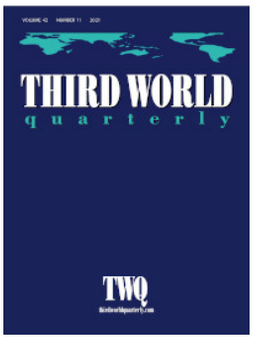

Third World Quarterly

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ctwq20

Behind the scenes of science in action: a 'replication in context' of a randomised control trial in Morocco

Florent Bédécarrats, Isabelle Guérin, Solène Morvant-Roux & François Roubaud

To cite this article: Florent Bédécarrats, Isabelle Guérin, Solène Morvant-Roux & François Roubaud (2021) Behind the scenes of science in action: a 'replication in context' of a randomised control trial in Morocco, Third World Quarterly, 42:11, 2669-2689, DOI: 10.1080/01436597.2021.1977114

To link to this article: https://doi.org/10.1080/01436597.2021.1977114

THIRD WORLD QUARTERLY
2021, VOL. 42, NO. 11, 2669–2689
https://doi.org/10.1080/01436597.2021.1977114

Behind the scenes of science in action: a 'replication in context' of a randomised control trial in Morocco

Florent Bédécarrats[a], Isabelle Guérin[b], Solène Morvant-Roux[c] and François Roubaud[d]

[a]Nantes Metropolis, Nantes, France; [b]IRD-CESSMA (Centre for Social Science Studies on the African, American and Asian Worlds at the French National Research Institute for Sustainable Development), Université de Paris, Paris, France; [c]Graduate School of Social Sciences, University of Geneva, Geneva, Switzerland; [d]LEDa-DIAL (Joint Research Unit CNRS-IRD-Université Paris-Dauphine), Paris, France

ABSTRACT

This article is a 'replication in context' of a flagship randomised control trial (RCT) conducted in Morocco on microcredit. 'Replication in context' consists in combining the quantitative replication of an RCT with a contextualised analysis of its implementation and its political economy, in the sense of the interplay between different stakeholders with divergent and potentially conflicting interests, constraints and powers. 'Replication in context' draws on quantitative and qualitative data and uses the tools of statistics, political economy and sociology of science. This method allows us to describe the entire RCT production chain, from sampling, data collection, data entry and recoding, estimates and interpretations to publication and dissemination of the results. We find that this particular RCT does not respect the key principles of randomisation (imbalanced sampling and contamination) nor those of statistics (coding and measurement problems, poor-quality data and arbitrary trimming procedures). The qualitative analysis highlights the difficulties of implementing a randomised protocol in the real world. Beyond this particular case study, our analyses call into question the supposed superiority of randomised methods, echoing the growing unease in an academic field increasingly struggling to enforce the basic rules of ethics and scientific deontology.

ARTICLE HISTORY
Received 23 June 2020
Accepted 15 August 2021

KEYWORDS
Randomised control trials
sociology of science
power
epistemology
replication
microfinance

Introduction

In October 2019, Abhijit Banerjee, Esther Duflo and Michael Kremer jointly won the 51st Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. The three researchers were awarded 'for their experimental approach to alleviating global poverty' and for having 'turned development economics – the field that studies what causes global poverty and how best to combat it – into a blossoming, largely experimental field' (The Royal Swedish Academy of Sciences 2019, 2). There are several reasons to welcome the prize. Whereas the prize is overwhelmingly awarded to white North American men, this trio includes one woman

CONTACT Isabelle Guérin isabelle.guerin@ird.fr

(Esther Duflo is the second female winner) and one non-white (Banerjee is the second non-white winner). The award brings to the fore the issue of poverty and the collection of primary data, which has long been passed over by development economics. As a response to the thorny counterfactual issue, field experiments, also called randomised control trials (hereafter RCTs), theoretically offer the possibility of precisely isolating and quantifying the impact of a development intervention, all other things being equal.

However, there is also cause to raise questions about the validity and repercussions of the growing use of this method, which the prize may boost further. Among other issues, RCTs are criticised for their inability to extend beyond the particularities of the interventions studied (ie their external validity; Deaton and Cartwright 2018; Heckman 1991a), their incapacity to make an optimal trade-off between bias and precision and to measure externalities (ie their internal validity; Deaton and Cartwright 2018; Heckman 1991b; Ravallion 2020; Rodrik 2009), the lack of rigour in their interpretation, and the massive recourse to storytelling (Kabeer 2019; Labrousse 2020). Another type of critique focuses on the political economy of RCTs. Some scholars argue that RCTs are in fact a new research business model (Bédécarrats, Guérin, and Roubaud 2019b) combining several strategies (Donovan 2018): rhetorical (asserting their superiority), emotional (arousing compassion), organisational (implementing entities specifically dedicated to the management of experiments), and institutional (controlling access to journals and academic positions). This in turn transforms the field of development into a new mode of governmentality (Berndt 2015) and knowledge (Kelly and McGoey 2018), in which 'causal certainty' will eliminate 'human suffering' (Donovan 2018, 27). This paper contributes to this debate by proposing a new method: a 'replication in context'. A 'replication in context' consists in combining the quantitative replication of an RCT (ie. the reproduction of statistical analyses) with a contextualised analysis of the implementation of this RCT and its political economy, in the sense of the interplay between different stakeholders with divergent and potentially conflicting interests, constraints and powers. 'Replication in context' uses the tools of statistics, political economy and sociology of science and aims at retracing the entire RCT production chain, from sampling, data collection, data entry and recoding, estimates and interpretations to publication and dissemination of the results.

We chose an RCT on microcredit conducted in Morocco (Crépon et al. 2015), not only because the microdata were available, which is of course a pre-condition, but also because it is an emblematic RCT in the field of development, for several reasons explained below.

Ultimately, our paper explains three paradoxes. The first is the discrepancy between a multiple error-impaired RCT and its academic success. The second paradox concerns the diametrically opposed conclusions drawn by the RCT's two main actors: the authors and the funder. Crépon et al. (2015) and, more broadly, many 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 fundamental arguments put forward by its promoters to justify the scientific robustness of RCTs, and the high complexity of the protocol actually applied by the case study considered here.

The paper is organised as follows. The second section explains what we mean by 'replication in context' and details the method. The third section summarises the results of the replication and the different shortcomings observed. We find that this particular RCT does not respect the key principles of randomisation (because it falls victim to imbalanced sampling and contamination) nor those of statistics (we identify coding and measurement problems, poor-quality data and arbitrary trimming procedures). The fourth section contextualises the replication by exploring the difficulties of implementing a randomised protocol in the real world, the gap between theory and practice, and the interplay between different stakeholders at the origin of that gap. The fifth section delves further into the analysis and argues that this gap results from the researchers' belief in the superiority of their method (Ravallion 2020). In the conclusion, we draw more general lessons: beyond this specific study, our results call into question the supposed superiority of randomised methods, echoing the growing unease in an academic field increasingly struggling to enforce the basic rules of ethics and scientific deontology.

A 'replication in context'

The scientific community is increasingly recommending systematic replication to strengthen the robustness of empirical research. A replication is a 'study whose main purpose is to determine the validity of one or more empirical results from a previously published study' (Duvendack, Palmer-Jones, and Reed 2017, 47). In the field of development economics, replications are still rare (Sukhtankar 2017). Qualitative analysis has already been carried out on a few RCTs (Kabeer 2019; Morvant-Roux et al. 2014; Quentin and Guérin 2013; Rao, Ananthpur, and Malik 2017), providing alternative interpretations to those of RCT researchers. What we suggest here is different: a 'replication in context' mobilises qualitative data and qualitative analysis to contextualise the implementation of the RCT itself and explore the mechanisms underlying the scientific production of the experiment. Through a combination of methods, a 'replication in context' makes it possible to scrutinise science in action. In so doing, it offers a fresh look at the fabric, in the field, of experimental methods, from the collection of data to the publication of results.

We focus on an RCT involving rural microcredit supplied by a leading Moroccan microcredit institution (Al Amana, hereafter AAA), published by Crépon et al. (2015). There are a number of reasons for the choice of this case study (AAA-RCT, hereafter). Naturally, the availability of data, both quantitative and qualitative, was a primary requirement. The AAA-RCT micro-data are freely available, making for an extremely thorough replication. We also had the unique opportunity to access various qualitative data, which are described below. However, our choice was also motivated by the fact that this is a flagship RCT, for a number of reasons. Microcredit is a key RCT focus in the development field: at The Abdul Latif Jameel Poverty Action Lab (J-PAL) (the most active research centre in RCTs), 298 of its 1068 RCTs (complete or in progress as of 10 March 2021) are on financial issues.[1] An initial summary of RCT findings on microcredit was published in 2015 in a special issue of the *American Economic Journal: Applied Economics* (Banerjee, Karlan, and Zinman 2015). This summary was seen as a decisive contribution to settling a long-standing debate on microcredit impact (Ogden 2017). Five years after its publication, the special issue has already been cited 4963 times (Google Scholar, March 10, 2021). The AAA-RCT is one of six studies in the special issue, and exhibits many strengths that should make it an exemplary RCT.

The basic principle of an RCT is, in theory, simple. It consists in randomly assigning individuals to two groups within a homogeneous population. The first receives a 'treatment' (here, a microcredit); the second receives a placebo, a different intervention, a lesser degree of exposure to the treatment, or nothing (here, nothing). After a certain period of time, the two groups are compared in order to assess the effectiveness of the intervention (or to analyse two distinct modalities, or a different degree of exposure to the treatment). Here, the objective is to compare areas with and without microcredit, and this implies a number of conditions. The first is the existence of areas that are free of any intervention and remain so throughout the intervention. Here, the context seems ideal, since at the time the study was launched, the microcredit provider was planning to expand into remote rural areas, which are therefore assumed to be microcredit free. In addition, measures had been taken to prevent competing microcredit organisations from entering the treatment area. As stated by the authors in the introduction to their article: '[The study] takes place in an area where there is absolutely no other microcredit penetration, before or after the introduction of the product, and for the duration of the study' (Crépon et al. 2015, 124). Randomisation is a second key condition: when randomised on a large sample size, the treatment and control populations are expected be identical in both observable variables (poverty, number of children, education, etc.) and 'unobservable' variables (eg confidence in business, attraction for entrepreneurship, etc.).

Another condition is that take-up is high enough for there to be enough difference between the control and treatment groups, and in turn sufficient statistical power to detect significant effects. The use of models to predict which populations are more likely to take up the intervention may be one way to compensate for low take-up. These methods are just beginning to emerge, and the method elaborated by the AAA-RCT is considered a pioneer (Banerjee, Karlan, and Zinman 2015). The AAA-RCT also innovates by measuring externalities, ie the effects of microcredit on non-clients. Other basic conditions naturally include surveying the same populations before and after the intervention, as well as the stability of the intervention (if it changes over time, then the analysis is not clear as to what it is assessing). Last but not least, other basic rules of statistical and econometric analysis include appropriate coding of variables, elimination of outliers that may affect the mean (known as trimming) and, when trimming changes the results, a discussion explaining the choices that are ultimately made.

Given that the AAA-RCT is presented as innovative, while being equated with the canonical ideal of an RCT, the special issue's introduction relies extensively on Crépon et al. (2015) 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. The paper of Crépon et al. (2015) is clearly an academic success, with 445 citations five years after publication (Google Scholar, 10 March 2021). We also note that the paper was written by some of the most prestigious RCT proponents, which is taken as a guarantee of quality.

Our replication (published elsewhere as a companion paper: Bédécarrats et al. 2019a) shows that the basic principles of an RCT are not in fact respected by the AAA-RCT in terms of sampling, balance between treatment and control populations, and non-contamination. The theoretical foundations of RCTs (simplicity of the method, comparison of comparable villages with and without microcredit, rigorous analysis and good data quality) do not hold: the protocol is extremely complex, the samples are biased, the data quality is poor and the statistical treatments (trimming) are not rigorous.

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

- Two of us participated in a qualitative field study designed specifically to complement the AAA-RCT. In 2009, at the time of the endline survey, 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). This provided a unique opportunity to observe the actual implementation of the RCT.
- This qualitative study was commissioned and financed by the same donor as the experiment: the Agence Française de Développement (French Development Agency or AFD). This gave us 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 the RCT research team, email exchanges, and academic articles published by AFD researchers drawing lessons based on their experience.[2] 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 some members of the RCT research term. 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 RCT research team declined the invitation to collaborate. Although this obviously restricted the analysis, access to the grey literature nevertheless captured their point of view.

These various data explain the results of our replication and the gap between theory and practice. Our observations of the experiment's implementation in real time and over time take us behind the scenes of 'science in action'.

Scientific research is often seen as a rational process, motivated primarily by the pursuit of rigour, objectivity and excellence, and regulated by performance: the most rigorous theories best able to explain and predict the world are likely 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 (Desrosières 2002; Latour and Woolgar 1979). Evidently, empirical data, technical and analytical skills, and possibly theory, come into play. Yet the power of persuasion and the 'rhetoric', ie ensuring that a result is convincing and disseminated, is just as fundamental (McCloskey 1998). The unique data to which we had access enabled us to reconstruct the entire results production chain, from the design of the protocol to the dissemination of the results. We show that these results are contingent on a series of trade-offs, constraints, interpretations, actor interactions and power games.

The RCT and its 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. AAA had already begun to expand into rural areas. The RCT focussed on 'remote areas', where AAA was then expanding. The

experiment took place in 162 villages where 4465 households were surveyed at baseline between 2006 and 2007.

The main results of the AAA-RCT can be summarised as follows. The programme 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 benefitted 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 discussed 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.

We conducted three of the four types of replication in the typology proposed by Clemens (2017): a 'replication test' (which consists in 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 consists in reiterating the same study on other populations. The results of our replication are detailed elsewhere (Bédécarrats et al. 2019a). We summarise the main results here.

Protocol complexity and confusion about who and what is being evaluated

Far from the simplicity that is one of the major assets of the RCT methodology, the protocol used differs significantly from this canonical framework. Compared to a classical RCT, it is particularly complex since it relied on a prediction model that was supposed to compensate for the low take-up rate, and it was adjusted during implementation to add 1433 more households to the sample at endline, to further cope with the low take-up. The different steps and components of the AAA-RCT protocol are described in the Appendix and summarised in Figure 1. Complexity is not in itself a problem. The problem is that, ultimately, as we will see below, it is difficult to get a clear idea of the impact of what is ultimately being estimated, and on whom (Bernard, Delarue, and Naudet 2012).

Substantial and significant imbalances between treatment and control groups

As noted above, the random draw theoretically ensures that the populations are similar. The purpose of randomisation is to minimise imbalances between the treatment and control populations, so the differences observed at endline stem exclusively from the effect of treatment. Yet we found substantial and significant imbalances in the baseline for a number of important variables, including the RCT's outcome variables. Possibly in relation to this, we estimated implausible 'treatment effects' on certain variables, eg on household head, gender and spoken language. We also found sampling errors. For example, sex and age composition for 20% of the households interviewed at baseline and reportedly re-interviewed at endline

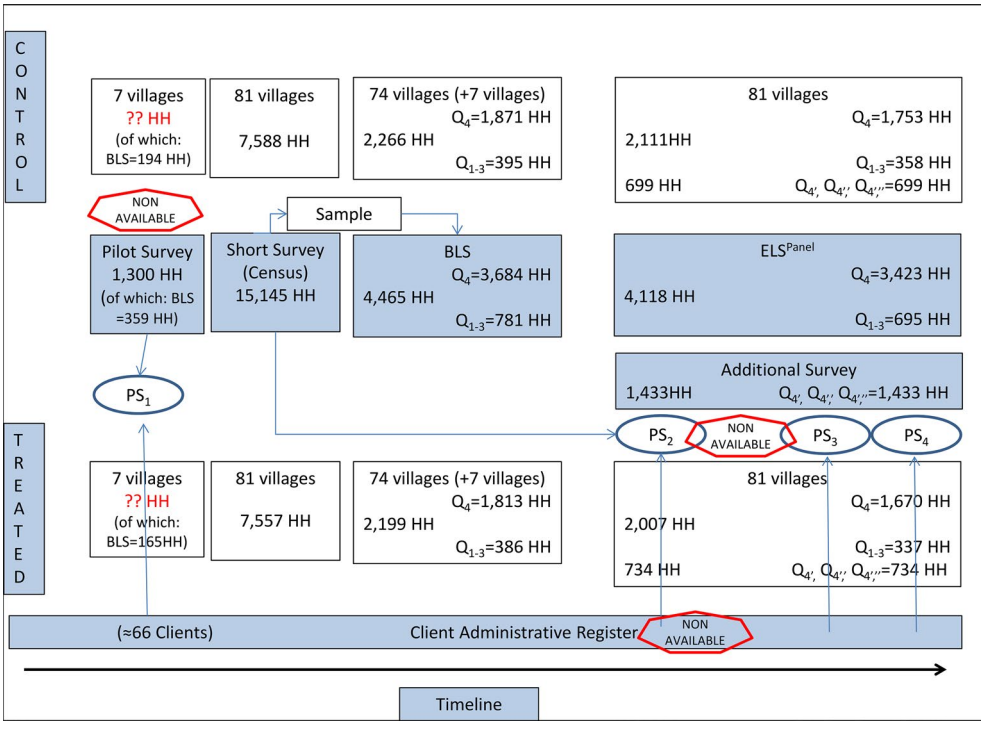

Figure 1. The AAA-RCT survey protocol.

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 (Q4 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.

differs to such an extent that it is implausible that the same households were re-interviewed in these cases. In addition, we found that sample characteristics differed in substantial ways from the population's characteristics. The number of household members in the sample grew on average from 5.17 to 6.13 between the baseline and endline surveys. The national census, however, reported that Moroccan rural households shrank from an average of 6.03 members in 2004 to an average of 5.35 members in 2014. Such discrepancies raise questions about the sample's representativeness, and hence undermine the external validity of this study.

Contamination

Another basic principle of RCTs is to compare a population with and without intervention (treatment group/control group). Measuring who had the intervention (in this case, microcredit) and who did not is crucial to ensure that there was no 'contamination'. Replication highlights two results. First, data on access to microcredit are unreliable, which is critical to the integrity of the analysis. Rather than using their own survey for this variable, the researchers used the data provided by the microcredit provider, believing it to be more reliable. Yet there is no reason why this should be the case, for reasons explained in detail in our companion paper (Bédécarrats et al. 2019a). Furthermore, if we use the survey data, we observe that both populations already had access to formal credit at baseline, albeit different credit than AAA.

Access to credit remained stable in the treatment group between baseline and endline, while it was decreasing in the control group, for various reasons explained later. Our results challenge the very meaning of this RCT: what was tested was not the impact of the introduction of microcredit in 'virgin' areas, but rather the replacement of other formal sources with one microcredit source in the treatment group and credit rationing in the control group.

Coding errors

We documented numerous coding errors. For instance, the appraisal of agricultural assets at endline omitted two types of assets (tractors and reapers), which happen to be the most valuable assets owned by surveyed households. These assets had been included in the household asset appraisal at baseline, and their removal at endline was never mentioned by the authors. Inclusion of tractors and reapers in asset appraisal increases the sample's average value of agricultural assets per household by 470% (from 1377 Moroccan dirhams to 5111 Moroccan dirhams). The identified coding errors altered some 80% of the observations. Yet assets are a key variable in the impact analysis.

Arbitrary trimming

Trimming (ie discarding households that display remarkably high values on some variables) can significantly alter the results of statistical and econometric analysis. Here, not only did the researchers use an inconsistent trimming technique, but the – arbitrary – choice of threshold modified the results considerably. Crépon et al. (2015) reported a balanced sample at baseline after removing extreme values on 24 variables over 459 observations (10.3% of the sample). At endline, however, they trimmed 27 observations (0.5% of the sample) differently by removing them entirely. Moving the endline trimming threshold by just 0.2% (removing a dozen observations, more or less) produces radically different results in terms of sales, expenses, investment and profits. No other trimming threshold would have produced results consistent with their published findings, and no other paper in the same special issue used a similar trimming method or threshold.

The authors produced a reply to our replication, entitled 'Rejoinder', rejecting most of the errors we documented. They referred to our analysis, but they do not appear to have replicated or closely analysed its statistical content, and their rejoinder therefore contains numerous factual errors and omissions. We published a review of their main arguments in response to our replication.[3] We found that all the coding, measurement and sampling errors documented in our replication still hold.

We have seen that some basic principles of randomised experimentation and statistical analysis were not respected. Qualitative analysis helps us understand how such errors could have occurred, and also reveals that other basic principles were not respected either.

Behind the scenes of data collection

The sociology of science shows that scientific production, even in the life sciences, is inextricably linked with 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-RCT was conducted, in theory, in a very favourable context: a highly reputable research laboratory (J-PAL); a donor with expertise in both microcredit and research; and a leading microfinance institution (MFI) on the Moroccan and global markets (in 2005, AAA was ranked among the 30 'best' MFIs worldwide by Mix Market, a worldwide microcredit platform) 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. In fact, the different actors involved did not necessarily react as foreseen in the theoretical protocol. Such was the case for respondents and their families, but also competing microcredit organisations, local loan officers, the research firm in charge of data collection, investigators and staff in charge of data entry. Actors' interactions gave rise to various discrepancies with key aspects of the theoretical protocol, which the research team did not anticipate or monitor properly when that was precisely its role.

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 focussing 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).[4] In practice, however, the final sampling did not adhere to these rules. The 'isolation' criterion was not always fulfilled: our qualitative study finds a diversity of contexts, including suburban villages that do not meet the criteria of 'remote areas' at all (Morvant-Roux et al. 2014).

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 the RCT research term and AFD had anticipated risks of 'contamination' (where the control population has access to the intervention) in the shape of competing MFIs entering control villages, a much-discussed point (see eg AFD 2008, DI_3). However, our replication turns up another form of contamination: the target villages already had access to microcredit at baseline. At endline, competing MFIs had disappeared, AAA was the main MFI, and the RCT ultimately studied the substitution of AAA for other formal credits. Our qualitative survey helps 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 diminished considerably.

Another major issue is the low take-up. The RCT research 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

RCT research team made several modifications. The first modification was to change 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 had to resort to multiple strategies to convince villagers to take out microcredit. These included awareness campaigns, expanding the selection criteria and pushing back the usual village borders in the hope of finding more clients. As these measures proved insufficient, the team modified the sampling method as mentioned above (modification of prediction models, and addition of new households at endline). Villages with zero take-up were dropped (AFD 2009, DI_4: 2). These adjustments call into question the study's external validity. The AFD team clearly raised this issue in their own paper published at the end of the experiment: which product was evaluated, since the supply changed as the experiment progressed; and what were the evaluation criteria, since the sampling rules changed constantly and were unable to predict borrowing propensities (Naudet, Delarue, and Bernard 2012)?

Poor data quality

Data collection and entry were subcontracted to consultancy firm Team Maroc, specialised in engineering but with no experience of statistical surveys whatsoever. An AFD field mission team 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.[5] The interviewers therefore made use of impromptu translators, including local leaders (*mokadem*), raising problems of comprehension and response bias (AFD 2008, DI_3).

Another concern was the number of respondents in households and extended households, which 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 from one period to the other, 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 perhaps poorly supervised), may simply have interviewed households available at the time of their visit. At the end of their field mission, the AFD team made carefully formulated recommendations to improve the quality of the data collected. Their report was even followed by a letter to the 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. The RCT research team 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 (ie 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 RCT research 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 a master's degree) for this type of work (AFD 2008, DI_3). But they were obviously not (or not sufficiently) trained 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 brought up the data collection problems.[6] Additional supervisors (outside Team Maroc) were recruited by the RCT research team 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 had 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 organisations that needed to be convinced not to enter survey areas, but were actually already there and subsequently withdrew, either at the instigation of AAA (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 to increase take-up); a non-specialised consultancy firm; highly educated but undertrained investigators who did not always speak the local language (impairing data quality); members of the households surveyed who interfered during the interviews; and local leaders brought on board as impromptu translators (and thus likely to generate response bias).

Behind the scenes of scientific knowledge production and dissemination

Our behind-the-scenes exploration of the data collection explains some of the errors found by our replication. We now need to explore behind the scenes of the scientific fabric 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 (McCloskey 1998). This art of formulation and persuasion is at the heart of the struggle among researchers, laboratories and schools of thought. However, the struggle is fundamentally asymmetrical and cumulative, similar to Bourdieu's concept of capital (Bourdieu 1975). A statement is all the more credible if it is made by researchers whose credibility is already recognised. And it is much easier for an already credible researcher to resort to assertiveness and denigration (Latour and Woolgar 1979). 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 1975).

Loose compliance with statistical good practices

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). They also disregarded the debates specific to data collection. In most quantitative empirical research protocols, there is a division of labour between data collectors and analysts: the former are statisticians, the latter economists (econometricians or mathematicians). 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 2002). Here, the multiple errors in data collection and data entry reflect a clear lack of experience with data collection best practices, 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. The RCT research team belongs to J-PAL, in which there is a strict division of labour 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), and that is probably what happened here.

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, it might be asked whether this stage was properly conducted.

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, on the contrary, 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 research 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 (12 public presentations, national and international, between 2009 and 2013). It was published as an academic paper relatively quickly, in both French (Naudet, Delarue, and Bernard 2012) and English (Bernard, Delarue, and Naudet 2012). In that paper, the AFD team considers that RCTs are ultimately 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 is emerging in favour of methodological pluralism and the need to no longer consider RCTs as the 'gold standard', but a 'good standard'.

In other words, the experience enabled AFD to build up expertise on the topic, both internally – at least for a while, AFD stopped financing RCTs and 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 two versions of the AFD article have each been cited only 22 times to date (Google Scholar, 10 February 2021).

Not only do Crépon et al. (2015) make no mention of the AFD's publications, they also pass 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 (simple comparison of means) is presented as an argument for superiority over more traditional methods, such as macroeconomic models and microeconometrics based on observational or quasi-experimental data. But this argument of simplicity contrasts with the extraordinary complexity of the final protocol described in the previous sections, due to both the requirements of randomisation and the multitude of actors involved. It is precisely this complexity that makes RCTs particularly prone to 'tweaking'. Lastly, the question might be asked as to why the authors of the study can afford to make this omission, despite the AFD's repeated (and public) warnings. Given their high profile, it may be assumed that they can get away with it: with their capital already built up, they can place themselves above criticism (Bourdieu 1975), even if such criticism has been made public.

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 Wydick 2016). Contextualisation is therefore key to explain the type of population studied (Pritchett and Sandefur 2015). 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 focussed, among other aspects, on the use of microcredit (Morvant-Roux et al. 2014). Contrary to client statements (reported by the RCT research team: Crépon et al. 2015, 134), we observed a massive use of microcredit (from 60% to 80% depending on ecotype systems) for everyday consumption, durable goods and housing. Our work turned up two reasons for the low use of microcredit for non-farm entrepreneurship (less than 10%) and livestock (10–30%): lack of market opportunities for the former (except in suburban areas, and then in small proportions) and limited expansion opportunities for the latter due to strong labour 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 labour, is likely to have strong spillover effects (unlike small retail trade, whose substitution effects are well known). Our conclusions therefore differed significantly from those

of Crépon et al. (2015), 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 RCT research team several times if we could discuss and compare our methods and results. They were not interested in collaborating. Note also that the frequent use of microcredit for activities that do not generate direct income was already widely recognised (see eg Collins et al. 2009). This significantly alters the causal chains underlying the impact processes (and therefore the theory of change used by many RCT proponents, both in Crépon et al. (2015) and in the general introduction to the special issue (Banerjee, Karlan, and Zinman 2015).

Bypassing certain rules of scientific ethics?

A question may also be raised concerning respect for certain basic rules of scientific conduct. This problem appears to be growing in the scientific community as a whole (Heckman and Moktan 2020). 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 need to be respected, starting with the management of conflicts of interest between authors and members of journal editorial boards. Editorial favouritism is a recognised and demonstrated process, particularly among economists (Fourcade, Ollion, and Algan 2015). It is usually based on close social ties between editors and authors, such as being members or former members of the same faculty, having the same PhD, and co-publication or PhD supervision (see eg Colussi 2018). Here, the article was published in a journal founded by one of the authors – *American Economic Journal: Applied Economics*. The author was editor-in-chief of the publication at the time and co-author of two articles in the special issue in which Crepon et al. 2015 was published. Two of the three editors of the special issue were 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) were also members of J-PAL, and four others were associate researchers of J-PAL or PhD students supervised by J-PAL members. Is there not a conflict of interest here, which might explain why the article was published despite its many shortcomings?

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 was based on unique data, allowing for a 'replication in context' of a randomised trial, combining the replication of quantitative data with a qualitative analysis of the implementation of this randomised trial in the field and in real life. 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 problematic; the contradictory results of the two research teams concerned: the RCT research team used the study as a basis for drawing universal conclusions, while the funder has dropped RCTs as its main method of impact assessment; and a 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. Other evaluation methods are also not free of bias or errors.[7] 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 (Necker 2014). 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 2020). 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 dismissive attitude underpins their epistemological isolation and their claim to 'find solutions' (Fourcade, Ollion, and Algan 2015).

The scientific conduct observed here clearly reflects these two aspects, both the growing difficulty of economists to respect the basic principles of the scientific profession (Necker 2014) and their dismissive attitude towards other disciplines and methods, which is expressed here amongst economists themselves (Kvangraven 2020).

What can we learn from this analysis and from the use of RCTs in the field of microcredit and more broadly in development economics? An analysis of five other microcredit RCTs, published in the same special issue, finds similar difficulties: low take-up and compliance resulting in low statistical power; sampling constraints resulting in highly specific populations; and loose interpretation of results implying a highly specific, but not at all explicit, theory of change and poverty (Bédécarrats, Guérin, and Roubaud 2020). An overview of the contribution of RCTs in the field of global health, sanitation and governance also highlights numerous limitations. In the end, it seems that the randomised method is most useful for testing behavioural responses to different forms of intervention, whereas impact analysis, which was its initial *raison d'être*, is much more challenging (Morduch 2020).

Our purpose is not to reject RCTs, since they remain appropriate and legitimate for certain precisely circumscribed policies. However, they should still be conducted by the book, take their feasibility and ethical implications seriously by aligning with best practices established in the medical world, and interface with other methods. Although RCTs remain fit and proper for certain precisely defined policies, other methods can and should be used too.

Quantitative observational data (ie of a population for which the researcher does not control the treatment) can be just as statistically rigorous (Ravallion 2020). It is also crucial to rebalance research efforts to take in other components of the analytic chain: what might have been gained (in theory) in terms of causal attribution, and overinvestment in this area, has left other equally important aspects by the wayside. First and foremost, there is the question of data quality, all too often sacrificed out of a lack of interest and competence. At the same time, closer attention should be paid to the question of sample designs. All too often, the implications of the use of complex sample designs are overlooked.

Last but not least, when dealing with complex causal chains, which is the case with many development interventions, qualitative methods (semi-structured interviews, focus groups, participant observation, ethnography, case studies, life stories, etc.) are often the only way to really address the thorny question of causality (White and Masset 2018). This also supposes an epistemological break: the objective is no longer to set out to lay down universal laws, but to explain causal links specific to a particular time and place.

Acknowledgements

We sincerely thank the three anonymous reviewers and the editors of the journal for their constructive comments.

Disclosure statement

This research is not the result of a for-pay consulting relationship. One of the four authors participated in it while working for the Agence Française de Développement Evaluation Unit, which is part of the AFD Research Department. The AFD Research Department was the main funder of the initial RCT replicated in this paper. AFD evaluation and research activities are supposed to be independent of the institution's operational and financial interests. Microfinance is a very small part of AFD's portfolio and does not constitute a substantial financial interest for the institution. The other co-authors did not receive any funding for this replication analysis, but two of them did receive funding from AFD to perform the qualitative component of the RCT, the central topic of this paper, between 2009 and 2010.

Funding

Solène Morvant-Roux is funded by the Swiss Foundation for National Science (SNSF) Professorship Sponsorship Program [grant number PP00P1 163774].

Notes on contributors

Florent Bédécarrats holds a PhD from the University of Paris-Sorbonne. Since late 2019, he has been heading the data management unit at Nantes Metropole. Most of his work on this book was done while being in charge of coordinating scientific impact evaluations at the French Development Agency. He held this position from 2013 to 2019. From 2007 to 2013 he was in charge of research and development activities at CERISE, a platform of microfinance support organisations. Previously, he worked for three years in Latin America, in a solidarity-based company for tourism and culture in Brazil, for a network of microfinance cooperatives in Mexico, and for an international non-governmental organisation in Guatemala.

Isabelle Guérin PhD, is a socioeconomist, Senior Research Fellow at the French Institute of Research for Sustainable Development (IRD), Associate at the French Institute of Pondicherry and former member of the School of Social Sciences at the Institute for Advanced Study, Princeton (2019–2020). She specialises in the political and moral economics of money, debt and finance. Her current work focuses on the financialisation of domestic economies, looking at how financialisation produces new forms of inequalities and domination, but also alternative and solidarity-based initiatives. Her work draws most often from her own field-based original data, combines ethnography and statistical analyses and is interdisciplinary and comparative in nature. Her work also includes a permanent thinking about the conditions of data production and the combination of methods.

Solène Morvant-Roux is currently Assistant Professor at the University of Geneva. She is funded by a grant from the Swiss National Research Foundation. She is the Principal Investigator of several research projects on financialisation through debt in Mexico and in Switzerland.

François Roubaud is an economist and statistician, Senior Research Fellow at the French Institute of Research for Sustainable Development (IRD), a member of the DIAL research unit in Paris and its former head (2000–2004). He holds a PhD in economics from the Paris-Ouest Nanterre University and is a graduate of the Paris Graduate School of Economics, Statistics and Finance (ENSAE). In statistics, he initiated the mixed surveys approach (household-enterprise) to measure the informal economy, in particular the *1-2-3 survey*, and developed the governance modules

grafted on official household surveys now used to monitor SDG 16. Both are recognised as international standards and implemented in dozens of LDCs (in Africa, Latin America and Asia). In development economics, his main fields of expertise are labour market and informal economy, corruption, governance and institutions, and impact evaluation and political economics of development policies.

Notes

1. Data available from the J-PAL website (https://www.povertyactionlab.org/evaluations), accessed on 7 March 2021.
2. 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' and internal documents are noted 'ID'. Lists of both are given in the Appendix.
3. Both papers (the response of Crépon et al. and our response to their response) can be found here: https://dial.ird.fr/publications/documents-de-travail-working-papers#chapitre_2
4. The notes and reports cited here are listed in the Appendix.
5. 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. See http://www.axl.cefan.ulaval.ca/afrique/maroc-1demo.htm, last accessed on 28 March 2019.
6. '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).
7. In the field of development, see for instance (Camfield, Duvendack, and Palmer-Jones 2014).

ORCID

Florent Bédécarrats http://orcid.org/0000-0003-1001-5540
Isabelle Guérin http://orcid.org/0000-0002-4476-0074
Solène Morvant-Roux http://orcid.org/0000-0003-2609-268X
François Roubaud http://orcid.org/0000-0003-2234-5256

Bibliography

Banerjee, A., D. Karlan, and J. Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics* 7 (1): 1–21. doi:10.1257/app.20140287.

Bédécarrats, F., I. Guérin, S. Morvant-Roux, and F. Roubaud. 2019a. "Estimating Microcredit Impact with Low Take-up, High Contamination and Inconsistent Data?" *International Journal for Re-Views in Empirical Economics* 3: 1–22. doi:10.15456/iree.2019071.090421.

Bédécarrats, F., I. Guérin, and F. Roubaud. 2019b. "All That Glitters Is Not Gold. The Political Economy of Randomized Evaluations in Development." *Development and Change* 50 (3): 735–762. doi:10.1111/dech.12378.

Bédécarrats, F., I. Guérin, and F. Roubaud. 2020. "Microfinance RCTs in Development: Miracle or Mirage?" In *Randomized Control Trials in the Field of Development: A Critical Perspective*. 186–226. London: Oxford University Press.

Bernard, T., J. Delarue, and J.-D. Naudet. 2012. "Impact Evaluations: A Tool for Accountability? Lessons from Experience at Agence Française de Développement." *Journal of Development Effectiveness* 4 (2): 314–327. doi:10.1080/19439342.2012.686047.

Berndt, C. 2015. "Behavioural Economics, Experimentalism and the Marketization of Development." *Economy and Society* 44 (4): 567–591. doi:10.1080/03085147.2015.1043794.

Bourdieu, P. 1975. "The Specificity of the Scientific Field and the Social Conditions of the Progress of Reason." *Social Science Information* 14 (6): 19–47. doi:10.1177/053901847501400602.

Camfield, L., M. Duvendack, and R. Palmer-Jones. 2014. "Things You Wanted to Know about Bias in Evaluations but Never Dared to Think." *IDS Bulletin* 45 (6): 49–64. doi:10.1111/1759-5436.12112.

Clemens, M. A. 2017. "The Meaning of Failed Replications: A Review and Proposal." *Journal of Economic Surveys* 31 (1): 326–342. doi:10.1111/joes.12139.

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

Colussi, T. 2018. "Social Ties in Academia: A Friend is a Treasure." *The Review of Economics and Statistics* 100 (1): 45–50. doi:10.1162/REST_a_00666.

Crépon, B., F. Devoto, E. Duflo, and W. Parienté. 2015. "Estimating the Impact of Microcredit on Those Who Take It up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics* 7 (1): 123–150. doi:10.1257/app.20130535.

Deaton, A. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Washington: The World Bank.

Deaton, A., and N. Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21. doi:10.1016/j.socscimed.2017.12.005.

Desrosières, A. 2002. *The Politics of Large Numbers: A History of Statistical Reasoning*, 1st edition 1993. Translation and edited by C. Naish. Cambridge, MA: Harvard University Press.

Donovan, K. P. 2018. "The Rise of the Randomistas: On the Experimental Turn in International Aid." *Economy and Society* 47 (1): 27–58. doi:10.1080/03085147.2018.1432153.

Duvendack, M., R. Palmer-Jones, and W. R. Reed. 2017. "What Is Meant by" Replication" and Why Does It Encounter Resistance in Economics?" *American Economic Review* 107 (5): 46–51. doi:10.1257/aer.p20171031.

Fourcade, M., E. Ollion, and Y. Algan. 2015. "The Superiority of Economists." *Journal of Economic Perspectives* 29 (1): 89–114. doi:10.1257/jep.29.1.89.

Grosh, M., and P. Glewwe, eds. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. Washington, DC: The World Bank.

Heckman, J. J. 1991a. *Randomization and Social Policy Evaluation*. Cambridge, MA: National Bureau of Economic Research.

Heckman, J. J. 1991b. *Randomization and Social Policy Evaluation*. Cambridge, MA: National Bureau of Economic Research.

Heckman, J. J., and S. Moktan. 2020. "Publishing and Promotion in Economics: The Tyranny of the Top Five." *Journal of Economic Literature* 58 (2): 419–470. doi:10.1257/jel.20191574.

Jatteau, A. 2018. "The Success of Randomized Controlled Trials: A Sociographical Study of the Rise of J-PAL to Scientific Excellence and Influence." *Historical Social Research/Historische Sozialforschung* 43: 94–119.

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. doi:10.1080/19452829.2018.1536696.

Kelly, A. H., and L. McGoey. 2018. "Facts, Power and Global Evidence: A New Empire of Truth." *Economy and Society* 47 (1): 1–26. doi:10.1080/03085147.2018.1457261.

Krauss, A. 2018. "Why All Randomised Controlled Trials Produce Biased Results." *Annals of Medicine* 50 (4): 312–322. doi:10.1080/07853890.2018.1453233.

Kvangraven, I. H. 2020. "Impoverished Economics? A Critical Assessment of the New Gold Standard." *World Development* 127: 104813. doi:10.1016/j.worlddev.2019.104813.

Labrousse, A. 2020. "The Rhetorical Superiority of Poor Economics." In *Randomized Control Trials in the Field of Development. A Critical Perspective*, edited by F. Bédécarrats, I. Guérin, and F. Roubaud, 227–255. Oxford: Oxford University Press.

Latour, B., and S. Woolgar. 1979. *Laboratory Life: The Construction of Scientific Facts*. New York: Sage.

McCloskey, D. 1998. *The Rhetoric of Economics*, 1st ed. 1985 ed. Madison: The University of Wisconsin Press.

Morduch, J. 2020. "The Disruptive Power of RCTs." In *Randomized Control Trials in the Field of Development: A Critical Perspective*, edited by F. Bédécarrats, I. Guérin, and F. Roubaud. 108–125. Oxford: Oxford University Press.
Morvant-Roux, S., I. Guérin, M. Roesch, and J.-Y. Moisseron. 2014. "Adding Value to Randomization with Qualitative Analysis: The Case of Microcredit in Rural Morocco." *World Development* 56: 302–312. doi:10.1016/j.worlddev.2013.03.002.
Naudet, J.-D., J. Delarue, and T. Bernard. 2012. "Évaluations d'impact: Un outil de redevabilité? Les leçons tirées de l'expérience de l'AFD." *Revue d'économie du développement* 20 (4): 27–48. doi:10.3917/edd.264.0027.
Necker, S. 2014. "Scientific Misbehavior in Economics." *Research Policy* 43 (10): 1747–1759. doi:10.1016/j.respol.2014.05.002.
Ogden, T. N. 2017. *Experimental Conversations: Perspectives on Randomized Trials in Development Economics*. Cambridge, MA: MIT Press.
Pamiès-Sumner, S. 2015. *Development Impact Evaluation. State of Play and New Challenges (A Savoir)*. Paris: AFD.
Pritchett, L., and J. Sandefur. 2015. "Learning from Experiments When Context Matters." *American Economic Review* 105 (5): 471–475. doi:10.1257/aer.p20151016.
Quentin, A., and I. Guérin. 2013. "La randomisation à l'épreuve du terrain." *Revue Tiers Monde* 213 (1): 179–200. doi:10.3917/rtm.213.0179.
Rao, V., K. Ananthpur, and K. Malik. 2017. "The Anatomy of Failure: An Ethnography of a Randomized Trial to Deepen Democracy in Rural India." *World Development* 99: 481–497. doi:10.1016/j.worlddev.2017.05.037.
Ravallion, M. 2020. "Should the Randomistas (Continue to) Rule?" In *Randomized Control Trials in the Field of Development: A Critical Perspective*, edited by F. Bédécarrats, I. Guérin, and F. Roubaud, 47–78. Oxford: Oxford University Press.
Rodrik, D. 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In *What Works in Development?: Thinking Big and Thinking Small*, edited by J. Cohen and W. Easterly, 24–47. Washington, DC: The Brookings Institution.
Rozas, D. 2014. *Ending the Microfinance Crisis in Morocco: Acting Early, Acting Right*. Washington, DC: IFC.
Sukhtankar, S. 2017. "Replications in Development Economics." *American Economic Review* 107 (5): 32–36. doi:10.1257/aer.p20171120.
The Royal Swedish Academy of Sciences. 2019. "Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019." Understanding Development and Poverty Alleviation.
White, H., and E. Masset. 2018. "The Rise of Impact Evaluations and Challenges Which CEDIL Is to Address." *Journal of Development Effectiveness* 10 (4): 393–399. doi:10.1080/19439342.2018.1539387.
Wydick, B. 2016. "Microfinance on the Margin: Why Recent Impact Studies May Understate Average Treatment Effects." *Journal of Development Effectiveness* 8 (2): 257–265. doi:10.1080/19439342.2015.1121512.

Appendix

Description of the AAA-RCT protocol

A set of 162 villages were 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 randomly assigned to the treatment group and the other to the control group. In the 81 treatment villages, AAA started to provide microcredit, opening branches and offering microcredit services 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. This population was then divided into two groups: the 'high borrowing propensity' households, cor-

responding to the last quartile of a propensity score estimated from the preparatory survey, and the 'lower borrowing propensity' group. All households in the first group were included in the sample, along with five households in the second group. Two surveys were then conducted on the sampled households: the pre-AAA baseline survey, before the intervention, conducted in four successive waves between 2006 and 2007, and the endline survey conducted two years later, after the intervention. To this sample of 4465 households, an additional sample of 1433 households with a 'very high borrowing propensity' 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 baseline and endline. They could only be fully interviewed at the endline. The three samples (initial 'high borrowing propensity' and 'lower borrowing propensity', plus the subsequent 'very high borrowing propensity') 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 effects. This particularity is presented as an essential and unprecedented contribution to this experiment.

List of unpublished documents

Presentations at 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', 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 and emails

(DI_1): J-PAL, Evaluation de l'impact d'un programme de micro-crédit en milieu rural: Al Amana au Maroc, research project submitted to the AFD Research Department, 21 January 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 April 2008, 7 pages.
(DI_4): Réunion Copil 2009, Comité de pilotage de l'Evaluation d'impact d'un programme de micro-crédit en milieu rural au Maroc, minutes of the meeting of 29 January 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 July 2008.