disregard good practices in terms of quality data collection and entry (the "first stage"), just as they disregard related ethical issues (Abramowicz and Szafarz, 2020), and excessively focus on the "second stage", or the

econometrics one which addresses bias issues, selection and identification of a counterfactual (Bédécarrats et al. 2019a, 30). Most quantitative empirical research protocols require a division of labor between data collector (statisticians) and analysts (economists, econometricians), and it is rarely the case that these two can be ensured by the same people as they require distinct training and skills (Bédécarrats et al. 2019a, 30). The former are responsible for accurate measurement, the latter for measurement relevance and analysis as well as for the interactions and relations in the data. As Desrosières (2008b) argues, both economists' and statisticians' activities are essential for the production of sound research results, even though the latter have less social prestige than the former. A similar argument regarding the "superiority" of economists within the academic, the social and the political worlds is put forward by Fourcade et al. (2015). Bédécarrats et al. (2019a, 30) and Jatteau (2016) also mention the incentives that economists have to publish in leading academic journals which shapes the way they advance their methods and results.

3.4. A heterodox economics critique: RCTs as tools of evidence and of government

Early in the growing dominance of RCTs in the fields of economics, development and policy evaluation, the economist Agnès Labrousse initiated the critique of the method in the French literature. As opposed to that of the standard economics literature, her critique is based in the history of economic thought, which involves not only a history of science perspective but also an epistemological one, and takes consideration of social institutions and their broader history. In a seminal article, published in 2010 in the *Revue de la régulation*, she sheds light on the take-off that RCTs had in development economics, both as "tools of evidence and of government". While the term "*preuve*" in French is literally translated into "proof" in English, we chose to translate it as "evidence" as more common in the English discourse on RCTs and policy evaluation in particular, or as is suggested by the name of the evidence-based policy movement. The same goes for Jatteau, whose thesis and recently published book of the same name are called "*Faire preuve par le chiffre*". Precisely, Labrousse mentions that the English notion of "evidence" links to that of empirical corroboration and hierarchization of proofs. The word "*outil*" is also translated as "instrument" by Labrousse in English, but we arbitrarily translated it as "tool".

3.4.1. A "technology transfer" and new contribution to the mainstream

Terminological considerations aside, she argues that there was a "technology transfer" from clinical studies to development economics. This new development economics took the place of the late Washington consensus which was founded on growth theory and regression analysis of macroeconomic observational data. Criticizing it, the later-called "*randomistas*" offered a radically opposite new approach, based on concreteness and realism, and put forward its supposed comparative advantage in terms of causal identification. RCTs were also claimed as advantageous in that, being direct interventions in the field, they are more realistic than lab experiments and are available in a larger "stock" than natural experiments. Interestingly, Duflo (2009, 51) criticizes the latter, referring to the lamp-post bias, as they reduce researchers to only "evaluate what they can evaluate". As we will dig into further, some irony can be discerned in this statement given the limited scope of RCTs (for instance their ability to estimate "at most" very local causal impact through randomization, such as we referred to by citing Cartwright

and Deaton above) and the practical barriers to implementing them such as their cost. Also, as previously mentioned, Deaton (2009) made clear the actual respective advantages of these different methods. Also, by being deliberate transformations of reality, RCTs inscribe themselves in the field of "research-action", analyzed in the socio-anthropology of development (Hugon & Seibel 1988, 13; cf. Olivier de Sardan 1995, 192-199). Furthermore, RCTs focus on more modest objectives and objects, corresponding to the image of "economists as plumbers" or engineers that Duflo still maintains today[10]. Concretely, these objects are "social microstructures" and the behavior of individuals in their environments. Let us note that RCTs may have large "knowledge effects", such as the 2004 "Worms" article by Kremer and Miguel shows. They are new contributions to the mainstream insofar as they can be assimilated to a non-standard approach relative to the then highly theoretical and abstract paradigm, but they ultimately took the institutional place of the mainstream through for instance their publications in the top economics journals and by virtually not referring to the heterodox stream.

3.4.2. Epistemological limits

Labrousse (2010) puts forward a number of the epistemological limits and blind spots of RCTs, at a time when the method was still developing and defining itself. She suggests that these are due to the fact that their proponents then essentially defined their method relative to other techniques, and that debates concerning it focused mostly on the generalization of its results rather than on its investigation logic and heuristics. Labrousse argues that when Duflo refers to RCTs as a pragmatic approach, the term pragmatic is to be understood in the common rather than the philosophical sense. However, the approach does resemble the abduction investigation logic developed by the pragmatist philosophers Peirce and Dewey, for instance in the apparent random character of the discoveries made by the method, and in its empirically founded rather than *a priori* behavioral hypotheses. This effort to build the method on an empirical basis relates to that of having researchers in close contact with the field and "co-experimenting" with actors, in order to give the experimentation a collective, deliberative and "creative" dimension (Duflo 2009, 54), which goes against the "confirmationism" commonly found among econometricians. However, as Labrousse argues, for such iterative processes between observation and theory to really deploy their knowledge effects, the relation of RCTs to theory must be affirmed and clarified, such as in the explanation of the results obtained and their underlying mechanisms. This goes with the fact that the only evidence that RCTs are able to provide relates to the effectiveness of policy interventions, or their impacts, which is understood as a certain type of causality, rather than their underlying causal or explanatory mechanisms. Additionally, conceptual reflections are often under-developed. Concepts such as "development", "human development", "autonomy", or "poverty", are often used in the works of Banerjee and Duflo (2011, 2006; Duflo 2010a, 2010b) but are rarely developed or even simply defined, in line with Duflo referring to experimentations as "conceptually transparent" (2010b, 18). Moreover, the randomized approach is essentially a "micro" one, to which the observation of macroeconomic or structural phenomena is reduced. Since its results also have a high historic specificity,

[10] Such as in her 2020 inaugural lecture of the Sciences Po Paris School of Public Affairs held online, called "Good economics for harder times". This is the same title as her 2019 book co-written with Banerjee and numerous other recent lectures. URL: https://www.youtube.com/watch?v=lR6iHnBOP0c. [Accessed on April 30th, 2021].

Labrousse comments the approach as incapable of answering all questions relating to development, especially those concerning the generalization of its results.

3.4.3. RCTs as social constructs rather than purely objective techniques

Of particular interest to our research, Labrousse considers RCTs as social constructs rather than as purely objective techniques, with their own temporalities, objectives and practical constraints. She cites Dominique Pestre, of the science studies field: "the facts are *made*[11], they are the product of complex and impure actions accomplished in well-specified spaces" (Pestre, 2006, 97). She argues that the constraints of the field do indeed require adaptations from the ideal scientificity of RCTs. This is something that we notice from our study of the RCTs undertaken within the FEJ, with for instance some resistance exerted by the experimented participants, which ends up undermining the randomization protocol. She adds that the construction process of RCTs is filled with arrangements or "tinkering" and contingencies, just like every other methodology "in action", and in opposition to the technical neutrality veil that is usually put upon them. These arrangements and non-neutrality, according to Labrousse, augment rather than diminish the scientificity of the methodology, as they reflect creativity and the development of new hypotheses through the contact with the field and its constraints. Finally, Labrousse notes the economic constraints that condition the use of RCTs, and warns about the use of the method becoming a pre-requisite for receiving evaluation funding, such as we observed in the FEJ. She also stresses the related problems of guest-authorship, whereby some authors provide their names to a report that they might not even have read as way of scientific backing, which is again something apparent with the RCTs of the FEJ.

3.4.4. A new technology of government of populations

RCTs also have a political dimension, insofar as they are set at the "intersection of the politics and the research world" (Duflo 2010b, 18). Labrousse actually transposes the analysis of Desrosières on statistics to that of RCTs, to say that they are a tool for the rationalization of the conduct of human affairs, substituting the reason of measure and calculation for the arbitrariness of passions and power relations (Desrosières 2008a, 22). Just as statistics, RCTs, both in the social sciences and in the management of the social world, have a role of "de-ideologization and objectivation" that allow treating social facts as things, as per Durkheim's expression, which applies both to the natural scientist and to the engineers pursuing progress and shaping nature to human projects (*ibid.*). For Labrousse, this powerful economic technology that RCTs hold explains their "rise in power". Additionally, this rise of the economist as a plumber, as per Duflo's metaphor (Duflo 2009, 27-30), extends standard economics to social devices. To finish, she argues that the technology of population government that underly RCTs resembles more the cameral tradition than the neo-liberal one translated by Foucault's "governmentality" notion.

[11] In French "les faits sont *faits*."

3.5. Political sciences approaches with a particular focus on the French context and the FEJ

Agathe Devaux-Spatarakis' research (2014, 2017) focuses on RCTs and the evidence-based policy making context in France. Using a political science approach, she studies RCTs as an instrument and social institution aimed at organizing a common learning between scientific and public actors. She studies how the inscription of these stakeholders in their respective "strategic action fields" conditions the use of RCTs on the French territory. They then decline themselves in a variety of "institutional sites" that are witness to oppositions in the practices, interests and the learning models of its actors. She does 15 case studies of RCTs undertaken in France, some of which are some of the FEJ that we study. On her view, evidence-based policy making that relies on RCTs as evidence-production methods made its way to France through the creation of the FEJ in 2009 (Devaux-Spatarakis 2017, 3-4). Her study reveals tensions inherent to the application of RCTs on new devices of social intervention, and questions the capacity of the method in terms of producing a common learning between public and scientific actors. Just to give an example, she notices that political actors "cherry-pick" the evidence that the FEJ produces, independently of its quality (Devaux-Spatarakis 2017, 11). For instance, she shows that many local experiments were rolled out nationally before the evaluation was finished, sometimes even before it provided any intermediary results (Devaux-Spatarakis 2017: 11-12). Another observation is one of political misuse of evaluation results, where the former French President Nicolas Sarkozy, known for his "number politics" (*politique du chiffre*) terminated a bill based on RCT results that evaluators had questioned in terms of internal validity due to participants being subject to the Hawthorne effect (Behaghel, Crépon & Le Barbanchon 2011, cited in Devaux-Spatarakis 2017, 12-13).

Another political science approach of RCTs in the French context and the FEJ is that of Bourgois (2010), who describes RCTs as a particular form of social experimentation that faces ethical, operational and financial obstacles. According to him, the experimental approach emerged in France due to a progressive flexibilization of judicial, political and scientific barriers. These coincide for instance, on the political level, to new requirements in terms of the performance of public action, or on the scientific level, to the emergence of a heteroclite network of researchers structured around the experimental approach. He confirms the strong incentive that the FEJ exerted on the use of RCTs at its start, while noting that given the latitude given to project holders, these ultimately favor mixed methods. For him, the FEJ played an important role in renewing the debate in France around evaluation, given the central spot it was provided and the numerous debates that were triggered by the RCT method. Overall, he argues that the emergence of RCTs corresponds to and reinforces certain norms and values in public action.

4. Methodology, data and sources

4.1. An approach inspired by Science and technology studies (STS)

In order to provide answers to our research question, namely why did the FEJ virtually cease relying on RCTs as an evaluation method in spite of their proclaimed advantages, we use an approach inspired by the Science and technology studies (STS). STS are an interdisciplinary field of research in the social sciences which studies scientific practices and technologies as they are carried out on the ground, within particular institutional set-ups and social, professional and economic arrangements. The field was majorly shaped by the pioneering works of Latour, Callon, Law and Woolgar (cf. Latour 1987, 1999; Callon & Latour 1991; Callon 1988, 1997; Callon, Law & Rip 1986; Latour & Woolgar 1988), who follow scientific practices as they are made, within a socio-technical network that shapes them and that they symmetrically shape back. If there is one *a priori* hypothesis upon which this approach is based, it is that scientific practices, technologies and ultimately knowledge do not exist in a vacuum, or that they are always somewhat entangled into a web of socio-professional practices. Our approach is thus similar in that we aim to observe RCTs as they are made in the field in order to get hold of their limitations. In other words, we will pay attention to the various material constraints that shape them and that they shape back, as they intervene in the field.

The Science and technology studies disciplinary field had several "waves" of particular research programs, one of which is the history and sociology of quantification (cf. Bruno & Didier 2013; Desrosières 1993, 2008a, 2008b; Ogien 2013; Martin 2020). As previously mentioned in the literature review, this research program was mainly driven by the works of Alain Desrosières, an INSEE administrator who also contributed to the understanding of the history, sociology and also of the politics of statistics, both in the theoretical and the practical realm (cf. Desrosières 1993, 2008a, 2008b). According to his definition (2008a), quantifying is first convening and then measuring. Statistics, and quantification in general, thereby have both a political and a scientific dimension, hence their description as tools for both "governing" and "proving" (2008a). As Labrousse (2010) has shown, this analysis applies well to RCTs, which can thus also be qualified as quantification tools. They do not rely on quantification per se, as statistics do, but on experimentation, which involves the production as well as the reliance on quantified figures. Precisely, RCTs are a quantitative evaluation method that operates quantification at two levels: first, throughout the experimentation at any moment where participants' characteristics are measured quantitatively, and second, by the production of quantitative measures of certain effects as a result of the experimentation process. Let us also note that the notion of evaluation itself is ambivalent, as it refers both to a scientific practice and an institutional process or government practice. As such, RCTs can be qualified as quantification tools that have both a scientific and institutional or governmental dimension. The term "device", as a translation of the French term "*dispositif*", may be more appropriate to qualify RCTs as it refers to both the greater structure within which RCTs are placed and to the notion of Foucault (2004), close to the one of "*gouvernementalité*", which also fits particularly well to the stakes of RCTs. The notion of device is much larger than that of tools, and as per Callon and Latour's notion of "technical devices", they perform the social world, which itself is composed of both human and non-human objects interrelated in a network. Relatedly, the

notion of "government by instruments" as used by Le Galès and Lascoumes (2005), applies well to RCTs. Our aim will therefore be to study the RCTs undertaken within the FEJ with these notions in mind, grasping them as both scientific evidence and government devices which perform and are performed by the social world in which they intervene.

4.2. A "field" study within a unique institutional framework: evaluation reports as main research material

In a similar fashion to a science studies approach, our research consists in "observing" a variety of the RCTs conducted within the FEJ, "as they were made" or "constructed" in the field, including the various arrangements and contingencies that characterize them. In other words, our aim is to do a case study of several RCTs of the FEJ "in action". Some limitations should be noted right away: within the time and scope of this master's thesis research, it was not feasible to study the RCTs as they were made directly in the field. However, most evaluation reports of the completed RCTs are publicly available on the FEJ's website. We thus decided to use these reports rather than the actual RCTs as our main research material, which is a relatively important deviation from a typical science studies approach. As such, our observations of RCTs in the making will be made via the screen of their evaluation reports, which means that some information regarding how they were actually carried out in the field might be omitted or presented in a way that disforms their reality. However, we believe that the formatting of RCTs by the communication exerted in their evaluation report adds another dimension of interest to our study of the social construction of RCTs. Indeed, although the evaluation reports act as certain veils or formats that respectively cover and shape the reality of the RCTs conducted, they are also a second layer of formatting of the RCTs that makes the connection between how they were made in the field and how they are presented to a wider community of actors. In other words, we can conceive these reports as hinges between the RCTs that took place in the field and their presentation to the public, such as the political and policy-making community for instance, for which RCT results are of interest. This second dimension is important insofar as making the evaluation results publicly available is an element upon which the FEJ counts to foster policy learning among a variety of actors. Evaluation reports thus consist in a rather rich material that allows us to identify, insofar as they are presented in the reports as such, the various practical arrangements and contingencies of the RCTs undertaken within the FEJ, as well as the way in which their measures are presented. Let us note as well that the case of the FEJ is particularly interesting in our sense since it involves a variety of projects with diverse contents as well as the potential to use a diversity of methods to serve its purpose, namely to evaluate such projects in order to produce knowledge about them and eventually upscaling them. As such, the unique framework of the FEJ provides a rich set of projects with different natures which we will see both format and are formatted by the RCTs that are put in place to evaluate them.

4.3. Data collection and analysis

4.3.1 Constructing our sample of RCTs to study

In order to study the RCTs undertaken within the FEJ, "as they were made", we had to select experimentations funded by the FEJ that used RCTs as an evaluation method. For that purpose,

we decided to first count in our study sample the RCTs put forward in the main reports of the FEJ that provide general feedbacks about its experimentations. These include a document of 2014, *De l'éducation à l'insertion: dix résultats du Fonds d'expérimentation pour la jeunesse*, and two others from 2017, including a report of the activities of the FEJ between 2015 and 2017, and a progress report (*Note d'étape*) that presents the experimentations and their results at that same period. We thus first included in our sample the experimentations that were mentioned as using RCTs in these two reports. We checked the past activity reports that were available, namely those dating from 2010 to 2013, but none indicated any other RCTs that were not mentioned in the subsequent reports. As such, we selected all the RCTs that were put forward in the FEJ's main feedback and activity reports up until 2017. This amounts to 9 RCTs, some of which were mentioned in more than one report. For instance, the experimentation *10 000 permis de conduire pour réussir* was mentioned both in the 2014 and the 2017 reports, even though it was carried out between 2009 and 2012. Overall, this allows us to account for the RCTs that the FEJ made particularly visible in its public reports. We also had a look at the 2014 and 2017 activity reports focused on the specific project "*La France s'engage"* funded by the FEJ, but no use of the randomized evaluation method was mentioned. Other than indicating that the FEJ did not favor RCTs as an evaluation method for this project launched in 2014, it does not provide us with any relevant information for our inquiry on the practical execution of RCTs. Our sample also excludes the latest 2019 FEJ feedback report on discrimination and inequalities, as it was not freely accessible.

In addition to these 9 RCTs presented in the FEJ's main reports up to 2017, we decided to add to our study sample the experimentations that explicitly mentioned that they were based on randomization in the experimentation sheets available on the FEJ website. We proceeded in this way since no list of the experimentations including RCTs was available, aside from general statistics presenting the proportions of quantitative and qualitative evaluations carried out by the FEJ, sometimes specifying the proportion of RCTs carried out among these, such as in the 2009-2011 activity report (c.f. FEJ 2011, 29). Anyhow, we were able to conduct a terms search upon the totality of the experimentation sheets available on the FEJ website. Searching for the terms "random", "tirage au sort", "aléatoire" and "contrôl" allowed us to view a restricted sample of the experimentations that might include our objects of interest, namely RCTs. Carefully reviewing these evaluation sheets to make sure that we selected experimentations that really did include RCTs, we were able to add 10 RCTs to our sample, making it amount to 19 RCTs in total. As such, one of the limitations of our sampling method is that we were only able to select those experimentations that explicitly mentioned in their sheets available on the FEJ website that they were using RCTs. This necessarily excludes experimentations that might have used RCTs without mentioning it in their sheets, or at least without mentioning it there as per the terms that we searched for. Just to give one example, this includes for instance an experimentation carried out in 2010 called "Talens Project", evaluated by the Paris School of Economics, whose sheet fails to indicate that the evaluation in question is a randomized one, although this information appears immediately on the evaluation report[12]. We encountered three

[12] C.f. the experimentation sheet of the Talens Project on the FEJ website (URL: https://www.experimentation-fej.injep.fr/534-projet-talens.html) in comparison to its evaluation report (URL: https://www.experimentation-fej.injep.fr/IMG/pdf/ap2_209_eva_rf_201501.pdf). [Both URLs were accessed on April 22nd, 2021].

of such cases when searching for the evaluations conducted by the Paris School of Economics, which we studied for indicative purposes but did not include in our formal RCT sample. Due to time constraints, we did not systematically verify all experimentation sheets and evaluation reports available on the FEJ website, which amount to more than 700. As such, we decided to stick to our sample of 19 RCTs to study "up close" the way they were carried out. While not representative of all the RCTs conducted in the framework of the FEJ, we nevertheless deem it telling of the various practical modalities of RCTs in the field and the obstacles that they might encounter.

4.3.2. Looking at the chain of production of evidence by RCTs

Having this sample of RCTs and their associated evaluation reports computed, we analyzed their respective chains of evidence production. By chains of evidence production, we mean the different steps that contribute to the production of quantitative evaluation results by the RCTs and their publication. From the experimentation sheets and the evaluation reports we were able to gather information ranging from the institution in charge of the evaluation to the evaluation report's structure. Precisely, we constructed a table including textual fragments from our 19 RCT's evaluation reports, which we organized in terms of the following variables:

- nature of the project holder
- institutional and professional background of the evaluator
- geographic zone and size of experimentation territory
- project thematic
- authors and publication date of the evaluation report
- targeted participant group
- nature of the project and its implementation modalities
- intended duration of experimentation
- stated objectives of randomization
- randomization protocol
- modalities of evaluation throughout the experimentation (including data collection and analysis)
- sample selection method
- sample size (including potential evolutions)
- size of treatment and of control group
- evaluation results
- use of complementary quantitative and qualitative methods
- mention of encountered problems, of their nature and potential solutions
- ex ante or ex post verification of balance between the treatment and control groups
- stakeholders
- whether the experimentation resulted in the publication of a paper outside the FEJ website
- budget

This list emerged from noticing some common and diverging elements among the RCTs of our sample. We deem our approach similar but wider than that of Bédécarrats et al. (2019a), who count in their RCT chain the sampling, data collection, data entry and recoding, estimation and interpretation, and publication and dissemination of results. This allows them to identify

problems in terms of internal and external validity, ethics and interpretation of results as well as general inconsistencies with the method. While these are all concepts according to which we end up assessing the RCTs we "observe", we do so by looking at how they are entangled with the practical conditions within which RCTs are carried out.

4.3.4. Encoding identified problems

As such, by looking at the practical conditions in which RCTs are carried out, we identified a variety of problems, obstacles or resistances that the RCTs of our sample seem to encounter in the field. These include internal, external validity and ethical concerns but are not restricted to these now commonly acknowledged problems of RCTs. Actually, by using evaluation reports as our "observational data", we are able to assess the way in which evaluators deal with such problems as presented by themselves. Again, we are not only dealing with the content of the RCTs per se, but with their social and institutional framework as well as with the elements of rhetoric mobilized in the presentation of their implementation and results. Based on the textual fragments that we collected from the evaluation reports, we conducted some textual analysis both by terms search and by direct interpretation in order to compile different "problem indicators" of the RCTs. Our indicators are both quantitative and, to the extent possible, qualitative. They are grouped as 17 main indicators, which are the following:

1) Promoted by the FEJ
2) Size of experimentation territory
3) Institutional nature of project holder
4) Professional characteristics of evaluator
5) Nature of project
6) Nature of sample
7) Sample Size
8) Significance of results
9) Randomization and observation units
10) Comparability of treatment and control group
11) Heterogeneity of effects
12) Use of complementary evaluation methods
13) Internal validity
14) External validity
15) Causality
16) Ethical concerns
17) Budget

These main indicators are decomposed into sub-indicators specifying some aspects of the main categories they refer to. Our indicators are quantitative insofar as it was possible to fill them with a 0 or a 1 for every of the 19 RCTs of our sample to indicate whether they respectively fulfilled or not the information contained in the indicator. This allows us to identify directly which and how many RCTs are concerned with each problem. Also, we qualified these indicators whenever relevant. For instance, in the indicator 11), "use of complementary evaluation methods", we specified which quantitative or qualitative additional evaluation methods were used, or in the indicator 12), "Internal validity", we specified the type of bias noted by the evaluator.

In the next section where we analyze our data and present our findings, we will cover the content of these indicators and sub-indicators in greater details, which will allow us to put forward some of the limitations that RCTs face in the field, in particular that they fail to prove themselves as "gold standards" in practice and that they have a highly limited scope and thereby policy relevance. These indicators are concerned with the RCTs themselves rather than with an extensive organizational analysis of the FEJ. As such, even if the limitations of the RCTs put forward do explain partly their abandonment by the FEJ, it is far from being a full explanation. Thus, the insights of our research are limited to bringing out the contingencies and arrangements of RCTs undertaken by the FEJ as proof of their limitations, which partly plays a role in explaining that they were abandoned by the FEJ. A complete organizational analysis of the FEJ thus falls out of the scope of our research, but we suggest it as complementary further research to ours in order to fully explain the puzzle that we observed. Before moving to the analysis of our results, let us mention that we would have liked to also get direct insights from persons who worked directly with the FEJ and in particular its Scientific Board, but our request unfortunately met no answer. Also, for reference we provide in-text citations of the evaluation reports, which we listed in the Appendix.

5. Analysis – Findings

5.1. Some preliminary comments about our RCT sample

5.1.1. An over-representation of RCTs undertaken at the beginnings of the FEJ

To begin with, let us present some of the main characteristics of the 19 RCTs that we sampled such as previously described. First, we notice that none of them were launched after 2017. The latest RCT that we have noted to occur is one that was conducted from 2017 to 2019 ("Bob Emploi"), which consisted in a randomization on a large sample (approximately 250 000 participants) and used machine learning methods as means to analyze the collected data. More interestingly, all but two were conducted between 2009 and 2012, which corresponds to the first project call of the FEJ. Given that we selected RCTs that were put forward on the FEJ's website and general reports, this tells us that most of these were undertaken at the beginnings of the FEJ. Moreover, it suggests that the FEJ did indeed carry out less and less RCTs throughout its course, or at least that even if more than we are able to see were conducted, they are not put forward as exemplary experimentations among all the ones pursued. Actually, when we look further into the experimentations in which the evaluators belonged to the Paris School of Economics or the J-PAL, we can confirm that these were also undertaken around 2009. As such, the phase where RCTs were most prominent in the FEJ does seem to coincide with its beginnings, gradually disappearing until none of them seem to be carried out anymore today. This is coherent with the fact that the latest general activity reports from the FEJ lack any mention of RCTs.

5.1.2. A variety of thematic, project-holders and evaluators

Other than that, we observe a large variety of thematic among the RCTs sampled: mobility, health, professional insertion, study contracts, school success, school dropout, professional

orientation, fight against harassment at school, housing, contractual autonomy income, and fight against discrimination. This is proof of the richness of the different thematic that the FEJ tackles, and upon which it attempts to produce policy knowledge using RCTs. However, given the limited scope of each RCT's results and how few are undertaken within each thematic, it is difficult to see how they can individually contribute to significant policy learning for each of the various thematic.

Also, the RCTs show variety in terms of the size and types of territories on which they were undertaken. Only two of them, the "Bob emploi" and the "10 000 permis de conduire pour réussir" were deployed on the national level. The others rather concerned samples belonging to a particular region or to towns in the same or different regions. Only in a minority of cases the RCTs were undertaken on a strictly local level, for instance for an experimentation concerning a single boarding school ("Internat de Sourdun") or a single school in a given municipality. Otherwise, in most cases where the experimented device was provided to a wider variety of subjects (most often times schools or individual participants), these extended beyond the purely local sphere. Sometimes, the pre-existence of policy devices similar to the ones being experimented on the experimentation field were noted as diminishing the statistical power of the evaluation results. This is the case for instance in the experimentation on apprenticeship undertaken in Corrèze, where the local mission was already accompanying a lot of participants in their apprenticeship. This undermines the randomized method protocol insofar as the effect of the treatment is already there, making the differences between randomly drawn treatment and control group weak and ultimately preventing the measurement of a statistically significant and precise quantified impact (cf. page 4 of the "Évaluation d'un programme de prévention des ruptures dans l'apprentissage" evaluation report).

In terms of the institutional nature of project holders, these vary uniformly between NGOs, universities and public authorities (local missions, territorial communities and regional agencies). In terms of the evaluating structure, a majority of them are national research centers, such as the National Centre for Scientific Research ("*Centre national de la recherche scientifique*", CNRS), the federation "Labor, Employment and Public Policy" ("*Travail, Emploi et Politiques Publiques*", TEPP) of the CNRS, or again the Research centre in economics and statistics ("*Centre de recherche en économie et statistique*", CREST), also affiliated to the CNRS. In only a single case is the evaluation carried out by an NGO, the EpiSud, in the "Accès Santé Jeunes" experimentation, which was aborted throughout its course. Evaluations are rarely carried out by a single evaluating structure, such as with evaluations taken care of by both the Paris School of Economics, the J-PAL and CNRS-affiliated structures. Often this makes sense, insofar as for instance the J-PAL is based in the Paris School of Economics, or when researchers in charge of the evaluation belong to several of these institutions.

5.1.3. Stars as signs of a somewhat vague legitimacy and representation in activity reports

First of all, out of the 19 RCTs that we selected, 16 are designated by a star on the FEJ website. As the website itself describes, the star is supposed to identify the experimentations "put forward by the FEJ". As such, we have as first indication that a majority of the RCTs selected are deemed worthy by the FEJ of being put forward as some of their best or most successful

experimentations. Our sample does not account for all the RCTs undertaken by the FEJ, but having a look at the three other RCTs that were conducted by the Paris School of Economics within the FEJ, which were all starred, gives us an indication that overall, most RCTs of the FEJ are put forward on their website. Again, this suggests that most of the RCTs undertaken by the FEJ are deemed as successful, or at least worthy of being given attention, as per our interpretation of the star signal vaguely defined by the FEJ. For public visitors of the website for instance, the star may signal the quality of these experimentations and their results and provides them, to an extent, with a certain legitimacy. However, this signal remains very vague in terms of exactly what the FEJ intends to put forward in the experimentations that are identified with a star.

As another clearer sign of providing visibility to the RCTs conducted, the FEJ presents some of them in its main activity reports. One that particularly stands out is the "10 000 permis de conduire pour réussir", which was presented in a positive light in both the 2014 and the 2017 activity reports among others, even though it dates back to 2009. Otherwise, half of the RCTs in our sample are presented in either one or the other of these latest publicly accessible reports. Interestingly, out of the five that are presented in the 2017 report, four of them date from before 2015. Again, this shows that few RCTs were recently undertaken and worthy of being put forward in the FEJ's public reports, which goes in the direction of our initial observation of the disappearance of RCTs among the FEJ.

5.1.4. About the aborted RCTs: attrition, lack of mobilization and political support, ethical concerns, temporality issues and Hawthorne effect

The three RCTs that are not signalled by a star ("Accès Santé Jeunes", "École de la deuxième chance", and "Lutte contre l'absentéisme scolaire") were all aborted, due various reasons making the experimentation infeasible. For instance, "Accès Santé Jeunes" was aborted along the way, after the random draw of the treatment and control groups, due to a too important attrition rate. Indeed, only 16 participants ended up making a demand to access the evaluated device, which was a health expenditure support system. For its results to be statistically significant at 5% and with a power of 80%, the experimenters expected their experimentation to reach a population of 438 individuals, split equally into treatment and control groups.

Different problems explain the abortion of the "École de la deuxième chance" experimentation, which was intended to include 4000 participants and supposed to last for a total of 6 years, from 2008 to 2014, which is twice longer than the maximum of three years during which the FEJ intends to fund the experimentations. The explanation for the termination of this experimentation, a year and a half after its start, is provided in the first report of the Scientific Board of the FEJ (2010, 34-35). One of the multiple causes is said to be the difficulty of mobilizing actors on a new approach (the randomized one) because of their lack of understanding of it and of its stakes, even if the Scientific Board stressed the importance of measuring the impact of such "second chance" schools on the social and professional insertion of the youth. The FEJ justified the stakes of this impact evaluation by the cost of such schools (three times superior to that of other formations) and by the fact that research on the international level did not then show any systematic effectiveness on their part. The argument is then the one typically invoked for the need to conduct an impact evaluation, namely one that

stresses the importance of knowing the actual impacts of a certain policy in order to understand whether it is a waste of resource and if it attains its objectives, or the "efficiency" and "effectiveness" arguments. Anyhow, the FEJ regrets the lack of mobilization and respect of the evaluation protocol by the actors "on the front line" (local missions and advisors) in spite of a long awareness phase, which resulted in having too little participants in the evaluation. Moreover, the fact that the demand for access to these "second chance" schools was lower than their offer did not justify, on an ethical basis, the random allocation of participants to their access. As we will come back to this type of ethical issues further below, let us simply note that it participated in decreasing the political support given to the experimentation. To give a contrasting example, the report of the experimentation "Prévention des ruptures dans l'apprentissage" mentions that the project holders were highly enthusiastic about the evaluation (they were the ones who demanded it) and highly compliant to the evaluation protocol without the need for evaluators to intervene for instance to accompany them or put the different actors of the field in relation. As they argue, these conditions "were highly favorable to the conduct of a randomized controlled evaluation", and the encouragement of bilateral relations between actors in the field made the evaluation as "neutral as possible" (cf. page 41 of the "Prévention des ruptures dans l'apprentissage" report). Lastly, the evaluation is said to have failed because it started when the device to be evaluated was already well-established. In other words, the evaluation and the implementation of the experimented device, namely the "second chance" schools, did not develop at the same time, in an integrated manner, which makes the evaluation difficult. In other words, the experimented device was not well adapted to the evaluation and vice versa, which explains the ultimate abandonment of the evaluation. This puts forward a non-negligible aspect of experimentations, particularly important when RCTs are used: the evaluation device essentially determines the object that it is able to measure, and thereby the final evaluation results. In the case of RCTs, this means that they only exist insofar as they are able to directly shape the nature of the experimented device, namely by making it allocable on a randomized basis, which obviously restricts the type of experimented device that may be evaluated. We will come back to this issue in more detail when we address the scope of RCTs further below.

The third aborted experimentation of our sample, "Lutte contre l'absentéisme scolaire", started in 2009 and was terminated a year after the pilot phase due to the impossibility of "gathering good conditions to conduct a rigorous impact evaluation" (cf. page 2 of the "Lutte contre l'absentéisme scolaire" evaluation report). First, while not directly related to the evaluation itself but to the generalizability of the experimented device after the pilot phase, the evaluators were afraid that participants would not comply to the experimentation protocol, which consisted in providing financial incentives for classes to organize collective projects, due to lack of motivation shown by teachers during the pilot phase. In terms of the possibility to conduct the impact evaluation itself, the evaluators were reaching specific "ethical and scientific conditions" (*ibid.*). The former, which according to the evaluators did not pose an issue, consisted in avoiding risks of exclusion and stigmatization of certain students. Since the incentives provided are collective and concern the totality of a class, the evaluators argue that there are no reasons for such individual risks to apply. On the other hand, the conditions required for the "evaluation to be scientifically valid" or "objective" (cf. pages 2 and 3 of the

report) were not considered to be possibly met. This "scientific condition" only concerns avoiding that the evaluation itself has an influence on the behavior of participants, otherwise called a Hawthorne effect. As stated by the evaluators, this effect can be avoided provided that the experimented device be "sufficiently accepted" by participants such that they do not modify their behavior with a view of modifying its results.

Without further analyzing these obstacles and the way in which they were treated by the evaluators, we already see that carrying out an RCT in the field is far from the method's theoretical simplicity. Whether the randomization has already been enacted or even before so, the possibility of actually measuring anything for the sake of evaluation faces resistance from a multitude of dimensions. Let us remind here that one of the objectives of the FEJ is to fund evaluations and projects that are developed and carried out concomitantly, in order to produce knowledge about the impacts of certain projects intended to be beneficial to the youth, in view of potentially generalizing them. As shown by these three aborted experimentations, which all intended to use the RCT method, the realities of the experimentation field make this objective hard to attain. Problems are already identified in terms of attrition, general compliance to the experimentation protocol by the mobilized actors, political support, ethical considerations, temporality of the evaluation and the experimented device, and Hawthorne effect. These are problems that are also met and attempted to be dealt with in the other RCTs that did fully terminate, which we will now further analyze.

5.2. The issue of internal validity in practice

As it is usually claimed as one of the main advantages of RCTs, we will start by presenting our findings in terms of internal validity and the way in which it was dealt with "in the field", as suggested by the evaluation reports of our RCT sample. Let us remind that the internal validity of an RCT concerns its ability to provide an unbiased estimate of a causal impact, under a probabilistic understanding of causality or one that equates causality to the attribution of an observed effect to a single cause or variable in a given environment. As explained previously, the internal validity of an RCT thus rests on their ability, through their experimental design, to make a treatment and a control group identical in all observed and unobserved dimensions but for the one that it is interested in measuring, namely the only left correlation between the treatment and the outcomes of interest. As a matter of fact, and in reflection to Deaton's (2009) critique, RCTs never automatically equate the characteristics of the treatment and the control groups that they randomly draw, in the sense that they only do so on average given a large enough sample, as per the law of large numbers.

5.2.1. The rhetoric of causality

Interestingly, there are two things among these considerations that are absent of all evaluation reports associated to our sampled RCTs. First, none mention the specific understanding of causality that RCTs claim to identify. Other than mentioning that RCTs enable the "estimation of a causal impact", "the measurement of a causal effect", "the causal effect of…" and the like, none specify further the causality that is at stake. Although these types of statements do make clear that the causality in question concerns an impact, they do not specify its content whatsoever. Second, the method of RCTs is often justified or claimed advantageous relative to other methods by their ability to remove selection bias, to make two groups "comparable" or

even “identical”, to construct a counterfactual, and sometimes even long formal and abstract demonstrations of the internal validity of the method are invoked[13]. To give just one example, the evaluation report of the flagship experimentation “10 000 permis pour réussir” states that the random allocation of participants to the two comparison groups makes them have the “same observable and unobservable characteristics”, or again that randomization is “the only way to ensure that the groups have the same composition and are actually comparable”, followed by a citation of L’Horty and Petit (2011), both authors of the evaluation report (cf. pages 20-21 of the “10 000 permis pour réussir” evaluation report). Even if in this case the sample may be deemed large enough (approximately 7000 participants are counted) this does not guarantee that the two groups’ characteristics end up being, in fact, identical. Another example, from the “Internat de Sourdun” experimentation, is statements such as, after randomization “it suffices to compare the outcomes of the treatment and control group to determine the effect” of the experimented device in question (cf. p.13 of the “Internat de Sourdun” evaluation report). Again, this lack of conceptual precision regarding the notion of causality, even if it is specified as “impact causality”, as well as the affirmation of an apparent automaticity of the randomized method to construct comparable groups, without specifying that it is statistical comparability, are present in all the evaluation reports assessed, with no exception.

5.2.2. Ex ante *or* ex post *verification of the comparability of treatment and control groups*

In the 19 evaluation reports, it is only after such elements are presented that the possibility that the two randomly drawn groups might not be exactly identical is addressed. In other words, after statements are made about the ability of RCTs to make two groups properly comparable and thereby establish a causal effect, in a rather blunt and seemingly uncontroversial way, some reports do stress the need to check the actual comparability of the two randomly drawn groups. In some cases, the evaluation reports subsequently specify that randomization only makes two groups identical “in theory” or on “average”, or that it makes them comparable “statistically”, or if groups were of “infinite size” which is recognized to never obtain in practice, (cf. “De la santé à l’emploi” and “Internat Sourdun” evaluation reports). To then verify that the two groups are actually statistically comparable or “identical on average”, the evaluators measure different characteristics of the treatment and the control group that were constructed through randomization, either before or after the treatment is provided. In that case, the compared characteristics are necessarily observed ones, and randomization cannot do anything anymore about the possibly left endogenous unobserved variables. Out of our 19 “observed” RCTs, 17 of them do conduct either *ex ante* or *ex post* (or both) verifications of the characteristics of the two randomly drawn groups. The only two that do not are two out of the three RCTs that were aborted, the third having still conducted an *ex ante* comparison of some sociodemographic characteristics of the participants. In most cases it is indeed sociodemographic characteristics that are measured and compared, the most common being age and gender, or when the observation unit concerns students of the same class, the class grades or the social origin. Here, two things should be noted: first, that RCTs, already sometimes costly in themselves (in terms of financial, time and human resources concerning the gathering of participants, the allocation

[13] Cf. the demonstration of the Rubin potential outcome framework in page 58 of the “Inscrire les contrats en alternance” evaluation report as an example.

of treatment and the collection of data for instance) thus necessarily require additional costs (again not only financial but human and temporal) of data collection in order to verify the comparability of the control and the treatment group. For instance, if this is not already implied in the initial evaluation protocol, additional phone interviews or the gathering of sociodemographic data through administrative documents have to be organized in order to verify that the randomized allocation actually came close to succeeding. Second, once some data is collected, encoded and "ready to be compared", it has to be actually compared, and this most usually calls into statistical or econometric analysis methods, which come with their own issues that we are not going to discuss here. In terms of methods used to compare the characteristics between the two groups, we usually find standard statistical measurement with hypothesis testing of the statistical significance of the differences, and in some cases, we find techniques such as student or Welch t-tests (cf. "De la santé à l'emploi" and "Mobilité et Accompagnement des jeunes vers l'emploi" evaluation reports).

5.2.3. The presence of biases, the need for tinkering and the heterogeneity of impact

As an outcome of such verification of the balance in the characteristics of the treatment and control groups, some statistically significant differences, and therefore bias, may be identified. This is the case of only four RCTs, as mentioned in their evaluation reports, all others apparently stating no identified statistically significant differences in the characteristics of the groups established through randomization. For these four RCTs that do identify an imbalance in the constructed treatment and control group, let us mention some of the solutions that they employ to solve this issue. For instance, in the "Médiation sociale en milieu scolaire" and "Pass'accompagnement" experimentations, the solution consists in controlling for the variables that are statistically significantly different between the two groups, such as the education level, age and gender of students. Such controlling takes place in the realm of linear regression analysis, which is usually used when a simple comparison of outcomes post-randomization is discarded. An interesting case is that of the "Inscrire les contrats en alternance" experimentation, whereby the method of randomization is used as a means to "*limit*" selection bias to then use another quantitative evaluation method, namely the matching one and in particular the "PSMATCH2" developed by Leuven and Sianesi (cf. page 67 of the "Inscrire les contrats en alternance" evaluation report, [emphasis added].) As such, RCTs are used in this case not as the ultimate evaluation method but only as an intermediary one to minimize selection bias, which acknowledges that it does not eliminate it automatically, and to then use another quantitative method. Here, we see that other quantitative methods are not only used as "tinkering" tools for RCTs but, on the opposite, it is RCTs that may be used as "tinkering" tools for other methods. As such, RCTs should rather be conceived as, in practice, being complementary to other quantitative methods and vice versa.

While such alternative quantitative methods are initially invoked in order to palliate the hazards of randomization or, in other words, "to make randomization work", they are able to bring something new to the evaluation that RCTs ultimately cannot account for, namely the heterogeneity of impact. To say it otherwise, while alternative statistical methods are called into as way of "tinkering" the RCTs or helping them reach internal validity, they are also able to measure certain things that RCTs are in themselves incapable of measuring. This is the case for the measurement of heterogeneity of impact, or the differentiated effects of the treatment given

certain characteristics of the group, when linear regression methods with controls are used. By design, RCTs are intended to measure the impact of a certain treatment on a treatment group which is only characterized by the average characteristics of the group in question. For instance, the generic results of an RCT could be stated as the following: a treatment has an impact on a certain outcome measured with numbers, and this holds for individuals or entities characterized by the average characteristics of the randomly selected sample. As such, by design, an RCT cannot differentiate the ways in which the treatment actually impacted individuals of the selected sample with different characteristics, which might be interesting if we want to take the results of the RCT out of the specific context in which it was undertaken. We are referring to the link that there might be between heterogeneity of impact and external validity, which we will not dig deeper into here, but which allows us to point out another limitation of the scope of RCTs and their potential policy relevance.

In terms of bias in general, we note that most RCTs face experimental biases that are not directly linked to the incapacity of RCTs to automatically produce comparable groups. One of the most common biases noted is that of attrition, whereby internal validity or the statistical comparability of the two groups is compromised due to participants leaving the experimentation or failing to provide consistent response rates throughout. Other common biases noted are contagion, whereby the control group ends up being influenced by the treatment, or again Hawthorne effects, which we previously discussed. Some contagion effects were noted in the "De la santé à l'emploi" experimentation, where treated and control participants are part of the same local mission and are therefore susceptible of exchanging information on the experimented device, which corresponds to providing informational advice on administrative procedures for social security (cf. page 20 of the "De la santé à l'emploi" evaluation report). Peer effects may also be noted when there is a discrepancy between the randomization and the observation unit, such as in the "Mallette des parents" experimentation, where the treatment is randomly allocated to schools, but the outcomes are observed at the individual level. Again, solutions or "tinkering" are enacted when such problems are faced. For instance, in the "Mallette des parents" experimentation, the peer effects which could be considered as bias are accounted for by including them into an econometric model. In our observed RCTs in general, other quantitative methods, usually econometric ones, are used: multivariate linear regression, instrumental variables, difference-in-differences, or Heckman models (such as in the "Stimuler les capacités cognitives" experimementation). In the case of attrition bias, some methods such as reconfiguring the sample of analysis or cutting off of it some participants are used. For instance, in the "Pass'accompagnement" experimentation where the attrition rate was high, a method that equalizes the response rate between the treatment and control group is used, by only considering in the treatment group, whose response rate was higher, the participants that were most easily contacted. Another example is from the "Bob Emploi" experimentation, where due to attrition the evaluators decide focusing only on the reduced sample that is left. While the validity of such methods is not to be discussed here, they show both that 1) internal validity is rarely, if ever, attained in practice, and 2) "tinkering" such as reconfiguring the sample or using other quantitative methods, which come with their own issues, usually happens to palliate the practical limitations of RCTs.

5.3. The issue of external validity in practice

Before digging into the way in which, as suggested by their evaluation reports, RCTs deal with external validity or the generalizability of their results, we shall say a few words about their statistical power and significance, which relates to their sample sizes. As we will see, these sample sizes themselves depend on the nature of the experimented device as well as on the financial means available in the experimentation.

5.3.1. Tinkering the sample for statistical power

First of all, we noticed that checking the statistical power of the experimental device in question was not the concern of all experimentations. For instance, many experimentations fail to even mention the notion of statistical power and thereby to justify that their sample size might be "large enough" not only to minimize imbalances between control and treatment groups but to reach statistically significant effects, such as we mentioned above in the case of the aborted "Accès Santé Jeune" experimentation. As such, only in rare cases do the results produced by the RCTs of our sample appear to be totally significant, in the sense that the results are statistically significant in all the outcomes measured. In other words, many of the evaluation protocols fail to detect significant effects in all the outcomes of interest. For instance, the evaluation report of the experimentation "Mobilité et accompagnement des jeunes vers l'emploi" mentions that the quantitative part of the evaluation, which used an RCT, was not able to detect an effect due to the fact that the observation sample was too small (250 participants in the treatment, 75 in the control), rather than because the "true" effect of the treatment is weak (cf. page 7 of the "Mobilité et accompagnement…" evaluation report). In other cases, the lack of statistical power due to insufficient sample size is not left as is. Indeed, some more "tinkering" takes place to recover internal validity, for instance in the experimentation "Prévention des ruptures dans l'apprentissage" conducted in Corrèze, where a second apprenticeship training center "entered the evaluation" for the purpose of increasing its statistical power (cf. page 41 of the "Prévention des ruptures…" evaluation report). In other cases, and for rather interesting reasons, the lack of significance of the results was not conceived as a failure of the evaluation. In the experimentation "De la santé à l'emploi", the results are deemed insignificant due to contagion bias from the treatment to the control group. However, this bias ultimately consists in making the control group "also experience an amelioration in its health coverage and practices" (cf. page 4 of the "De la santé à l'emploi" report). Here, we witness an amelioration in the individual lives of participants due to the RCT intervention, at the "cost" of its internal validity. The notion of "measuring and measured measure" of Bruno Latour can here be turned on its head as we see that the measure, RCTs, instead of measuring what it intends to measure, is blind to a measure that it produces unintentionally.

5.3.2. The economic conditions of the experimentation

Let us now turn to other practical conditions that affect the statistical power of RCTs. The sample sizes of our selected RCTs vary from 275 to 226 861 individual participants[14]. The latter

[14] The noted respective sample sizes of our 17 RCTs (removing two aborted ones) are the following: 7 143, 468, 1 528, 4 000, 1 520, 395, 275, 902, 1 102, 1 130, 15 450, 226 861, 981, 4 088, 900 and 487. Within the scope of our research, we avoided making a descriptive statistical analysis of this data and instead focused on individual cases and their contexts.

is the above-mentioned "large-scale RCT" using machine learning methods, "Bob Emploi". The ability to deploy an experimented device on such a large scale is due, in that case, to its nature and its cost: it is a website which is both cheap and easy to distribute to a large number of participants. In all our observed cases, the nature and the cost of the experimented device certainly play a role in the ultimate statistical power of the experimental design and the statistical significance of the results. While we have no access to the particular budget allocated to each experimentation and to the cost of each of their experimented devices, some evaluation reports did mention when the budget was detrimental to the validity of results, or in the opposite case when it was beneficial. Take into consideration the experimentation "10 000 permis pour réussir", which received a budget of 10 million euros overall, mainly supported by the foundation of TOTAL, who is one of the private partners of the FEJ. The experimentation consisted in providing a subsidy of 1000 euros to individuals to accompany them in obtaining their driving license, and the experimentation was able to reach out to 10 000 participants. In that case, the budget of the experimentation essentially constitutes the nature of the experimented device (a monetary subsidy), as well as its sample size, which ultimately affects the statistical power of its evaluation. In the case of the "Internat de Sourdun" experimentation, the cost of the project, which was of 154 000 euros per year provided for three years and half enabled to attain the highly ambitious objective of the experimentation, which was to precisely follow up on each participating student throughout the whole experimentation. The costs corresponded to the recruitment of large effectives of research assistants, the use of a survey provider, and important logistics (cf. page 21 of the "Internat de Sourdun" evaluation report). Another example is the experimentation "Mallette des parents", for which the very low cost of the treatment (few informational meetings between teachers and parents) facilitated the deployment of the experimentation on a large scale. In the experimentation "Médiation sociale en milieu scolaire", the opposite abides: its limited budget restricted the data set that was ultimately collected and analyzed, namely those concerning only a single year for a given subset of the sample (cf. page 70 of the "Médiation sociale en milieu scolaire" evaluation report). In that same experiment, by "worry of economizing the budget", only two out of the three schools belonging to the randomly selected school site were studied (cf. page 72 of the report). It also specifies that, for instance, the budget enabled having follow-up phone interviews with only 10% of the studied sample (cf. page 78 of the report). Lastly, its evaluation report clearly states that it was able to gather sample sizes that are usually considered as the "lower bound" for the statistical power of this kind of experimentation, due to its budgetary constraints for financing its mediators (cf. page 80 of the report).

5.3.3 The FEJ's considerations on external validity

This detour by statistical power and sample sizes enabled us to show, before digging deeper into the issue of external validity, that the results of an RCT themselves are not guaranteed to be significant, and thereby difficultly generalizable, in addition to not being assuredly unbiased, as the previous section on internal validity showed. Moreover, it showed that the evaluation itself, and thereby its results, is highly dependent on its economic cost as well as on the nature of the experimented device that is evaluated. With that in mind, external validity already seems rather compromised since the internal validity and statistical power of RCTs is not automatically guaranteed. However, we have also seen that, as per Cartwright and Deaton's

critique among others, even if RCTs are internally valid and statistically powerful, this does nothing to their external validity, insofar as their results are only valid in the restricted experimental context in which they were carried out. As such, other methods, such as Jatteau presented, both quantitative and qualitative, are to be used. The explicit use of alternative methods to RCTs for the purpose of serving external validity is rarely, if ever, noticed in our sample of RCT reports. An exception is the "Bob emploi" experimentation, which uses generic machine learning methods to address heterogeneity and thereby external validity (cf. pages 90-99 of the "Bob emploi" report). However, and we will not discuss them further, such recent machine learning methods have their own limits in terms of results generalization, just as replication methods do (Müller 2020b). Moreover, we found that not all of the 19 RCT reports that we studied explicitly addressed or even simply mentioned the issue of external validity. Indeed, in 6 cases, the terms of "external validity", "generalizability" and their derived forms are completely absent from evaluation reports. In light of the FEJ's insistence on the use of evaluation to decide whether an experimental project is to be generalized or not, we found the lack of systematicity in addressing the external validity issue surprising. A peculiar case is that of the "Groupement de créateurs" experimentation, for which the evaluation report, drafted by the CREST, the Paris School of Economics, the J-PAL and Sciences Po, includes a section titled "external validity and policy recommendations", but is empty (cf. page 50 of the "Groupement de créateurs" report). On the other hand, in some cases we observed some efforts to include in the evaluation report, and thereby in the experimentation, some means or at least some considerations for achieving external validity. For instance, in some cases it is checked whether the observed sample is representative of a wider population, such as the "Mallette des parents" experimentation which concerns the Academy of Versailles or an acclaimed large and representative sample of the French student population. In other cases, the observation sample is typically un-representative of a larger population. This is the case for instance in the "Internat de Sourdun" experimentation, which actually has no intention of producing results applicable to another context, as it intends mostly to measure the impact of the particular education formation that it provides. Overall, and as we will develop in the next related section on the policy relevance of RCTs, we find that the 19 RCTs that we observe do not take remarkable consideration of the issue of external validity or the generalizability of their results, in spite of it being a primordial objective of the FEJ.

5.4. The policy relevance of RCTs

As suggested in our literature review, the issue of external validity in RCTs is essentially related to that of their policy relevance, insofar as the "very local" results of RCTs are weakly relevant to inform policy decisions on greater scales or different contexts. Ultimately, this limitation rests in the fact that RCTs are only able to measure "what works" or rather "what happens" in a very particular spatial and temporal context, which is commonly referred to as the effectiveness or impact of a certain policy intervention. As such, RCTs are typically unable, or rather not *made* to provide any evidence regarding the underlying mechanisms that explain the occurrence of a certain effect. As is stipulated in the methodological guides of the FEJ (both the latest and the first one), qualitative methods are made for that latter purpose. In the 2009 methodological guide for instance, it is clearly stated that quantitative approaches are made to answer questions such as "does the program deliver the expected effects?", or "is the policy

device effective in terms of the pursued objective?", and to thereby provide quantitative measures of the program's performance (Conseil Scientifique du FEJ 2009, 4. [Own translation]). On the other hand, qualitative methods are able to answer questions such as "how does the device allow attaining such objectives?", and to spot the obstacles or levers on which to act to generalize the experimentation (*ibid.*). The latest methodological guide provides similar indications but with reference to "before and after comparison" and "*in itere*" methods instead of explicitly qualitative ones. It also emphasizes on the complementarity of such methods with quantitative ones, in order for the evaluation to be truly adapted to the experimented device, which is considered not only in its nature but in its purpose regarding its potential benefit to the youth (cf. Kerivel 2017, 8-11). As Piccioto (2020, 268) states, "qualitative methods guided by theories of change examine what has happened and why", and help "discriminate between design issues and implementation problems", which is ultimately out of reach for RCTs in themselves. Qualitative and quantitative methods, including RCTs in particular, therefore inform different and complementary dimensions of the policy making process. The following sections will address the precise policy relevance of RCTs, and how this issue is dealt with by the FEJ.

5.4.1. The complementarity between RCTs and qualitative methods in the field

First, and as suggested above, RCTs may be needed to be complemented with other qualitative methods to answer policy-relevant questions. As the FEJ ultimately promotes quantitative, qualitative and mixed approaches, let us assess whether alternative qualitative approaches are used with the RCTs of our sample. Out of the 19 experimentations to which they correspond, ten of them also have a qualitative aspect. In some cases, the qualitative and quantitative aspects of a same experimentation seem completely independent, as suggested by the fact that they are usually presented in separate reports, and that the quantitative one usually makes almost no mention of the other. In rarer instances, the qualitative and quantitative aspects are properly complementary. This is seen in a report submitted in 2017, for the experimentation "Mobilité et Accompagnement des jeunes vers l'emploi", which is subtitled as "a sociological and statistical evaluation" and drafted by both a sociologist and two economists (respectively Anne Denis, Julie Le Gallo and Yannick L'Horty). Here, the qualitative part of the evaluation not only provided its own results, it allowed directing the quantitative evaluation, which was essentially an RCT, towards achieving statistically significant results. Indeed, the RCT was not able to detect any significant effect due to a low sample size, but since the qualitative evaluation, notably through monographic studies, showed that the treatment (a support program for youth towards employment) had high and low intensity levels, the quantitative evaluation ultimately turned toward a regression analysis of the impact of the treatment as a binary variable (cf. pages 7-8 of the "Mobilité et Accompagnement report). Here, we can see how a qualitative method enabled directing the experimentation towards a more policy relevant direction, by providing evidence that the RCT was in itself unable to provide.

5.4.2. Literature reviews as theoretical foundations of experimentations

A critique that can be made about RCTs concerns their under-reliance on theory (Deaton, 2009), which links to their inability to provide explanations for the measured effects as well as to produce generalizable results. As observed in our RCT reports sample, when a qualitative

evaluation is not used to provide mechanistic evidence about the underlying processes explaining the occurrence of certain impacts, a literature review is sometimes provided as way of theoretical foundation to the quantitative evaluation. For instance, for the "Pass'accompagnement" experimentation, evaluated by the CREST and the J-PAL by use of an RCT, the evaluation report starts with a literature review that not only motivates the research question, but provides some theoretical mechanisms underlying the relation between housing and employment that were established in other studies. This quantitative evaluation intends to measure the actual impact of a personalized support on the access to both housing and employment, and thereby constitutes a complement to the mechanistic evidence on the same topic provided in the literature. While in such cases the qualitative and quantitative aspects, including the RCT, of a same policy topic seem well integrated, the presentation of such literatures is a minority among the evaluation reports that we sampled.

5.5. The limited scope of RCTs

The limited scope of RCTs, in terms of their results, is thus partly revealed by questioning their policy relevance. We will now discuss their scope and its limits in a more general way, by looking at what objects they are actually able to measure "in practice", as suggested by their evaluation reports.

5.5.1. A focus on individuals and their behaviors

First, we notice that a majority of the RCTs in our sample randomize at the individual level, meaning that they randomly allocate the policy treatment in question to individuals. In only six cases the randomization unit is a collective entity, such as a school or a classroom most commonly. Nevertheless, the observation unit is unanimously individual. The only case that seems to consider, to an extent, "social" effects or rather peer effects, which correspond to interactions between individuals, is the experimentation "Mallette des parents", which includes longitudinal data on the friendships between students. This "social" aspect of the evaluation remains limited insofar as it studies the effect of the interaction between the treatment and relationships between students, or the level of class integration, on an outcome that is ultimately individual, namely school dropout. While RCTs are meant to measure impacts, and thereby measurable outcomes, it does not go without saying that they should only measure individual outcomes. Indeed, we could possibly conceive of measuring outcomes on collective or institutional levels. This calls upon the definition that Desrosières provides of quantification, which we remind means: first convening, then measuring. With RCTs, the convention regarding the object to quantify seems to rest on individuals and their measurable outcomes, rather than on collective entities or social structures.

Moreover, we notice that as RCTs' scope is limited to individual outcomes, it is also limited to individual incentives and behavior, which relates to the notion of behaviorism that we approached in our literature review. In some experimentations, this notion of behaviorism is quite important insofar as their evaluations sometimes include an important behavioral dimension. This is the case of the "Groupement de créateurs" one, which includes a survey based on behavioral games, constructed by specialists in neurosciences and experimental psychology of the École Normale Supérieure's Cognitive Neurosciences Lab. The collection of psychometric data is justified by the fact that the randomized evaluation intended to measure

the impact of the experimented device on the "decisional autonomy and motivation" of participants (cf. page 16 of the "Groupement de créateurs" report). Besides, this experimentation shows the involvement of experts other than the standard evaluators usually called upon by the FEJ, namely economists mostly and more recently sociologists (such as in the above-mentioned "Mobilité et accompagnement des jeunes vers l'emploi" experimentation). Another sign of the influence of behavioral economics, rather than direct behaviorism is shown in the "Pass'accompagnement" experimentation, whose evaluation report's literature review is largely anchored in the behavioral economics literature, and whose measured results depend on models developed in it. Overall, this shows that a particular conceptualization of the individual, its relation to its environment and to others abides in the realm of RCTs. To put it otherwise, RCTs conceive of the social world in a particular way, one that reduces it to individuals that are not inscribed in greater social structures but who rather react and act upon stimuli from their external environment.

5.5.2. The measuring and measured measure: only "small" randomizable projects are evaluated

In addition to being focused on individuals and their behaviors, the nature of RCTs ultimately determine, or construct, the quantitative measures that it produces. Again, we can here consider RCTs as "measuring and measured measures", as per Bruno Latour's expression. As our sample of RCTs show, they only apply to what we call "small" randomizable projects. These take the following forms: access to internet websites, individual support programs, informational meetings, provision of access to an online platform, classroom or individual monetary incentives, personalized training, etc. For a start, we call these projects "small" insofar as they are only short-term and small-scale interventions in a given environment, rather than larger-scale, more complex policy interventions in unstable environments (Picciotto 2020, 267). Even if these "small" projects are then generalized, they remain focused on individuals rather than on their broader environment. Relatedly, these "small" projects are those concerned and measured by RCTs precisely *because* they are randomizable, or possibly allocated to individuals on a random basis. Moreover, mono-causality, such as is measured with RCTs, is rare in the "real world". RCTs thus intervene in this world by attempting to institute some mono-causality, or, by nature of their experimental design, by attempting to make a single variable vary in a complex environment. As is stressed in the latest methodological guide of the FEJ, RCTs are more than a strict evaluation approach as they form an integral part of the elaboration of the experimented device, and other methods are appropriate to measuring the effects of such a device on a system or understanding its implementation conditions (Kerivel 2017, 8-10). As such, RCTs are not fit for any experimentation, just as not every experimentation can be fit into RCTs. The "Médiation sociale" experimentation is an example of the resistance of reality against the RCT's forceful implementation of mono-causality. In that experimentation, the treatment, which consists in having mediators in classrooms, varies according to the age of such mediators. As such, the internal validity of the RCT collapses due to the fact that the treatment it wanted to implement ends up not being mono-causal.

As a last limitation of the scope of RCTs in that vein, let us mention that in many cases, they are not even able to measure the actual impact of the treatment, but rather the impact of the intention to treat individuals through the given treatment, such as with the "Bob emploi"

experimentation which consists in the provision of access to a website that aims to help young people in job search. The same applies to the experimentations "De la santé à l'emploi" and "AQ3E", whereby additional evaluation devices are put in place to measure the impact of the actual treatment rather than that of the intention to treat.

5.6. Ethical challenges and their solutions

The ethical stakes of RCTs are an issue that is explicitly addressed in some of the FEJ's general reports and methodological guidance. For instance, the first methodological guide acknowledges that RCTs temporarily break with the "equality principle" (Conseil Scientifique du FEJ 2009, 6). It extensively addresses this ethical obstacle and urges experimentations to take precautions to override it, such as avoiding depriving anyone from accessing a resource to which they have a right, and by only providing a selective and temporary access of a new resource in order to measure its effectiveness and eventually generalize it (Conseil Scientifique du FEJ 2009, 11-13). Similarly, the FEJ's activity report of 2009-2011 clearly stresses that the professionals offering young people to participate in an experimentation should ensure that they obtain their informed consent (FEJ 2011, 32). More than that, it is said that for an RCT to be allowed, the demands for accessing the experimented device should exceed its host capacity (FEJ 2011, 32). In other words, this "excess demand" condition means that no participant should be deprived of access to treatment if it is available. This goes in the direction of the equipoise principle which, as explained in our literature review, is meant to provide the best possible treatment to all participants based on the available knowledge. Additionally, the 2009-2011 activity report refers to the "Groupement de Créateurs" experimentation and quotes Nila Ceci-Renaud and Juliette Seban, who were likely involved in it[15], to justify the fairness of RCTs. They do so by defending the "excess demand" condition, as well as by arguing that non-randomized experiments, where treatment is usually allocated on a first-come-first-served basis, are not fairer than random draws, which put all volunteers on an equal footing (FEJ 2011, 33).

In practice, as per our RCT reports sample, the concern for the ethical stakes of RCTs are less obvious. Indeed, only five of them explicitly mention a certain concern for the fairness of the randomization, including four who show anticipation of this issue by stating that the participants have been asked for their informed consent before entering the RCT. For instance, the evaluation report of the "Groupement de Créateurs" experimentation mentions that the evaluation team assumed the "moral responsibility" of informing the volunteers that participated in interviews before the random draw of their selection results (cf. page 14 of the "Groupement de Créateurs" evaluation report). In other cases, the experimentation itself was modified due its ethical stakes, such as when experimenters end up providing the treatment to individuals who were initially randomly assigned to the control group, due to fairness concerns. For instance, the evaluation report of the "Inscrire les contrats en alternance" experimentation

[15] In general, there is a lack of transparency concerning the precise composition of evaluation teams for each experimentation funded by the FEJ and the actual roles of each evaluators in the field. Indeed, no such information is publicly accessible via the FEJ website and if the evaluation reports do not specify names, they only mention the evaluation structures to which evaluators belong to. However, we know that Juliette Seban for instance is the author of another evaluation report and that she is a researcher at the Paris School of Economics and involved in the J-PAL, hence our supposition.

states that by "ethical concern", automatically provides tutoring (the treatment being "reinforced tutoring") to anyone who faces a contract break, and stipulates that the "purely random allocation" of treatment has been respected (cf. page 5 of the "Inscrire les contrats en alternance" evaluation report). Another interesting example is the "10 000 permis pour réussir" experimentation, whose evaluation report mentions that if "exceptionally", and "in the case that an experimenter considers it essential that a person enters the program and benefit from the driving license subsidy, due to personal history, situation or project", then that person should be integrated to the program and thus "escapes the random draw" (cf. page 22 of the "10 000 permis pour réussir" evaluation report). This option is called a "joker", and experimenters only have a maximum of one per thirty applications to participate in the experiment. Obviously, such modifications in the evaluation protocol undermines its internal validity, which is dealt with in various ways such as certain that we mentioned above.

5.7. The socio-professional body of RCTs in the FEJ

5.7.1. Economists as evaluation experts

Last but not least, our analysis of the RCTs we selected focuses on the socio-professional body that frames them. Given the scope of our research and methodological protocol, majorly characterized by a focus on the evaluation reports of these RCTs, we only had a limited but not uninteresting view upon the socio-professional background of the evaluators involved in them. First of all, from the information provided in the evaluation reports, it does not seem that in the case of RCTs, actors other than experts in evaluation per se, who broadly are economists, are usually involved in the evaluation. Indeed, we could have hoped to see experts in the policy topic of the experimentation in question to be involved. For instance, we could have expected specialists in education policy to be involved in the evaluation of experimented device that concern school students, but this does not appear to be the case in our RCT sample. However, we must note certain limitations with respect to our observation of what actually happens in the field and who the actors involved are, provided that we did not have access to any relevant information on that matter other than what was available on the FEJ's website and the evaluation reports.

5.7.2. RCTs of the FEJ held by a tight socio-professional body: distance from the field and signs of scientific legitimacy

Moreover, we notice that the RCTs of our sample were under the responsibility of a tight professional body, with the evaluators in charge of each RCT often coming from the same institution. The most common are the CNRS, the TEPP, the CREST, the Paris School of Economics and the J-PAL, some of which being related, as mentioned above. In particular, we noticed that in a majority of evaluation reports (13 out of 19) the authors were either Bruno Crépon, Marc Gurgand and Yannick L'Horty. First, let us note that the fact that their names as provided as authors of the evaluation reports does not tell much about the actual role that they played on the field. However, given the large number of the RCTs that they respectively sign, as well as the significant involvement that is required for RCTs in the field, it is unlikely that they were significantly involved in each individual experimentation. Their names, just as the J-PAL logos sometimes apparent on the evaluation reports that some of them sign, might just be signals of a certain scientific legitimacy of these reports and their results, especially when these

are also published papers on the J-PAL website. The disconnection from the field that Bédécarrats et al. (2019a; 2019b) and Jatteau (2016; 2018a) noticed from researchers at J-PAL can thereby be found as well in the case of RCTs within the FEJ.

In other respects, we noticed that these three figures are economists, respectively affiliated to the CREST, the Paris School of Economics or J-PAL, and the TEPP. While they can all be considered experts in terms of evaluation, they can above all be considered proponents of RCTs, as Devaux-Spatarakis (2014) extensively discussed in her thesis. The three of them also all had official roles among the Scientific Board of the FEJ, with Marc Gurgand for instance being its first president. As these known proponents of RCTs are over-represented in our sample of RCTs, we can say that this method was supported and held by a tight socio-professional body, composed of economists working in related institutions and significantly involved in developing the method. Indeed, in spite of the apparent decline of RCTs among the FEJ, these three figures still appear to be involved with RCTs today. For instance, Bruno Crépon and Marc Gurgand are both scientific directors at the J-PAL, which unsurprisingly almost exclusively uses RCTs. A broader and more precise sociological analysis of the professional and institutional body composing the evaluating structures of the FEJ is unfortunately out of the scope of this research. We therefore suggest it as a line of further research, in the continuation of the works of Jatteau (2016), such as what he has done for the J-PAL, or of Devaux-Spatarakis (2014, 2017) and Bourgois (2010), who already extensively analyzed the FEJ but not in terms of its most recent evolutions.

6. Conclusion: Policy recommendations based on findings

As suggested by the citation that opened this thesis, Esther Duflo had expected RCTs to revolutionize the world of social policy just as they revolutionized that of medicine in the previous century. If a first conclusion can be drawn from our analysis of the RCTs that were undertaken within the FEJ regarding this hope of the "Nobel" winner, it is that it was defeated. Indeed, even if the FEJ played a significant role in promoting, encouraging and financing RCTs, it ultimately failed to revolutionize the world of social policy through them, at least from the perspective of this institution.

In light of our specific research question, which asked for an explanation of the abandonment by the FEJ of the "gold standard" for policy evaluation that RCTs supposedly are, we can answer that, as revealed by their practical use, RCTs are far from being a "gold standard" and ultimately lack policy relevance. Indeed, our findings showed that RCTs are far from an ideal evaluation method to attain the goals of the FEJ, namely the implementation of projects that benefit the youth, and more importantly the production of general knowledge about youth policy. In summary, we found that first, in line with the standard critical literature, RCTs rarely meet the conditions for them to be internally and externally valid in practice. Entangled with these issues, RCTs also revealed to pose ethical problems to the participants that were involved in them. We also found that they were very rarely used on their own, with other quantitative methods invoked to "tinker" them and palliate their methodological limitations, and other qualitative methods invoked to enlarge the scope of their results. As we "saw" them intervene in the field, we also found that the scope of RCTs was limited by the nature of their protocol, which enables them to focus only on individualized and "small" policy interventions.

Additionally, we saw that they were constrained by economic conditions and that they were held by a tight socio-professional body. We also noticed that they are associated to a particular rhetoric on the notion of causality, which remains obscure but provides them an appearance of scientific legitimacy. Ultimately, all of these interrelated issues concerning the material conditions in which RCTs are made enable seizing their limitations both in terms of methodology and of policy relevance.

In light of these findings, we propose two main recommendations:

- First, and addressing ourselves to the **policy evaluation community**, we suggest that **methodological pluralism should always be favored against the exclusive use of a single method, such as RCTs**. This is because 1) other methods are always needed to support the methodology of RCTs and 2) they enable enlarging the limited scope of the measures produced by RCTs. Other quantitative methods usually help RCTs get closer to internal validity but also to measure objects that they were not able to measure on their own. The latter point also particularly holds for the complementary use of qualitative methods, which allow grasping an aspect of the social world that RCTs are blind to.
- Second, addressing ourselves to the **academic community**, in particular that of **economists**, we strongly urge that **more transparency and conceptual matter should be provided to the notion of causal identification and to the limitations of RCTs, in particular in terms of their scope.** Instead of being put forward as an undebatable objective, the question of causal inference should be set within the broader problem of causality in the social world, especially in light of the limited scope of RCTs, which too often present themselves as self-sufficient. Overall this also links to the **need to cross disciplinary boundaries within the social sciences**, in order for a social science such as economics to learn from others rather than, as it has majorly done until now, taking only the natural ones as models.

APPENDIX: RCTs and their evaluation reports

Name of experimentation (from our 19 RCTs sample)	**Link to experimentation sheet**	**Link to evaluation report [All last accessed on April 28th, 2021].**
10 000 permis de conduire pour réussir	https://www.experimentation-fej.injep.fr/588-benevole-citoyen-passeport-pour-l-emploi-c-est-permis.html	https://www.experimentation-fej.injep.fr/IMG/pdf/APPC_Rapport_Final_Evaluation_Quanti_V2.pdf
Accès Santé Jeunes (16-25 ans)	https://www.experimentation-fej.injep.fr/763-acces-sante-jeunes-16-25-ans.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_EVA_AP2_058.pdf
AQ3E Améliorer la qualité des emplois exercés par les étudiants	https://www.experimentation-fej.injep.fr/445-aq3e-ameliorer-la-qualite-des-emplois-exerces-par-les-etudiants.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_EVA_AP1-336_AQ3E.pdf
De la santé à l'emploi "presaje"	https://www.experimentation-fej.injep.fr/767-de-la-sante-a-l-emploi.html	https://www.experimentation-fej.injep.fr/IMG/pdf/RE_AP2_076_EVA_RF_201404-2.pdf
Élaboration d'un pilote permettant de mieux caractériser l'impact des écoles de la 2ème chance sur le devenir des jeunes	https://www.experimentation-fej.injep.fr/538-elaboration-d-un-pilote-permettant-de-mieux-caracteriser-l-impact-des-ecoles-de-la-2eme-chance-sur-le-devenir-des-jeunes.html	N/A
Évaluation d'un programme de prévention des ruptures dans l'apprentissage/ Sécurisation du parcours des jeunes	https://www.experimentation-fej.injep.fr/920-evaluation-d-un-programme-de-prevention-des-ruptures-dans-l-apprentissage.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_EVA_APDIIESES-09.pdf

s'engageant dans l'apprentissage: bilan d'une expérimentation aléatoire contrôlée conduite en Corrèze		
Évaluation de l'internat de Sourdun	https://www.experimentation-fej.injep.fr/603-evaluation-de-l-internat-de-sourdun.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Evaluation_HAP-01_Sourdun_EEP.pdf
Mobilité et Accompagnement des jeunes vers l'emploi (MAJE).	https://www.experimentation-fej.injep.fr/1363-mobilite-et-accompagnement-des-jeunes-vers-l-emploi-maje.html	https://www.experimentation-fej.injep.fr/IMG/pdf/rapport_final_evaluation_ap5-essaimaje-tepp.pdf
Groupement de créateurs	https://www.experimentation-fej.injep.fr/849-groupement-de-createurs.html	https://www.experimentation-fej.injep.fr/IMG/pdf/rapport_gc_draft_final_22092016_vfinale.pdf
Inscrire les contrats en alternance dans une logique de parcours sécurisé	https://www.experimentation-fej.injep.fr/424-inscrire-les-contrats-en-alternance-dans-une-logique-de-parcours-securise.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_Evaluation_AP1_263.pdf
Lutte contre l'absentéisme scolaire	https://www.experimentation-fej.injep.fr/960-lutte-contre-l-absenteisme-scolaire.html	https://www.experimentation-fej.injep.fr/IMG/pdf/rapport_final_Eval_AP1_354_-_PSE_juin2010.pdf
Mallette des parents Orientation en 3ème	https://www.experimentation-fej.injep.fr/848-malette-des-parents-orientation-en-3eme.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_EVA_HAP_09_Mallette.pdf

Médiation sociale en milieu scolaire	https://www.experimentation-fej.injep.fr/1166-mediation-sociale-en-milieu-scolaire.html	https://www.experimentation-fej.injep.fr/IMG/pdf/RF_EVA_APSCO4_20_c.pdf
Mon parcours emploi / "Bob Emploi"	https://www.experimentation-fej.injep.fr/1709-mon-parcours-emploi.html	https://www.experimentation-fej.injep.fr/IMG/pdf/rapport_evaluation-crest_lfse_2411.pdf
Pass'Accompagnement	https://www.experimentation-fej.injep.fr/351-pass-accompagnement.html	https://www.experimentation-fej.injep.fr/IMG/pdf/RF_EVA_AP1_043_V2.pdf
Promotion de l'apprentissage et sécurisation des parcours des jeunes apprentis en France	https://www.experimentation-fej.injep.fr/369-promotion-de-l-apprentissage-et-securisation-des-parcours-des-jeunes-apprentis-en-france.html	https://www.experimentation-fej.injep.fr/IMG/pdf/RF_EVA_AP1_112.pdf
RCA - Missions locales	https://www.experimentation-fej.injep.fr/1169-rca-missions-locales.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_EVAL_RCA-ML_Quanti_Quali.pdf
Soutien au dispositif "Coup de pouce CLE" et évaluation	https://www.experimentation-fej.injep.fr/604-soutien-au-dispositif-coup-de-pouce-cle-et-evaluation.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_EVA-quanti_HAP-02.pdf
Stimuler les capacités cognitives pour éviter l'échec scolaire	https://www.experimentation-fej.injep.fr/428-stimuler-les-capacites-cognitives-pour-eviter-l-echec-scolaire.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_EVA_AP1_280.pdf
Other experimentations (evaluated by the Paris School of Economics)	**Link to experimentation sheet**	**Link to evaluation report [All last accessed on April 28th, 2021].**

Extension et évaluation du dispositif Créa Jeunes / "Les effets du dispositif d'accompagnement à la création d'entreprise *Créajeunes* : résultats d'une expérience contrôlée "	https://www.experimentation-fej.injep.fr/605-extension-et-evaluation-du-dispositif-crea-jeunes.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Rapport_Final_Evaluation_ADIE-CREAJEUNES_J-PAL_PSE_Mai_2014.pdf
Orientation des jeunes au lycée via des dispositifs de parrainage / Evaluation de l'impact du programme de parrainage d'aide a` l'orientation de l'association Actenses	https://www.experimentation-fej.injep.fr/942-orientation-des-jeunes-au-lycee-via-des-dispositifs-de-parrainage.html	https://www.experimentation-fej.injep.fr/IMG/pdf/Actenses_rapport_jan2013.pdf
Projet Talens	https://www.experimentation-fej.injep.fr/534-projet-talens.html	https://www.experimentation-fej.injep.fr/IMG/pdf/ap2_209_eva_rf_201501.pdf

Bibliography

Abdelghafour, N. (2017). Randomized Controlled Experiments to End Poverty? *Anthropologie & développement*. 46(47), 235-262. DOI: https://doi.org/10.4000/anthropodev.611.

Abramowicz, M., & Szafarz, A. (2020). Ethics of RCTs*:* Should Economists Care about Equipoise?. In Bédécarrats F., Guérin, I., and Roubaud, F. (Eds), *Randomized Control Trials in Development: A Critical Perspective*. Chapter 10. Oxford: Oxford University Press. (2020).

Acemoglu, D. (2010). Theory, general equilibrium, and political economy in development economics. *Journal of Economic Perspectives*, 24(3), 17-32.

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 Perspective*. 24(2), 3-30.

Angrist, J.D. & Pischke J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press.

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

Banerjee & Duflo (2011) *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* Public Affairs. 320p.

Banerjee, A., & Duflo, E. (2009), « L'approche expérimentale en économie du développement », *Revue d'économie politique*, n° 5, 691-726. DOI : 10.3917/redp.195.0691

Banerjee, A. (Ed.) (2007). *Making aid work.* Cambridge (Massachusetts)/London: MIT press.

Banerjee, A., & Duflo, E. (2006), "The Economic Lives of the Poor", *Journal of Economic Perspectives*, 21(1), 141-167.DOI : 10.1257/jep.21.1.141

Barbier, J.-C. & Matyjasik (2010) Évaluation des politiques publiques et quantification en France: des relations ambigües et contradictoires entre disciplines. *Revue Française de Socio-Économie*. 1(5), 123-140.

Bédécarrats, F., Guérin, I., Roubaud, F. (2020). *Randomized Control Trials in the Field of Development: A Critical Perspective*. Oxford: Oxford University Press

Bédécarrats, F., Guérin I., Roubaud, F. (2019a). Microcredit RCTs in Development: Miracle of Mirage?. Working Paper, pre-print version of Bédécarrats F., Guérin I, and F. Roubaud (2019), 'Microcredit RCTs in Development: Miracle or Mirage?', in Bédécarrats F., Guérin I and F. Roubaud (Eds), *Randomized Control Trials in Development: A Critical Perspective*, Chapter 7, Oxford: Oxford University Press (2020).

Bédécarrats, F., Guérin, I., Roubaud, F. (2019b). All that Glitters is not Gold. The Political Economy of Randomized Evaluations in Development. *Development and Change*, 50(3), 735-62. DOI: 10.1111/dech.12378.

Behaghel, L., Crépon, B., Le Barbanchon, T. (2011). *Evaluation de l'impact du CV anonyme*. https://www.parisschoolofeconomics.eu/IMG/pdf/CVanonyme_rapport-final_PSE-CREST-JPAL.pdf. [Accessed on April 24th, 2021].

Bérard J., Valdenaire M. (dir.). (2014). *De l'éducation à l'insertion : dix résultats du Fonds d'expérimentation pour la jeunesse*, La Documentation française/INJEP, Paris.

Bezès P. (2020). Le nouveau phénomène bureaucratique. Le gouvernement par la performance entre bureaucratisation, marché et politique. *Revue française de science politique*. 70(1), 21-47.

Bruno, I. & Didier, E. (2013) *Benchmarking. L'État sous pression statistique.* Zones.

Bourgois, L. (2010). *L'expérimentation, nouvel élixir de jouvence des politiques sociales ? De l'aide au développement des pays du Sud aux politiques de la jeunesse en France : une analyse politique de l'expérimentation sociale.* Mémoire de Master 2, Université Pierre Mendès France.

Callon, M. (1988). *La Science et ses réseaux. Genèse et circulation des faits scientifiques*, La Découverte, Paris.

Callon, M. (1997). « Défense et illustration des Science Studies », *La Recherche*, June, p. 90-92.

Callon, M., Law, J., Rip, A. (1986). *Texts and their Powers: Mapping the Dynamics of Science and Technology*, Macmillan, London.

Callon, M., Latour, B. (1991). *La science telle qu'elle se fait, anthologie de la sociologie des sciences de langue anglaise*, La Découverte, Paris, 391 pages.

Cartwright, N. (2010). 'What are randomised controlled trials good for?', *Philosophical studies*, 147(1), 59.

Cartwright, N. (2007). 'Are RCTs the Gold Standard?', *BioSocieties*, 2(1): 11-20.

Conseil Scientifique du FEJ. (2009). *Guide méthodologique pour l'évaluation des expérimentations sociales à l'intention des porteurs de projet*. [Online]. [Accessed on April 22nd, 2021]. URL: https://www.experimentation-fej.injep.fr/IMG/pdf/guide-pour-l-evaluation-des-experimentations.pdf

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials, *Social Science & Medicine*, 210, 2–21.

Deaton, A. (2009). Instruments of development: Randomization in the tropics, and the search for elusive keys to economic development. *National Bureau of Economic Research*. Working paper number 14690. DOI: 10.3386/w14690.

Desrosières, A. (2013) La mesure du développement : un domaine propice à l'innovation méthodologique. *Revue Tiers Monde*. 1(213), 23- 32.

Desrosières, A. (2008b). *Gouverner par les nombres : L'argument statistique II*. [Online, via OpenEdition]. Paris. Presses des Mines. [Accessed on April 21st, 2021]. DOI : https://doi.org/10.4000/books.pressesmines.341.

Desrosières, A. (2008a). *Pour une sociologie historique de la quantification : L'argument statistique I.* [Online, via OpenEdition]. Paris. Presses des Mines. [Accessed on April 21st, 2021]. DOI : https://doi.org/10.4000/books.pressesmines.909.

Desrosières, A. (1993) *La Politique des grands nombres. Histoire de la raison statistique.* Paris. La Découverte.

Devaux-Spatarakis, A. (2017). *Experiment-based policy making in France: political use of science and practices-based knowledges.* International Public Policy Association's 3rd International Conference on Public Policy, June 28-30, 2017 - Singapore. URL: https://www.ippapublicpolicy.org/file/paper/593fef41056ba.pdf

Devaux-Spatarakis, A. (2014a). *La méthode expérimentale par assignation aléatoire : un instrument de recomposition de l'interaction entre sciences sociales et action publique en France ?* Thèse, Sciences-Po Bordeaux.

Devaux-Spatarakis, A. (2014b) L'expérimentation "telle qu'elle se fait" : leçons de trois expérimentations par assignation aléatoire. *Formation emploi*. 126(2), 17-38.

Duflo, E. (2010a), *Le développement humain. Lutter contre la pauvreté* (I), Le Seuil / République des idées, Paris, 2010, 104 p.

Duflo, E. (2010b), *La politique de l'autonomie. Lutter contre la pauvreté* (II), Le Seuil / République des idées, Paris, 2010, 104 p.

Duflo, E. (2009). *Expérience, science et lutte contre la pauvreté*. Fayard, Paris. 60 pages. DOI : 10.4000/books.cdf.2690

Duflo, E. & Kremer, M. (2008). Use of Randomization in the Evaluation of Development Effectiveness, in Easterly, W. (dir.), *Reinventing Foreign Aid*, MIT Press, 93-120.

Favereau, J. (2020) Being Trapped in Paternalism: Randomized Field Experiments on Poverty. Presentation at the Positive and the Normative in Economic Thought International Conference, held online in December 2020. URL: https://posi-norm-eco.sciencesconf.org. [Accessed on April 30th, 2021].

Favereau, J. (2016) On the Analogy between Field Experiments in Economics and Clinical Trials in Medicine. *Journal of Economic Methodology*, Taylor & Francis (Routledge), 2016. ⟨hal-02092631⟩.

Favereau, J. & Brisset, N. (2016) Randomization of What? Moving from Libertarian to "Democratic Paternalism". GREDEG Working Papers Series. ⟨hal-02092638⟩

Fonds d'expérimentation pour la jeunesse (FEJ). (2019). Qui sommes-nous ? On the FEJ's internet website. URL : https://www.experimentation-fej.injep.fr/115-qui-sommes-nous.html . [Published online on February 8th, 2019]. [Accessed on April 24th, 2021].

Fonds d'expérimentation pour la jeunesse (FEJ). (2018). Animation du FEJ. On the FEJ's internet website. URL : https://www.experimentation-fej.injep.fr/11-animation-du-fej.html. [Published online on March 26th, 2018]. [Accessed on April 29th, 2021]

Fonds d'expérimentation pour la jeunesse (FEJ). (2014) *Rapport d'activité pour 2014*. La France s'engage. [Online]. [Accessed on April 22nd, 2021]. URL: https://www.experimentation-fej.injep.fr/IMG/pdf/FEJ_RA_2014_complet_BD.pdf

Fonds d'expérimentation pour la jeunesse (FEJ). (2012) *Rapport d'activités 2012*. [Online]. [Accessed on April 22nd, 2021]. URL: https://www.experimentation-fej.injep.fr/IMG/pdf/FEJ_RA_2012_FINAL.pdf

Fonds d'expérimentation pour la jeunesse (FEJ). (2011) *Rapport d'activité 2009-2011*. [Online]. [Accessed on April 22nd, 2021]. URL: https://www.experimentation-fej.injep.fr/IMG/pdf/FEJ_RA_20092011_CorpsRapport.pdf

Fonds d'expérimentation pour la jeunesse (FEJ). (2010a). *Rapport du Conseil scientifique du FEJ pour la période mai 2009 - décembre 2010.* [Online]. [Accessed on April 22nd, 2021]. URL: https://www.experimentation-fej.injep.fr/IMG/pdf/rapport-cs-fej-2010.pdf

Fonds d'expérimentation pour la jeunesse (FEJ). (2010b). Appels à projets : mode d'emploi. On the FEJ's internet website. URL: https://www.experimentation-fej.injep.fr/13-appel-a-projets-mode-d-emploi.html. . [Published online on March 16th, 2010]. [Accessed on April 22nd, 2021].

Foucault, M. (2004) *Sécurité, territoire, population*. Cours au Collège de France. 1977-1978, Hautes études/Gallimard, Paris.

Fougère, D. (2000). Expérimenter pour évaluer les politiques d'aide à l'emploi : les exemples anglosaxons et nord-européens, *Revue française des affaires sociales*, vol. 54, 2000, p. 111-144.

Fourcade, M., Ollion, E., & Algan, Y. (2015). 'The superiority of economists', *Journal of economic perspectives*, 29(1), 89–114.

Gautié, J. (2007). L'économie à ses frontières (sociologie, psychologie). Quelques pistes », *Revue Économique*, vol. 58, n° 4, p. 927-939.

Hugon, M.-A. ; Seibel, C. (éd.) (1988), *Recherches impliquées, recherche-action : le cas de l'éducation*, Bruxelles-Paris, De Boeck Wesmael.

Institut national de la jeunesse et de l'éducation populaire (INJEP). (2017a). *Rapport d'activité du Fonds d'expérimentation pour la jeunesse 2015-2017.* [Online]. [Accessed on April 22nd, 2021]. URL : https://www.experimentation-fej.injep.fr/IMG/pdf/rapport_fej_2015-17.pdf

Institut national de la jeunesse et de l'éducation populaire (INJEP). (2017b). *Présentation des expérimentations et des résultats du Fonds d'expérimentation pour la jeunesse.* Notes d'étape 2017. [Online]. [Accessed on April 22nd, 2021]. URL : https://www.experimentation-fej.injep.fr/IMG/pdf/presentation_des_experimentations_et_resultats_du_fej_vf.pdf

Institut national de la jeunesse et de l'éducation populaire (INJEP). (2017c). *Programme d'appui au déploiement de projets innovants d'utilité sociale. Sélection de projets, label, financement, accompagnement et évaluation*. Note d'étape, Septembre 2017. [Online]. [Accessed on April 22nd, 2021]. URL : https://www.experimentation-fej.injep.fr/IMG/pdf/programme_lfse_note_d_etape_sept_2017.pdf

Jamul Lateef Poverty Action Lab (J-PAL). (2021). Introduction to randomized evaluations. On the J-Pal website. Gibson, M., Sautmann, A. (authors). Last updated April 2021. URL: https://www.povertyactionlab.org/resource/introduction-randomized-evaluations. [Accessed on April 28th, 2021].

Jatteau, A. (2019) Les essais contrôlés randomisés. Une comparaison entre la médecine et l'économie. *Philosophia Scientiae*, 23(23-2), 85-110.

Jatteau, A. (2018a). Comment expliquer le succès de la méthode des expérimentations aléatoires ? Une sociographie du J-PAL. *SociologieS*. Dossiers, Les professionnels de l'évaluation. Mise en visibilité d'un groupe professionnel. [Online]. [Accessed on April 20th, 2021].

Jatteau, A. (2018b). De quoi les expérimentations aléatoires sont-elles le nom ? À propos de l'ouvrage de Jean-Michel Servet : L'économie comportementale en question. *Revue de la régulation*. [Online]. 23, Spring 2018. DOI: 10.4000/regulation.13148. URL: http://journals.openedition.org/regulation/13148. [Accessed on April 5th, 2021].

Jatteau, A. (2016). *Faire preuve par le chiffre ? Le cas des expérimentations aléatoires en économie*. Thèse de doctorat ENS Paris Saclay. Sous la direction d'Agnès Labrousse et Frédéric Lebaron.

Kabeer, N. (2019). Randomized Control Trials and Qualitative Evaluations of a Multifaceted Programme for Women in Extreme Poverty: Empirical Findings and Methodological Reflections. *Journal of Human Development and Capabilities*. 20(2), 197-217.

Kerivel, A. (from the Scientific Board of the FEJ). (2017) *Guide méthodologique relatif aux évaluations du FEJ*. INJEP, Paris. [Online]. [Accessed on April 22nd, 2021]. URL : https://www.experimentation-fej.injep.fr/IMG/pdf/methodes_evaluation_experimentation_guide.pdf

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

Labrousse, A. (2019). The rhetorics of Poor Economics. In Bédécarrats, F., Guérin, I., and Roubaud, F. (Eds) *Randomized Control Trials in Development: A Critical Perspective*. Chapter 8, Oxford: Oxford University Press. (2020).

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*. 7, 2-32. DOI : 10.4000/regulation.7818.

Le Galès, P., & Lascoumes, P., (dir.) (2005), *Gouverner par les instruments*, Paris, Presses universitaires de Sciences-Po.

The Lancet. (2004). The World Bank is finally embracing science. Editorial, 364(9436), 731-732. [Online]. [Accessed on April 21st, 2021] DOI: https://doi.org/10.1016/S0140-6736(04)16945-6

Latour, B. (1999) *Pandora's Hope. Essays on the Reality of Science Studies*. Harvard University Press, Cambridge, Massachusetts.

Latour, B. (1987) *Science in Action. How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge, Massachusetts.

Latour, B., Woolgar, S. (1988). *La vie de laboratoire. La production des faits scientifiques*. Paris, La Découverte (re-edited in 2006).

Lordon, F. (1997), « Le désir de « faire science » », *Actes de la recherche en sciences sociales*, vol. 119, p. 27-35.

Martin, O. (2020). *L'empire des chiffres*. Armand Colin.

Müller, S. (2020a). *The Unacknowledged Normative Content of Randomised Control Trials in Economics*. Presentation at the Positive and the Normative in Economic Thought International Conference, held online in December 2020. URL: https://posi-norm-eco.sciencesconf.org. [Accessed on April 30th, 2021].

Müller, S. (2020b). The implications of a fundamental contradiction in advocating randomized trials for policy. *World Development*, Elsevier, 127(C).

Müller, S. (2015). Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations. *The World Bank Economic Review*, 29(suppl. 1), S217–S225. DOI: https://doi.org/10.1093/wber/lhv027

Müller, S. (2014). *The external validity of treatment effects: an investigation of educational production*. Thesis presented at the University of Cape Town.

Nioche, J.-P. (1982). De l'évaluation à l'analyse de politiques publiques. *Revue française de science politique*. 32(1), 32-61.

Ogien, A. (2013). *Désacraliser le chiffre dans l'évaluation du secteur public.* Quae.

Olivier de Sardan, J.-P. (1995), *Anthropologie et développement – Essai en socio-anthropologie du changement social*, Paris, Karthala.

Penissat, É. (2011). Quantifier l'effet « pur » de l'action publique : entre luttes scientifiques et redéfinition des politiques d'emploi en France. *Sociologie et sociétés*, 43(2), 223-247.

Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity. *The World Bank Research Observer*, 33(1),34-64. DOI: 10.1093/wbro/lkx005.

Perez, C. (2000). L'évaluation expérimentale des programmes d'emploi et de formation aux États-Unis: éléments de critique interne.

Pestre, D. (2006), *Introduction aux Sciences Studies*, Paris, La Découverte.

Picciotto, R. (2020). Are the "Randomistas" Evaluators? In Bédécarrats F., Guérin I and F. Roubaud (Eds), *Randomized Control Trials in the Field of Development: A Critical Perspective*, Chapter 1, Oxford: Oxford University Press (2020).

Pritchett, L., & Sandefur, J. (2015) Learning from experiments when context matters. *American Economic Review*. 105(5), 471-75.

Ravallion, M. (2019). 'Should the Randomistas (Continue to) Rule?'. In Bédécarrats F., Guérin I and F. Roubaud (Eds), *Randomized Control Trials in the Field of Development: A Critical Perspective*, Chapter 1, Oxford: Oxford University Press (2020).

Ravallion, M. (2009). 'Should the randomistas rule?', *The Economists' Voice*, De Gruyter, 6(2), 1-5.

Rubin, D. B. (2005) Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469) 322-331.

Servet, J.-M. (2018). *L'économie comportementale en question*, Éditions Charles Léopold Mayer, 205 p.

Shaffer, P. (2015). Two concepts of causation: implications for poverty. *Development and change*, 46(1), 148-166.

Stern, E. Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the range of designs and methods for impact evaluations. Department for International Development Working Paper, 38. London, UK.

Supiot, A. (2015). *La Gouvernance par les nombres*. *Cours au Collège de France 2012-2014*. Paris : Fayard. Édition 2020. 512 pages.

Thaler, R. H. & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press. New Haven, CT. 293 pages.

Valedenaire, M. (2013). *The French Youth Experimentation Fund*. Powerpoint presentation for the International Workshop "Evidence-based Innovation: the Role of Evaluation and Social Experiments" held in Barcelona, September 26th 2013.

Vivalt, E. (2017). How much can we generalize from impact evaluations? Working Paper. Stanford, CA: Stanford University.

Woolcock, M. (2013). Using case studies to explore the external validity of 'complex' development interventions. *Evaluation*, 19(3), 229-248.

Public Policy Master's Thesis Series

This series presents the Master's theses in Public Policy and in European Affairs of the Sciences Po School of Public Affairs. It aims to promote high-standard research master's theses, relying on interdisciplinary analyses and leading to evidence-based policy recommendations.

The rise and fall of a gold standard. The case of Randomized Controlled Trials within the Experimentation for Youth Fund.

Anne-Pauline, de Cler

Abstract

In the economics, development and policy evaluation fields, randomized controlled trials (RCTs) have been put forward as a "gold standard" for measuring the impact of public policies. The French Experimentation for Youth Fund (FEJ) is a public-private institution that funds experimental projects and their evaluations, in view of producing knowledge about youth policy and eventually generalizing the projects. In its beginnings, the FEJ encouraged the use of RCTs and promoted them as an evaluation method to be privileged. However, their use gradually disappeared from the FEJ and they are no longer advocated as the best method to evaluate the projects it funds. Our research aims to explain the disappearance of the so-called "gold standard" that are RCTs within the FEJ. Using an approach inspired by the science and technology studies, we study the RCTs undertaken within the FEJ as they were constructed "in the field", via the screen of their evaluation reports. From our analysis of 19 cases of RCTs, we address the various practical conditions, arrangements and contingencies that characterize them. This allows us to bring out their limitations in terms of methodology and lack of policy relevance, which partly explains their abandonment by the FEJ. In light of our findings, we recommend the use of pluralistic methods for policy evaluation, as well as greater transparency expressed by the economist community regarding the practical limitations of RCTs.

Key words

Randomized controlled trials, policy evaluation, Fonds d'Expérimentation pour la Jeunesse, science and technology studies.