WORKING PAPER DT/2019-04

Lies, damned lies, and RCT: A J-PAL RCT on rural microcredit in Morocco

Florent BEDECARRATS

Isabelle GUERIN

Solène MORVANT-ROUX

François ROUBAUD

UMR DIAL 225
Place du Maréchal de Lattre de Tassigny 75775 • Paris •Tél. (33) 01 44 05 45 42 • Fax (33) 01 44 05 45 45 • 4, rue d'Enghien • 75010 Paris • Tél. (33) 01 53 24 14 50 • Fax (33) 01 53 24 14 51
E-mail : dial@dial.prd.fr • Site : www.dial.ird.fr

Lies, damned lies, and RCT: A J-PAL RCT on rural microcredit in Morocco[1]

Florent Bédécarrats Division des Evaluations de l'AFD 5, rue Roland Barthes, 75598 Paris, France bedecarratsf@afd.fr	Isabelle Guérin IRD, UMR CESSMA, 75013 Paris, France UMR CESSMA, 75013 Paris, France isabelle.guerin@ird.fr
Solène Morvant-Roux Graduate School of Social Sciences UNIGE-G3S, University of Geneva, Switzerland Solene.Morvant@unige.ch	François Roubaud IRD, UMR DIAL, 75010 Paris PSL, Université Paris-Dauphine, LEDa, UMR DIAL, 75016 Paris, France roubaud@dial.prd.fr

Working Paper UMR DIAL

February 2019

Abstract

How can we explain the academic success of a randomized study whose validity, both internal and external, is very problematic? Drawing on a study conducted on Moroccan rural microcredit by J-PAL, this article uses analytical tools from statistics, political economy and sociology of science to answer this question. It describes the entire study production chain, from sampling, data collection, data entry and recoding, estimates and interpretations to publication and dissemination of results. It highlights a particularly aggressive strategy carried out throughout the study process and in the field of research. This allows J-PAL researchers to put the past behind them, including by freeing themselves from a "data culture", rejecting criticism and bypassing the basic rules of scientific exercise throughout the research process. Well beyond J-PAL, our analyses question the supposed superiority of randomized methods while reflecting a growing unease within the academic field, which is less and less successful in enforcing the basic rules of ethics and scientific deontology.

Keywords: Randomized Control Trial (RCT), microcredit, Replication, Morocco, internal validity, internal validity, sociology of sciences

JEL Codes: A11, A14, B41, C18, C93, N27, O16.

Résumé

Comment expliquer le succès académique d'une étude randomisée dont la validité, tant interne qu'externe, est pourtant très problématique ? Prenant l'exemple d'une étude menée par le laboratoire J-PAL sur le microcrédit rural marocain, cet article mobilise les outils analytiques de la statistique, de l'économie politique et de la sociologie des sciences pour répondre à cette question. Il décrit l'ensemble de la chaîne de production de l'étude, depuis l'échantillonnage jusqu'à la publication et la dissémination des résultats, en passant par la collecte de données, la saisie et le recodage, les estimations et les interprétations. Il met en évidence une stratégie particulièrement offensive qui permet aux chercheurs de J-PAL de faire table rase du passé, y compris en s'affranchissant d'une « culture de la donnée », de refuser la critique et de contourner les règles de base de l'exercice scientifique tout au long du processus de recherche. Bien au-delà de J-PAL, nos analyses questionnent la supposée supériorité des méthodes randomisées tout en reflétant un malaise grandissant au sein du champ académique, qui parvient de moins en moins à faire respecter les règles de base de l'éthique et de la déontologie scientifique.

Mots-clefs : Randomized Control Trial (RCT), microcrédit, Réplication, Maroc, validité interne, validité externe, sociologie des sciences

[1] This working paper is a shorter version of Bédécarrats F., Guérin I, Morvant-Roux S., Roubaud F. (2019), « Lies, damned lies, and RCT : une expérience de J-PAL sur le microcrédit rural au Maroc », Document de Travail DIAL, No 2019-04. (http://www.dial.ird.fr/publications/documents-de-travail-working-papers) only available in French. The title refers to the famous phrase "*There are three kinds of lies: lies, damned lies, and statistics*", attributed to Benjamin Disraeli and popularized by Mark Twain.

Introduction

Borrowing from medical sciences, randomized control trials (RCTs hereafter) have been extended to development policies since the early 2000s and are now considered the gold standard in impact evaluation. As a response to the thorny counterfactual issue, they theoretically offer the possibility of precisely isolating and quantifying the impact of a development intervention, all other things being equal.

Although this method is very attractive, as evidenced by its widespread use (Ravallion, 2018), it is not above criticism. Among other issues, RCTs are criticized for their inability to extend beyond the particularities of the interventions studied (i.e. their external validity[1]), their incapacity to make an optimal trade-off between bias and precision and to measure externalities (i.e. their internal validity[2]), and their exposure to power games, at both the field experiment level and in the RCT industry as a whole (Bédécarrats et al., 2017). This article's purpose is to contribute to this debate by examining the implementation of a specific RCT on rural microcredit supplied by a vocal Moroccan microcredit institution (Al Amana, AAA hereafter) and published by Crépon et al. in 2015 (2015, hereafter CDDP, initials of the authors' names).

There are a number of reasons for our choice. Microcredit is a key RCT focus in the development field: at J-PAL (the research center most active in RCTs), 273 of its 947 RCTs (complete or in progress, on 27 February 2019) are on financial issues.[3] An initial summary of RCT findings on microcredit was published in 2015 (Banerjee et al., 2015) in a special issue of the *American Economic Journal: Applied Economics* (AEJ: AE). This summary was seen as a decisive contribution to settling a long-standing debate on the subject (Ogden, 2017). Four years after its publication, the special issue has already been cited 3,143 times (Google Scholar; 25 February 2019). CDDP is one of six studies in the special issue, and exhibits many strengths: focus on areas not yet studied (remote rural areas, and the Maghreb); a prediction model supposed to get round the recurrent problem of low take-up (highly problematic since it threatens statistical power and therefore the internal validity of RCTs); ideal conditions for an experiment, since AAA was just starting its expansion into remote rural areas, which were thereby expected to be free of any formal credit supply; and, lastly, an identification strategy designed to measure externalities. For these reasons, the AEJ: AE special issue's introduction relies heavily on CDDP to draw general conclusions, on both the impact of microcredit and the potential of RCTs, expanded on by the prediction model and measurement of externalities. CDDP is clearly an academic success, with 248 citations four years after publication (Google Scholar; 25 February 2019). And CDDP is signed by some of the most prestigious RCT proponents, which is taken as a guarantee of quality: Esther Duflo, one of the most well-known RCT leaders, and Bruno Crépon, who is also one of the pivots of the global RCT network (Jatteau, 2016: 311).

Last but not least, the unique opportunity to access various sorts of data, both quantitative and qualitative, is also instrumental in our choice.

- The AAA-RCT micro-data are freely available, making for an extremely thorough replication. The results of this replication, published elsewhere as a companion paper (Bédécarrats et al., 2019), turn up a number of internal and external validity problems.

[1] See, for instance, (Deaton and Cartwright, 2018; Heckman, 1991).

[2] See, for instance, (Deaton and Cartwright, 2018; Heckman, 1991; Ravallion, 2018; Rodrik, 2008).

[3] Data available from the J-PAL website (https://www.povertyactionlab.org/evaluations), accessed on 27 February 2019.

Two other types of data are used to *explain* these shortcomings:

Two of us participated in a qualitative field study designed specifically to complement the AAA-RCT. In 2009, we conducted 79 semi-directive interviews with different AAA stakeholders and players in the AAA environment (clients, non-clients, loan officers and key local stakeholders such as imams, grocers and local leaders) in a number of Moroccan regions (Morvant-Roux et al., 2014). The fact that the qualitative study was conducted in the last stage of the AAA-RCT (endline data collection) provided a unique opportunity to observe its actual implementation and opened the door to our use of the participatory observation method, well known to anthropologists. Our indirect, secondary role in the experiment's activities was conducive to close contact and collaboration with the main actors. This privileged observation post offered an exceptional opportunity to raise relevant questions about the microcredit program that the researcher proposes to evaluate.

This qualitative study was commissioned and financed by the same donor as the experiment, Agence Française de Développement (French Development Agency or AFD). A third author's work as political scientist and impact evaluation specialist at AFD gave him a seat in AFD's in-house discussions on the subject. The fourth author brought his household and microenterprise survey skills to the team.

We also had access to an almost exhaustive set of grey literature and internal documents produced throughout the implementation of the AAA-RCT, from design to dissemination: AFD notes, steering committee reports and PowerPoint presentations, project monitoring reports by J-PAL, e-mail exchanges, and academic articles published by AFD researchers drawing lessons based on their experience.[1] We also conducted a series of interviews, most of which were repeated over time, with some of the RCT's key stakeholders: AAA executive staff in charge of monitoring the RCT, AFD staff in the Research Directorate's Evaluation Department, and J-PAL researchers. However, exchanges with the latter were restricted to an interview with staff in charge of supervising the field surveys and a brief discussion at the final presentation of the two studies. Despite several requests, the J-PAL team declined the invitation to collaborate. Although this obviously restricted the analysis, access to the grey literature nevertheless captured their point of view.

While replication is slowly emerging in development economics (Duvendack et al., 2017), no replication has yet been combined with a qualitative analysis of the mechanisms underlying the scientific production of the experiment. The dataset used here, and our team members' observations of the experiment's implementation in real time and over time, effectively delves behind the scenes of "science in action". Scientific research is often considered a rational process, motivated primarily by the pursuit of rigor, objectivity and excellence, and regulated by performance: the most rigorous theories best able to explain and predict the world stand to be considered the best and to naturally prevail. However, the sociology of science shows that knowledge may well be a scientific fabric, but it is also a social, cultural, economic and political fabric. Turning data into scientific results is a complex process that entails a series of "translations" involving a multitude of factors. Evidently, empirical data, technical and analytical skills, and possibly theory, come into play. Yet the power of persuasion, i.e. ensuring that a result is convincing and disseminated, is just as fundamental. Researchers' power of persuasion depends on their belonging to more or less influential networks (Akrich et al., 2013) and on the power games and "capital" with which researchers are unequally endowed (Bourdieu, 1976). This power of persuasion makes use of scientific demonstration, but also a series of choices, adjustments and "tweaks", and translations and formulations. Once the final scientific result has

[1] The use of this grey literature led us to adopt a specific referencing system. The published documents mentioned in the body of the text have the usual notation. Public presentations in the form of slides are noted I_No. x and internal documents are noted D_No. x. These are listed in the References section at the end of the article

been obtained, all the intermediate steps and procedures that went towards its production are generally relegated to the level of technics (Latour and Woolgar, 1996).

This is what is observed here: scientific results are produced - here, the impacts of microcredit - and become dissociated from the techniques (here, sampling, surveys, coding and econometrics) that gave birth to them. The unique data to which we had access have enabled us to reconstruct the entire results production chain from sampling, data collection, data entry and recoding, estimates and interpretations to publication and dissemination of the results. And we show that these results are contingent on a series of trade-offs, constraints, interpretations, actor interactions and power games.

Ultimately, our paper explains three paradoxes. The first is the gap between a multiple error-impaired RCT and its academic success. The second paradox is the diametrically opposed conclusions drawn by the RCT's two main actors. CDDP, and more broadly J-PAL and RCT proponents, hold up experimentation as an example of scientific success from which more general lessons can be drawn. By contrast, AFD, the funder, concludes that RCTs are valid only for a very narrow set of interventions, while other methods should be preferred for many other interventions, including microcredit. The third paradox lies in the contrast between the method's supposed simplicity (attributing and quantifying causal impact simply by comparing treatment group with control group mean, based on the identification of an experimental counterfactual), which is one of the founding arguments put forward by its promoters to assert the generic superiority of RCTs over any other method, and the surprising complexity of the protocol actually applied by the AAA-RCT.

The paper is organized as follows. After presenting the main lines of the experiment, the first part summarizes the results of the replication and the different shortcomings observed. The second part describes the implementation of the protocol and the gap between theory (simplicity of the method, comparison of comparable villages with and without microcredit, and good data quality) and practice (extraordinarily complex protocol, biased samples and poor data quality). The third part argues that this gap results from a particularly offensive monopoly strategy in the field of impact evaluation. This allows the J-PAL team to make a clean sweep of previous research, freeing themselves from a "data culture", passing over criticism and circumventing the basic rules of scientific conduct. In conclusion, we draw more general lessons: above and beyond this specific study and J-PAL themselves, our results place a question mark over the supposed superiority of randomized methods in an echo of growing unease in an academic field increasingly struggling to enforce the basic rules of ethics and scientific deontology.

I. The replication

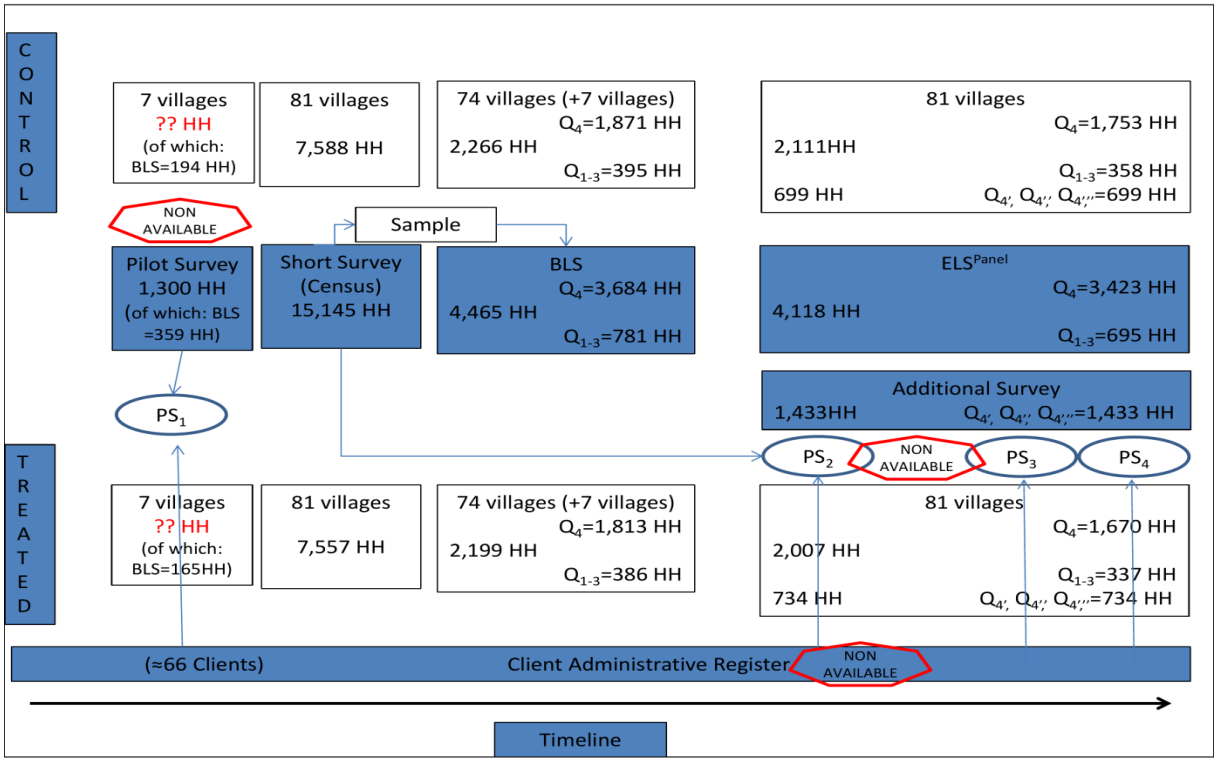
Between 2006 and 2010, a research team from J-PAL conducted an RCT in rural Morocco to measure the impact of microcredit provided by AAA, the then Moroccan market leader in the midst of a phase of expansion. As AAA had already begun to expand into rural areas, the RCT focused on remote areas. We first describe the protocol, and then summarize the results of our replication.

IA- The protocol

The initial protocol adhered to (in theory) the established standards for randomized trials. A set of 162 villages were randomly selected in the area where AAA planned to expand, within which 81 pairs of villages were formed based on observable variables recorded by a short preparatory survey prior to the RCT. For each pair, one village was assigned to the treatment group and the other to the control group, again at random. In the 81 treatment villages, AAA started its operations, opening branches and offering products similar to those offered in urban areas (collective and individual loans). In each of the 162 villages, a short preparatory survey was administered either to all

households in villages with up to 100 households or to 100 randomly selected households in larger villages (a total of 15,145 households and 25 variables). This population was then divided into two groups: the "high borrowing propensity" households, corresponding to the last quartile of a propensity score estimated from the preparatory survey, and the “others”. All households in the first group were included in the sample, along with five households in the second group.[1] Two surveys were then conducted on these households: the pre-AAA baseline survey, conducted in four successive waves between 2006 and 2007, and the endline survey conducted two years later. To this sample of 4,465 households, an additional sample of 1,433 households with a "very high propensity to borrow" was added to cope with the low take-up rate observed over time. The latter were drawn from among the households that formed the subject of the short preparatory survey, based on the calculation of a new better-adjusted propensity score taking into account the households that had actually taken out microcredit between the baseline and the endline. They could only be fully interviewed at the endline. The three samples were mixed after calculating new extrapolation coefficients to ensure representativeness at the level of each village. The propensity score was used as an instrument to assess the Local Average Treatment Effect, in addition to Intention to Treat and Treatment on the Treated. This particularity is presented as an essential and unprecedented contribution to this experiment. Figure 1 summarizes the different steps and components of the AAA-RCT protocol.

Figure 1: The AAA-RCT Survey Protocol



Sources: authors

Note: BLS = *BaseLine* Survey; ELS = *EndLine* Survey; HH: Households; PS_n = Score of propensity to borrow, estimated at various stages of the protocol. Q_1-Q_4 = quartiles of estimated propensities (Q_4 is the highest). Non available = not provided by the data available on line. The treatment-control distribution of the 1,300 HH in the pilot survey is unknown.

[1] The actual sample was smaller than the theoretical sample (4,860) due to the fact that some villages had fewer than 100 households.

CDDP's main results can be summarized as follows. The program had no impact on the creation of micro-enterprises – although it boosted existing enterprises, mainly in agriculture – or on various outcomes (income, capital, investment and profits). But neither household income nor consumption improved, due to the reduction in income from paid work outside the household. Positive impacts were also found to be heterogeneous. Microcredit benefited mainly the most profitable income-generating activities, with a negative impact on others. No impacts were observed on women's empowerment or externalities. Lastly, the main conclusion was that, overall, the impact of microcredit was limited and should not be overestimated. In addition, from a methodological point of view, the authors highlighted the sampling strategy they had developed to overcome the low take-up rate observed, suggesting that it could serve as a model for other experiments given the recurrent nature of this problem.

IB- Replication

We conducted three of the four types of replication in the typology proposed by Clemens (2017): a "replication test" (which consists of using the same specifications as the authors on the same sample to verify that the same results are found), itself subdivided into "replication-verification" (to estimate measurement errors: baseline data, recoding and programming) and "replication-reproduction" (to identify sampling errors), and a "robustness test" in the subcategory of robustness reanalysis (by modifying the analyses from recoded variables). The only type of replication not implemented is "robustness-extension", which applies the same procedures to other populations. The results of this replication are detailed elsewhere (Bédécarrats et al., 2019). We summarize the main results here:

Far from the simplicity that is one of the major assets of the RCT methodology and from which it derives its power of seduction, the protocol used differs significantly from this canonical framework. In the field, the experiment is surprisingly complex in terms of methodology (see Figure 1 and below). As a result, it is difficult to get a clear idea of the impact of what and on whom is ultimately being estimated (Bernard et al., 2012).

The results rely primarily on trimming. CDDP state in their paper that no trimming was conducted other than that reported at endline, which is inaccurate. In fact, the applied trimming procedures are inconsistent between the baseline and endline. CDDP trimmed 459 observations (10.3%) at baseline, removing only the most extreme values among those observations. At endline, however, they trimmed 27 observations (0.5%) differently by removing them entirely. Trimming at endline with slightly different thresholds has a strong impact on results. Thresholds below 0.5% produce results with no statistically significant impacts on either self-employment outputs (sales and home consumption) or profits. The logical interpretation would then be that microcredit has no clear impact on self-employment activities. Thresholds above 0.5% generate a statistically significant impact in terms of an increase in expenses and decrease in investment, but no statistically significant impact on profits. It would be harder to produce a coherent interpretation of such results, especially since a decrease in investment is inconsistent with an increase in assets.

Use of the same specifications as the original paper reveals imbalances at baseline. There are large, significant imbalances in the outcome variables used by CDDP to estimate impact at endline. Households in the treatment group made significantly less sales and profits from self-employment than households in the control group. They also invested more. In addition, there are imbalances at baseline with respect to a number of important variables, such as the area of owned land, access to basic services and women's empowerment.

Whether controlling for these imbalances or using the same specifications as CDDP, significant treatment effects are detected on outcomes wherein an impact of microcredit is hardly plausible: household head gender, absence of education and spoken language. This calls into question the reliability of the data and the integrity of the experimental protocol.

There is a series of coding errors in CDDP's do-files. The count of total borrowing at baseline omitted credits from MFIs other than AAA. This same count included solely outstanding loans at the time of the survey, instead of all the loans that had been outstanding in the previous 12 months, as specified in the variable's definition, stated in the published paper and computed at endline. The appraisal of agricultural assets at baseline omitted two types of assets (tractors and reapers), which happen to be the most valuable assets owned by surveyed households. Inclusion of tractors and reapers in asset appraisal increases the sample's average value of agricultural assets per household from 1,377 Moroccan Dirham to 5,111 Moroccan Dirham. The livestock assets total mistakenly included extra dataset columns, effectively adding in non-existent units. Total business earnings omitted some business sales. A number of cases were found of confusion between prices before, during and after harvest when computing agricultural sales and home consumption. A procedure was used for agricultural investment depreciation, which was omitted for other business investments of similar amounts. All in all, these coding errors affect 3,866 of the 4,934 observations (78.35%) used by CDDP (Table 3) for their ATE estimation of self-employment activities. Correcting the coding errors also substantially modifies the estimated average treatment effects.

Measurement errors can be observed in all dataset sections. Credit measures warrant particular attention, as they are essential to characterize the treatment. CDDP collected information from the microcredit institution's information system and appended it to the survey data. This administrative data proves to be largely inconsistent with the data on borrowings collected from the surveyed households. In the treatment group, 11% of respondents said they had borrowed from AAA, whereas the administrative data shows this figure to be 17%. CDDP explain this difference on average in terms of religious shame. Yet this explanation is implausible in most cases. The administrative data used by CDDP identified 435 households as clients, 241 of whom said they did not have an outstanding or matured loan from this MFI. A total of 46 of these 241 households declared a loan from another formal source, so the credit shame argument does not apply to them, although they might have confused credit provider. Now a "credit shame" argument for the remaining 195 households would imply a "credit pride" explanation for the 152 households who reported having a loan from AAA even though they did not appear in the MFI's registers.

CDDP systematically recoded credit from undetermined sources as connection loans from utilities (electricity or water companies), even when supplementary information provided by respondents was inconsistent with such a reclassification. When we reclassify undetermined sources as connection loans from utilities only where such reclassification is supported by the corresponding supplementary information provided by respondents, we find that the experiment is associated with significantly higher access to utility credit in treatment villages. This suggests a probable co-intervention that contaminated the results. That would explain the experiment's large, significant impact on access to drinking water and sanitation, which are not plausibly ascribable to AAA.

Other measurement errors affect variables used in the CDDP regressions and have a direct impact on the identification and impact estimates. Asset appraisal is a notable case. Each asset item was valued by imputing the median prices of all registered transactions for this item in the previous 12 months. Prices for many items are available for only a very small number of transactions, exposing the median to being skewed by outliers or implausible prices reported by the households. This method erratically shifts the unit value of each asset type by several tenths of a percentage point between the baseline and the endline.

There are also sampling errors. Households were sampled based on their answers to a short preparatory survey. But 50.5% of the households surveyed at baseline displayed considerable differences compared to the data collected by the preparatory survey for the exact same variables. Moreover, the sex and age composition of 20.5% of the households interviewed at baseline and supposedly re-interviewed at endline differs to such an extent that it is not plausible that the same units were re-interviewed in these cases. The borrowing propensity score used as the sampling

criterion at baseline totally fails to predict borrowing and is at odds with the revised borrowing propensity scores used as sampling criteria to add new households at endline.

The replication included recomputing the impact estimates by correcting the abovementioned coding errors with resampling, keeping only 3,268 households interviewed both at baseline and endline whose household gender and age composition was compatible between baseline and endline, and after trimming 0.5% of observations using the method reported by CDDP. This produced a smaller, less significant impact on sales and home consumption and a smaller impact on expenses. The impact estimate for profits is no longer significant. Despite these corrections, there are still large imbalances at baseline in sales and profits, household head gender and origin, access to electricity, water and sanitation, and opinions of women's empowerment. There are still disconcerting estimates for other outcomes that are hardly plausible and indicate a lack of data quality and alterations to the protocol and survey sampling.

Our reanalysis focused on the experiment's lack of internal validity, but several of the identified issues also raise concerns about its external validity. The average number of household members grew from 5.17 to 6.13 between the baseline and endline surveys. According to the national census, Moroccan rural households had an average of 6.03 members in 2004 and 5.35 members in 2014, displaying a decreasing pattern contrary to the experiment's observations. Compared to the Moroccan rural population, the study sample also has significantly fewer households headed by women and its measured consumption is 59% lower on average. If these differences do not stem from poor survey data quality, they raise the question as to who this sample was representative of.

A behind-the-scenes exploration of this research, starting with data collection, offers a preliminary explanation for these multiple errors.

II. Behind the scenes of data collection

The sociology of science shows that scientific production, even in "hard" sciences, is inseparable from multiple socio-political dynamics involving a broad diversity of actors, who in turn have multiple and sometimes divergent rationalities, interests and constraints. In this case, the AAA study was conducted, in theory, in an exceptionally favorable context: a highly reputed research laboratory (J-PAL), a donor with expertise in both microcredit and research, and a leading MFI on the Moroccan and global markets (in 2005, AAA was ranked among the 30 "best" MFIs worldwide by the Mix Market) in the midst of an expansion phase in supposedly virgin areas. Although all the conditions were in place for a study of outstanding quality, implementation proved much more complex and problematic. As in any study, the different actors involved did not necessarily react as foreseen in the theoretical protocol. As we shall see here, this was the case for the respondents and their families, but also competing microcredit organizations, local loan officers, the research firm in charge of data collection, investigators and staff in charge of data entry. Actors' interactions led to various discrepancies with the theoretical protocol, which the J-PAL team was unable to anticipate and monitor properly when that was precisely its role.

IIA- Distortion of the protocol: product and sampling

In terms of sampling, as seen above, the research team set out to randomly select 81 pairs of villages in areas where AAA planned to expand. In theory, two rules were supposed to dictate the choice of villages: an operational rule (to avoid disrupting AAA's expansion by focusing on areas far from the new branches, which were therefore isolated, low population density areas guaranteed in principle to be "virgin" credit zones) and a methodological rule (to form pairs of supposedly equivalent villages). The microcredit supply was assumed to be stable and fixed (J-PAL 2006, DI_1: 26). In practice, however, the final sampling did not adhere to these rules.

The "isolation" criterion was not always respected: our qualitative study finds a diversity of contexts, ranging from peri-urban to extremely isolated rural areas (Morvant-Roux et al., 2014). The comparability of remote villages with villages very close to urban centers can therefore be questioned.

The lack of initial credit, a key factor in isolating the effect of microcredit, as asserted by the research team (Crépon 2007, I_7), was not verified, as mentioned above. Both J-PAL and AFD had anticipated risks of contamination in the shape of competing MFIs entering control villages, a much-discussed point (see, for example, AFD 2008, DI_3). However, our replication turns up another form of contamination: the target villages already had access to microcredit at the baseline. At the endline, competing MFIs disappeared, AAA was the main MFI, and the RCT ultimately studied the substitution of AAA for other formal credits. Our qualitative survey helps to explain the withdrawal of other MFIs. AAA's director drew on his charisma and role as president of the Moroccan MFI network to convince the other MFIs not to compete with AAA in the study areas (unaware that they were already there, although they nonetheless withdrew, at least in part). In addition, mid-survey in 2008, the Moroccan microcredit sector was in the throes of a serious default crisis, which saw the entire sector's rural portfolio shrink from 65% to 47% of the total portfolio (Rozas, 2014). When interviewed in March 2018, the former AAA development manager confirmed that, after the crisis, the risk of contamination had considerably diminished.

Another major issue is the low take-up. The J-PAL team had proposed from the outset to focus on households with a "high probability" of taking out microcredit, but had not expected take-up to be so low. In the initial project document, the team anticipated a "very high" participation rate based on AAA's urban experience (J-PAL 2006, DI_1: 13). The problem emerged in the second wave (AFD 2007, DI_2). Participation was both low and heterogeneous, ranging from 0% to 55% depending on the village. To compensate for this low take-up, the J-PAL team made several "tweaks".

The first tweak was to modify the intervention (microcredit supply) by launching further information campaigns, introducing one-off bonuses for agents, and withdrawing the minimum quota for women. Take-up became an "obsession" for both research team and loan officers, who used the term themselves and went to great lengths to convince villagers to take out microcredit. Strategies included pushing back the usual village borders in the hope of finding more clients. This created a lot of confusion and discussion between J-PAL and AAA.

When these measures proved insufficient, the team tweaked the sampling method as mentioned above (modification of prediction models, and addition of new households at endine). Villages with zero take-up were dropped (AFD 2009, DI_4: 2). These adjustments, in actual fact distortions, place a question mark over the study's external validity. The AFD team clearly raised this issue in their own critical paper they published at the end of the experiment: which product was evaluated, since the supply changed as the experiment progressed; and what was evaluated, since the sampling rules constantly changed and were unable to predict borrowing propensities (Naudet et al., 2012)?

IIB- Poor data quality

As stated in the initial project document, data collection and entry were subcontracted to consultancy firm Team Maroc specialized in engineering, but with no experience of statistical surveys whatsoever. An AFD team field mission observed serious data collection dysfunctions at an early stage (AFD 2008, DI_3). These included translation problems because the interviewers did not speak Berber, a language spoken by a large part of the target population. The 2014 general census found that between 30% and 40% of the Moroccan population spoke Berber, with 16% of the population speaking only Berber (including distinct dialects). This rate rose to 80% and even 100% in some

remote rural areas.[1] The interviewers therefore made extensive use of impromptu translators, including local leaders (*mokadem*), raising comprehension and response bias problems (AFD 2008, DI_3).

Another concern was the number of respondents in households and extended families, which again appeared to be improvised depending on the presence and availability of people and their ability to speak. These observations probably explain in part the significant discrepancies between baseline and endline mentioned in the previous section. However, the size of the gap suggests another explanation: some households may not have been the same, as confirmed by our replication. Absence of a precise address calls for precise tracking techniques, which may have been neglected. Some interviewers, constrained by time (and poorly supervised), may simply have interviewed households available at the time of their visit.

At the end of their field mission, the AFD team carefully formulated recommendations to improve the quality of the data collected. Their report was even followed by a letter to François Bourguignon, Director of the Paris School of Economics (which hosts J-PAL Europe), dated 19 May 2008, in which AFD expressed its concerns about the potential repercussions of these shortcomings on the experiment's results. The letter also raised the data entry issues the team had observed: corrections, when made, appeared to be made arbitrarily without necessarily referring to the questionnaires. J-PAL responded to the letter on 19 July 2008, challenging the gravity of the problems and arguing that they did not call into question the internal validity of the experiment. Nevertheless, the next steering committee meeting (January 2009) decided that all the questionnaires already entered (i.e. all the baseline and some endline questionnaires) were to be sent to the French National Institute of Statistics and Economic Studies (INSEE) to be re-entered. This shows the severity of the problem (AFD 2009, DI_4).

The January 2009 steering committee report also highlighted the poor quality of the data (AFD 2009, DI_4). The J-PAL team put the problem down to a lack of financial and human resources (level of education and remuneration). However, all the investigators were highly educated (four years of university or master's degree) for this type of work (AFD 2008, DI_3). But they were obviously not (or not sufficiently) trained up in either household survey tools or this survey's particularities, as any survey would require. Moreover, AFD had granted in full the additional budget requested by J-PAL precisely to enable J-PAL to carry out its activities in the best possible conditions.

At the January 2009 steering committee meeting, the AFD team once again stressed the data collection problems: "The length of the questionnaires, poor interviewer motivation and the lack of supervision may affect the quality of the data collected" (AFD 2009, DI_4, our translation). At the request of J-PAL's project manager, additional supervisors (outside Team Maroc) were recruited to check data quality for the rest of the survey. But the survey was already well underway, since the endline survey had started.

A close examination of the different stages of the experiment, and in-depth knowledge of the field, including the different stakeholders, all of whom have their own motivations and constraints, provides initial insights into the errors listed in the first part of the paper: respondents who were far from convinced of the merits of microcredit and did not want it (thereby strongly distorting the protocol); competing microcredit organizations that needed to be convinced not to enter survey areas, but were actually already there and subsequently withdrew, either at the instigation of AAA

[1] See http://www.axl.cefan.ulaval.ca/afrique/maroc-1demo.htm, last access 23 November 2018.

(and its director) or because of the crisis (which ultimately changed the study focus to the substitution of AAA for other formal credits); loan officers "obsessed" with participation (ultimately changing the microcredit products); a non-specialized consultancy firm; highly educated, but poorly motivated and undertrained investigators who did not always speak the local languages (impairing data quality); members of the households surveyed who interfered during the interviews; and local leaders brought on board as impromptu translators (generating response bias). These different issues, which could have been largely avoided, point above all to a lack of investigative skills and know-how on the part of the J-PAL team and its service providers.

III. Behind the scenes of producing and disseminating scientific knowledge

Our behind-the-scenes exploration explains some of the errors found by our replication. We now need to explore behind the scenes of the scientific fabric in order to fully understand the three paradoxes presented in the introduction to this paper. As the sociology of science has shown, the use of assertiveness and the art of citation and positioning (relying on, refuting or denigrating existing evidence) can turn an assertion that might seem speculative into an irrefutable statement (Akrich et al., 2013). This art of formulation and persuasion is at the heart of the struggle between researchers, laboratories and schools of thought. However, the struggle is fundamentally asymmetrical and cumulative, similar to Bourdieu's concept of capital (Bourdieu, 1976). A statement is all the more credible if it is made by researchers whose credibility is already recognized. And it is much easier for an already credible researcher to resort to assertiveness and denigration (Latour and Woolgar, 1996). The process is all the more cumulative in that research is a specific "market" where "producers" are also "consumers": scientific credibility comes solely from peers (Bourdieu, 1976). We suggest that the gap between the approximation or even casualness of the CDDP method and its persuasive power stems from an overall strategy on the part of J-PAL to build and then secure a monopoly in the field of impact evaluation: this offensive strategy combines omission and assertiveness and consists in freeing themselves from a “data culture “, affirming the method’s irreproachability and passing over criticism, claiming universality and disregarding the importance of context, and finally circumventing certain basic rules of scientific conduct.

IIIA- Becoming free from a "data culture”

Proponents of RCTs in development economics imported the method from the world of medicine without due consideration of the critical discussions, conditions for their use and questions already raised about them in the public health sphere (Krauss, 2018; Labrousse, 2010). They also disregarded the debates specific to data collection. In most quantitative empirical research protocols, there is a division of labor between data collectors and analysts: the former are statisticians, the latter economists (econometricians or thematicians). With few exceptions (Deaton, 1997; Grosh and Glewwe, 2000), few people can occupy both ends of the spectrum. These are full-fledged jobs, requiring distinct skills and training. Statisticians are responsible for the accuracy of the measurement, economists for its relevance, its analysis and the relations and interactions between data. Both activities are essential for the final production of reasonable results, even if the former have less social prestige than the latter (Desrosières, 2013). Here, the multiple errors in data collection and data entry reflect a clear lack of experience and knowledge, as if the purely technical skills required in the second stage (econometrics: addressing bias issues, selection and identification

of a counterfactual) exempted researchers from all the know-how necessary for the first stage (collection of good quality data).
The disconnect between the researchers and the field is another illustration of this. This disconnect is the result of J-PAL's hierarchical organization, which makes for a strict division of labor between project managers, doctoral candidates and field staff (supervisors and investigators). The latter are ultimately given considerable responsibility for which they are not adequately trained (Jatteau, 2018). This division of labor is a technique frequently found in the field of natural and life sciences, but it does not prevent team leaders from staying in regular contact with the data production chain, including for *in vivo* experiments. And they are required to adhere to precise protocols to validate the rigor of the experiments conducted. This is not the case here, as evidenced by the dozens of RCTs that the most prominent RCT leaders are involved in (Bédécarrats et al., 2017). This disconnect was then exacerbated by J-PAL's exceptionally rapid expansion, to which we will return later.

Lastly, any quantitative survey requires a preliminary exploratory step, in the form of pilot surveys and possibly qualitative analyses, to be able to develop a protocol and questionnaires tailored to the local realities. This was planned in the project document (J-PAL 2006; DI_1), but given the poor quality of data collected, the question might be asked as to whether this stage was properly conducted.

IIIB- Ignoring criticism

Whereas J-PAL has used this study to build a universal narrative on the impact of microcredit, AFD has used it to build its expertise on the method and concludes, *a contrario*, that RCTs are inadequate in many cases to measure the impact of development projects. The gap between the conclusions of the two teams is patent: this is the second paradox highlighted in the introduction. But the J-PAL team overlooks it.

As early as 2009, while endline was still in progress, AFD began to publicly share its experience of RCTs, drawing on AAA and another study conducted in Cambodia at the same time (Delarue 2009, I_9). Their conclusions are clear: they highlight the method's challenges in terms of rigorously evaluating impact given the multiple breaches of protocol that the AFD team partially identified (problem of representativeness and product change) and the time constraints that compelled a focus on the short term.

This conclusion was then shared at a number of international events on numerous occasions (12 public interventions, national and international, identified between 2009 and 2013). It formed the subject of an article published relatively quickly, with the effort made to disseminate it in both French (Naudet et al., 2012) and English (Bernard et al., 2012). In this article, the AFD team considers that, ultimately, RCTs are only suitable for small projects, which the authors describe as "*tunnel projects*", with short-term impacts, clearly identified and easily measurable inputs and outputs, unidirectional (A causes B) linear causality and, lastly, not subject to the risks of low participation by targeted populations. Microcredit is not such a case. The AAA study also fed into a 2015 AFD report on impact evaluation (Pamiès-Sumner, 2015). It situates the AFD's lessons in the broader debates of the evaluation community, where a consensus appears to be emerging in favor of methodological pluralism and the need to no longer consider RCTs as the "*gold standard*", but a "*good standard*".

Although the AFD team was fully aware of the dubious quality of the protocol and data - throughout the study, as we have seen, they constantly encouraged J-PAL researchers to correct the situation - they deliberately chose not to openly discuss this aspect. This choice was probably intended to avoid a head-on clash with one of the most powerful global players in the field of evaluation and related

research (J-PAL), as well as the high-level researchers involved in the experiment, and possibly to lend more weight to their own criticism, which was of a different nature.

In other words, the experience has enabled AFD to build up expertise on the topic, both internally - AFD stopped financing RCTs, its evaluation committee endorsed the idea of giving preference to other methods, mixed if possible - but also externally by contributing to the international debates. Yet there can be no doubt as to the asymmetry of the positions: the English version of the article written by the AFD team has been cited only 15 times (Google Scholar; 25 February 2019) to date, almost exclusively by strong critics. The French version has been cited 17 times.

Not only does the article published by CDDP make no mention of the AFD's publications, it also passes over all breaches of the original protocol. All empirical scientific practices use "tweaking" in that field constraints imply adaptation, compromise, approximations and imperfections. However, RCTs have two specific features. The method's alleged simplicity and purity (simple comparison of means) are presented as arguments for superiority over more traditional methods, such as macroeconomic models and microeconometrics based on observational or quasi-experimental data. However, this argument of simplicity contrasts with the extraordinary complexity of the final protocol described in the previous sections, associated with both the requirements of randomization and the multitude of actors involved. It is precisely this complexity that makes RCTs particularly prone to "tweaking". Lastly, if the authors of the study can afford to make this omission, despite the AFD's repeated (and public) warnings, it is because they can get away with it: their capital already built up enables them to place themselves above criticism (Bourdieu, 1976), even if such criticism has been made public.

III3C-Claiming universality and omitting context

Given the many amendments to the sampling protocol, the target population's profile, as we have already seen, is particularly unclear (see also Cull and Morduch, 2017). Contextualization is therefore key to explain the type of population studied (Pritchett and Sandefur, 2015). Contextualization is also essential to bring the collected data to life, since the data "are never given" (Desrosières, 2013). Our replication has already led us to propose a radically different statement: the AAA-RCT does not compare clients and non-clients, but actually substitutes AAA for other sources of credit already available. This statement can be refined by the qualitative study conducted by two of us while endline was in progress. Our study focused, among other aspects, on the use of microcredit (Guérin et al., 2011). Contrary to client statements (taken up by the J-PAL team, CDDP, p. 134), we observed a massive use of microcredit (from 60% to 80% depending on the context) for everyday consumption, durable goods and housing. Our work turned up two reasons for the low use for non-farm entrepreneurship (less than 10%) and livestock (10%-30%): lack of market opportunities for the former (except in peri-urban areas, and in small proportions) and limited expansion opportunities for the latter, due to strong labor and grazing constraints. However, this observation was not incompatible with positive general equilibrium impacts (which our protocol did not enable us to measure): housing, a major consumer of raw materials and local labor, is likely to have strong spillover effects (unlike small retail trade, whose substitution effects are well known). Our conclusions therefore differed significantly from J-PAL, including in their theoretical analysis of the processes at work, albeit with their own limitations, inherent in qualitative analyses. Aware of this, we asked the J-PAL team several times if we could discuss and compare our methods and results. In a further illustration of their rejection of alternative methods of production of evidence, our requests went unheeded. Note also that the frequent use of microcredit for activities that do not generate direct income was already widely recognized (see, for example, Collins et al., 2009). This significantly alters the causal chains underlying the impact processes (and therefore the theory of change used by many RCTs proponents, both in the CDDP article and in the general introduction to AEJ: AE).

IIID- Bypassing certain rules of scientific ethics

In addition to denial of what already exists, we observe the sidestepping of certain basic rules of scientific conduct. While this problem appears to be growing in the scientific community as a whole and is not specific to J-PAL (Heckman and Moktan, 2018), it is particularly patent here.

In the research world, knowledge validation is based on the "peer review" principle, referring to the collective activity of researchers who critically and anonymously judge the work of their peers. Yet, for this to happen, numerous ethical rules must be respected, starting with the management of conflicts of interest between authors and members of journal editorial boards. Editorial favoritism is a recognized and demonstrated process, particularly among economists (Fourcade et al., 2015). It is usually based on close social ties between editors and authors, such as being or having been in the same faculty, having the same PhD, co-publication or PhD supervision (see, for example, Colussi, 2017). But what is observed here takes the practice to the next level and constitutes a real breach of the basic rules on conflicts of interest. The article was published in a journal founded by Esther Duflo - *American Economic Journal: Applied Economics*. She was editor-in-chief of publication at the time. She is not one of the editors of the special issue, but she is co-author of two articles. Two of the three editors of the special issue - Abhijit Banerjee and Dean Karlan - are members of the editorial board and co-authors of an article. Finally, nearly half of the authors of all the articles in the issue (11 out of 25) are also members of J-PAL, and four others are associate researchers or PhD students. There is an important deviation here from the basic principles of conflicts of interest that are supposed to regulate publication processes. And it can be assumed that this close proximity relaxed the usual requirements of evaluation procedures, which would explain why the article was published despite its many shortcomings.

In addition, the question could be put as to whether J-PAL's meteoric rise is compatible with scientific rigor. In sixteen years of activity, J-PAL has accumulated an impressive number of RCTs (947 complete and in progress on 27 February 2019). Their expansion has been surprisingly rapid by usual standards in the scientific field (Ravallion, 2018), bordering more on the sort of growth observed in the business world. Their expansion is clearly aimed at building a monopoly position on the impact evaluation market, but also at building up results to offset the lack-of-external-validity accusation repeatedly levelled at RCT proponents. Combined with highly centralized governance (see above), this growth implies that a handful of researchers head up a considerable number of experiments. This in turn places a question mark over their actual capacity to work on each RCT (and deepens the disconnect with the abovementioned field). At the time of writing (February 2019), Esther Duflo had 64 randomized RCTs under her belt, equal to just over four new RCTs a year. So how much can she really personally put into each of the RCT results she signs? AAA was Bruno Crépon's first RCT and his work was obviously judged satisfactory, since he has worked on 22 RCTs since then and has become one of the central cogs of the randomist network on the French labor market (Jatteau, 2016: 323). The AAA-RCT probably faced less of a risk of disconnect from the field than today, as J-PAL was still in its infancy (the study started three years after J-PAL was set up). The pressure to publish, however, was undoubtedly already strong and even stronger since J-PAL had to prove itself. RCT results are assessed in terms of a "publishable unit" and not publishing at least one article per RCT is well-nigh unthinkable given the amount of money invested (Jatteau, 2016: 423-432).

Conclusion

The purpose of this article was to describe the production chain behind a scientific result, from sampling, data collection, data entry and recoding, estimates and interpretations to publication and dissemination of results. It also aimed to shed light on three paradoxes: the academic success of the AAA-RCT results when its validity, both internal and external, is highly problematic; the contradictory

results of the two research teams concerned: J-PAL uses the study as a basis for drawing universal conclusions, while AFD has dropped RCTs as its main method of impact assessment; and an extraordinarily complex survey protocol when simplicity is precisely the key argument put forward for the alleged superiority of RCTs.

Our analyses reveal, in part, processes traditionally observed in the sociology of science. “Tweaking”, translation, strategic positioning (by means of references to existing literature or their omission), competition and asymmetry of positions are common currency in scientific production. Increasing pressure to publish is behind much abuse. We know just how much the diktat of "publish or perish" is shaping the academic world and causing many failings: salami-slicing, plagiarism, duplication, multiple signatures and even fraud. The observed abuse is merely an illustration of the limitations of an academic system that no longer has the means to assess the quality of the research conducted (Heckman and Moktan, 2018). Neither is it exceptional to find peer review rules sidestepped and conflicts of interest passed over, as we have also mentioned. More broadly, economists in general have a sense of superiority over other social sciences due, among other things, to their closer proximity to the experimental sciences, business and political power. This sense of superiority underpins their epistemological isolation and their claim to "find solutions" (Fourcade et al., 2015).

While J-PAL's practices confirm these different observations, they stand out for what could be described as hubris. Tweaking research protocol is often unavoidable and acceptable, but only if researchers exhibit reflexivity. Translation is inevitable, but only if its terms and conditions are specified and justified. Competition punctuates scientific life. Issue might be taken with excessive competition, but it is probably preferable to monopoly… Yet this is precisely what is at issue here: we are witnessing the pursuit of monopoly using specific techniques designed to impose an ad hoc definition of science, but also to dominate the entire scientific field (recruitment, training and reviews) and its contribution to public debate, as has been shown elsewhere (Bédécarrats et al., 2017; Jatteau, 2016; Ravallion, 2018). Such domination makes it hard for any form of criticism to be heard. The way CDDP was produced reflects and feeds into this position of domination. It enabled J-PAL researchers to ignore the knowledge built up in the field of household surveys, resulting in a patent lack of skills explaining some of the errors observed. This attitude enabled them to pass over multiple breaches of protocol, which seriously affect the validity of the experiment. It enabled them to ignore AFD's criticism, even when it had been publicly presented and published in real time. It enabled them to carry on claiming that the experimental method is simple, pure and therefore superior to all others, even though the final protocol displays various sampling errors and is extraordinarily complex.

Our purpose is not, however, to discredit the RCT method, but to recognize its true value by challenging the pedestal on which it now stands. While RCTs are likely to remain appropriate and legitimate for certain precisely circumscribed policies, they should still be conducted by the book. It is both necessary and possible to use other methods without compromising scientific rigor. As we have seen here, this pluralism should be a requirement, in particular to round out RCTs by contextualizing them, both before data collection and for analysis. Pluralism is also required for all development issues, projects and policies that are not suitable for RCTs, of which microcredit with its relatively well-targeted interventions is a good example given the low take-up and complexity of its effects (Bernard et al, 2012). Unfortunately, for many RCT proponents, and J-PAL in particular, "RCTs are not just top of the menu of approved methods, nothing else is on the menu," (Ravallion, 2018: 13).

How long can the pedestal hold, and when will the bubble burst (Picciotto, 2012)? It could be argued that the Moroccan case is singular and anecdotal. This then calls for more replication studies and detailed analyses of RCT implementation, such as that proposed in this paper. Other critical positions have been taken above and beyond our own analysis, some of them by renowned economists such as Nobel Prize winners Heckman and Deaton. Given the forces at work, these are valuable voices.

List of references

Akrich, M., Strum, S., Callon, M., Latour, B., 2013. Sociologie de la traduction: textes fondateurs. Presses des Mines via OpenEdition.

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

Bédécarrats, F., Guérin, I., Morvant-Roux, S., Roubaud, F., 2019. Estimating microcredit impact with low take-up, high contamination and inconsistent data? International Journal for Re-Views in Empirical Economics.

Bédécarrats, F., Guérin, I., Roubaud, F., 2017. All that Glitters is not Gold. The Political Economy of Randomized Evaluations in Development. Development and Change 0. https://doi.org/10.1111/dech.12378

Bernard, T., Delarue, J., Naudet, J.-D., 2012. Impact evaluations: a tool for accountability? Lessons from experience at Agence Française de Développement. Journal of Development Effectiveness 4, 314–327.

Bourdieu, P., 1976. Le champ scientifique. Actes de la recherche en sciences sociales 2, 88–104.

Clemens, M.A., 2017. The meaning of failed replications: A review and proposal. Journal of Economic Surveys 31, 326–342.

Collins, D., Morduch, J., Rutherford, S., Ruthven, O., 2009. Portfolios of the poor: how the world's poor live on $2 a day. Princeton University Press, Princeton.

Colussi, T., 2017. Social Ties in Academia: A Friend Is a Treasure. The Review of Economics and Statistics 100, 45–50. https://doi.org/10.1162/REST_a_00666

Cull, R., Morduch, J., 2017. Microfinance and Economic Development, Policy Research Working Paper 8252. World Bank, Washington DC.

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

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

Desrosières, A., 2013. Gouverner par les nombres: L'argument statistique II. Presses des Mines via OpenEdition.

Duvendack, M., Palmer-Jones, R., Reed, W.R., 2017. What Is Meant by" Replication" and Why Does It Encounter Resistance in Economics? American Economic Review 107, 46–51.

Fourcade, M., Ollion, E., Algan, Y., 2015. The superiority of economists. Journal of economic perspectives 29, 89–114.

Grosh, M., Glewwe, P. (Eds.), 2000. Designing household survey questionnaires for developing countries: lessons from 15 years of the Living Standards Measurement Study. The World Bank, Washington DC.

Guérin, I., Morvant-Roux, S., Roesch, M., Moisseron, J.-Y., Ould-Ahmed, P., 2011. Analyse des déterminants de la demande de services financiers dans le Maroc rural, in: Séries Analyse d'impact, No.6. AFD, Paris.

Heckman, J.J., 1991. Randomization and social policy evaluation. National Bureau of Economic Research Cambridge, Mass.

Heckman, J.J., Moktan, S., 2018. Publishing and promotion in economics: The tyranny of the top five, in: NBER Working Paper No. 25093. Cambridge, MA.

Jatteau, A., 2018. 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 [On ligne 13 March 2018, Accessed 19 October 2018].

Jatteau, A., 2016. Faire preuve par le chiffre? Le cas des expérimentations aléatoires en économie (PhD Thesis). Paris Saclay.

Krauss, A., 2018. Why all randomised controlled trials produce biased results. Annals of medicine 50, 312–322.

Labrousse, A., 2010. Nouvelle économie du développement et essais cliniques randomisés: une mise en perspective d'un outil de preuve et de gouvernement. Revue de la régulation. Capitalisme, institutions, pouvoirs 7.

Latour, B., Woolgar, S., 1996. La vie de laboratoire: la production des faits scientifiques. La Découverte, Paris.

Morvant-Roux, S., Guérin, I., Roesch, M., Moisseron, J.-Y., 2014. Adding value to randomization with qualitative analysis: the case of microcredit in rural Morocco. World Development 56, 302–312.

Naudet, J.-D., Delarue, J., Bernard, T., 2012. Évaluations d'impact: un outil de redevabilité? Les leçons tirées de l'expérience de l'AFD. Revue d'économie du développement 20, 27–48.

Ogden, T.N., 2017. Experimental conversations: Perspectives on randomized trials in development economics. MIT Press, Cambridge, Massachusetts.

Pamiès-Sumner, S., 2015. Development Impact Evaluation, State of Play and New Challenges (A Savoir). AFD, Paris.

Picciotto, R., 2012. Experimentalism and development evaluation: Will the bubble burst? Evaluation 18, 213–229.

Pritchett, L., Sandefur, J., 2015. Learning from experiments when context matters. American Economic Review 105, 471–75.

Ravallion, M., 2018. Should the Randomistas (Continue to) Rule? Center for Global Development Working Paper 492.

Rodrik, D., 2008. The new development economics: we shall experiment, but how shall we learn?, in: Brookings Development Conference. Presented at the Brookings Development Conference, The Brookings Institution, Washington, DC.

List of unpublished documents

Interventions to conferences and seminars

(I_1): Bernard Tanguy, “Rural micro-finance in Morocco: Lessons from an on-going randomized study”, Study presented to DIME Workshop, Dakar, Feb. 2010 (23 slides).

(I_2): Bernard Tanguy, Delarue Jocelyne, Naudet Jean-David, “On measuring and bridging through impact evaluations: Lessons from AFD’s experience”, presented to NONIE Conference, Bonn, March 2010 (12 slides).

(I_3): Bernard Tanguy, “Impact evaluation of a micro-credit program in Morocco: A donor’s perspective”, presented to DIME Conference, Dakar, May 2010 (24 slides).

(I_4): Bernard Tanguy, Delarue Jocelyne, Naudet Jean-David, “Impact Evaluations and Microfinance Interventions: A donor’s perspective”, presented to CGAP Conference, Nairobi, May 2010 (8 slides).

(I_5): Bernard Tanguy, Delarue Jocelyne, Naudet Jean-David, “On measuring and bridging through impact evaluations: Lessons from AFD’s experience”, presented to EES Conference, Prague, Oct. 2010 (12 slides).

(I_6): Bernard Tanguy, « Mesurer et comprendre par les évaluations d’impact : Leçons d’expérience de l’AFD », Méthodologie presented to Séminaire interne sur l’évaluation d’impact, Paris, Dec. 2010 (12 slides).

(I_7): Crépon Bruno, « Evaluer l’impact du micro-crédit en milieu rural », presented to Présentation aux partenaires, Casablanca, March 2007 (22 slides).

(I_8): Delarue Jocelyne, “Evidence and use: The impact evaluation of microfinance projects and their expected use”, presented to NONIE Meeting, Washington (DC), January 2008 (17 slides).

(I_9): Delarue Jocelyne, “Impact evaluations from a bilateral donor’s perspective”, presented to [meeting to be confirmed], Phnom Penh, June 2009 (11 slides).

(I_10): Delarue Jocelyne, « Les évaluations d’impact à l’AFD », presented to CIRAD - GT Impact, Montpellier, May 2010 (7 slides).

(I_11): Naudet Jean-David, “We Shall Learn But Shall We Use? A Sponsor Perspective on Impact Evaluation”,], March 2009 (10 slides).

(I_12): Naudet Jean-David, « Evaluations d’impact et expérimentation : éléments de positionnement de l’AFD », presented to Séminaire interne sur l’évaluation d’impact, Paris, Dec. 2010 (8 slides).

(I_13): Naudet Jean-David, « Tester ou évaluer les programmes de développement ? Quelques leçons d’expérience de l’AFD sur les évaluations d’impact », presented to NONIE 2012, Paris, March 2012 (11 slides).

(I_14): Pamies Sumner Stéphanie, « Les évaluations d’impact dans le secteur de la microfinance à l’AFD : quelques retours d’expérience », presented to Club Microfinance Paris, Paris, March 2012 (7 slides + report).

(I_15): Pamies Sumner Stéphanie, “Impact evaluations: lessons from AFD’s experience”, presented to SKY evaluation meeting, Phnom Penh, Nov. 2011 (9 slides).

(I_16): Pamies Sumner Stéphanie, « Les évaluations d’impact à l’AFD », presented to Master 2 analyse de projets de développement durable, Rennes, 2012 (14 slides).

(I_17): Pamies Sumner Stéphanie, “Impact evaluations at AFD: lessons learnt and perspectives”, presented to KfW Seminar, Frankfurt, January 2013 (26 slides).

Internal documents: protocols, steering committee minutes, field reports, mails

(DI_1): J-PAL, Evaluation de l’impact d’un programme de micro-crédit en milieu rural: Al Amana au Maroc, Projet de recherche soumis au Département de recherche de l’AFD, 21 janvier 2006, 39 pages.

(DI_2): Réunion Copil 2007, Evaluation de l'impact du microcrédit en milieu rural, note préparée pour le comité de pilotage du 1ier août 2007, AFD, 16 pages.

(DI_3): AFD, Compte rendu de mission sur les enquêtes de terrain de l'évaluation d'impact d'Al Amana, 29 avril 2008, 7 pages.

(DI_4): Réunion Copil 2009, Comité de pilotage de l'Evaluation d'impact d'un programme de microcrédit en milieu rural au Maroc, minutes de la réunion du 29 janvier 2009, AFD, 4 pages.

(DI_5): Lettre du directeur du département de la recherche de l'AFD à Mr. François Bourguignon, 19 May 2008.

(DI_6): Lettre de réponse du J-PAL au directeur du département de la recherche de l'AFD, 18 juillet 2008.