DOCUMENT DE TRAVAIL DT/2019-13

Microcredit RCTs in Development: Miracle or Mirage?

Florent BEDECARRATS

Isabelle GUERIN

François ROUBAUD

UMR LEDa
Place du Maréchal de Lattre de Tassigny 75775 • Paris •Tél. (33) 01 44 05 45 42 • Fax (33) 01 44 05 45 45
DIAL • 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 : dial.ird.fr

Microcredit RCTs in Development: Miracle or Mirage? [1 2]

Florent Bédécarrats[3], Isabelle Guérin[4], François Roubaud[5]

Abstract

Microcredit has long stood as a flagship topic for RCTs in development, starting with the publication of a special issue in a leading economics journal on six RCTs conducted in different world regions. This special issue was hailed as the first rigorous and conceivably definitive study on the impacts of microcredit. However, a detailed exploration of the implementation of these six RCTs reveals many limitations with respect to internal and external validity, ethics and interpretation. This paper uses analytical tools from statistics, political economy and development anthropology to discuss the extent to which the entire RCT chain strays from the ideal RCT principles (from sampling, data collection, data entry and recoding, estimates and interpretation to publication and dissemination of results). It also raises questions about the disparity between the academic and political success of this special issue and the many inconsistencies of method.

Keywords: Randomized Control Trial (RCT), Microcredit, Developing countries, Internal validity, External validity, Statistics, Development anthropology, Political economy, Ethics.

JEL codes: A11, A14, B41, C18, C93, N27, O16

Résumé

Le microcrédit est depuis longtemps un sujet phare des RCT dans le champ du développement. La publication d'un numéro spécial d'une revue économique de premier plan portant sur six RCT menés dans différentes régions du monde a constitué un point d'orgue dans ce domaine. Ce numéro spécial a été salué comme la première étude rigoureuse, et même pour certains le dernier mot concernant l'impact du microcrédit. Cependant, une analyse détaillée de ces six RCT et leur mise en œuvre révèle de nombreuses lacunes tant du point de vue de la validité interne qu'externe, de l'éthique et des interprétations qui en ont été tirées. Cet article mobilise les outils analytiques de la statistique, de l'économie politique et de l'anthropologie du développement pour discuter l'ensemble de la chaîne de production des différentes RCT (de l'échantillonnage, la collecte de données, la saisie et le recodage des données, les estimations et l'interprétation à la publication et la diffusion des résultats), notamment au regard de leurs principes théoriques. L'article propose des éléments d'interprétation pour expliquer le hiatus entre le succès académique de ce numéro spécial et de ses retombées en termes de politiques en dépit de ses défaillances méthodologiques de premier ordre.

Mots-clefs : Randomized Control Trial (RCT), Microcrédit, Pays en développement, Validité interne, Validité externe, Statistique, Anthropologie du développement, Economie politique, Ethique.

---

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

[2] We thank the participants of the March 2019 workshop which brought together most of the contributors to the book, as well as Solène Morvant-Roux, Jonathan Morduch and Martin Ravallion for their comments on an earlier version of the paper.

[3] AFD-EVA (Evaluation Division of the French Development Agency)

[4] IRD-CESSMA (Centre for Social Science Studies on the African, American and Asian worlds at the French National Research Institute for Sustainable Development)

[5] IRD-DIAL (Joint Research Unit on Development, Institutions and Globalization at the French National Research Institute for Sustainable Development)

I. Introduction

Since now two decades, a new field of research in development economics has emerged: the randomized control trials ((hereinafter referred to as RCTs). Initiated at the beginning of the 2000s, this approach has met with dazzling success, to the point of becoming a worldwide wave, not only in research but more generally in the international community dealing with development issues. On the first front, RCTs are today widely regarded as the benchmark, the *Gold standard* in the field of impact evaluation. They exercise a double domination. Theoretical first, since they claim to be the only ones able to solve "rigorously", but also simply, the puzzle of causal inference, which has become over the years the central subject and the unsurpassable horizon of econometricians (Cartwright 2007; Jamison 2017), the famous "*experimental ideal*" (Angrist et Pischke 2009); secondly empirical, because strong of its theoretical credibility, randomized experiments have become the only credible instrument allowing to prove "scientifically" the real effect of the policies and programs implemented on the ground, and consequently to select those which work and therefore susceptible to be replicated all over the world. Strong of these claims, and surfing on a promising context (*evidence based policy*), RCTs have attracted a growing number of international donors (public and private), struggling to prove the effectiveness of development aid. Not only have RCTs imposed themselves in the South, but by boomerang effect they have also given a new impetus to impact evaluations in the North, in particular in countries where the culture of scientific evaluation of public policies was limited, as for example in France, where the term "social experimentation" has become in a few years a key word.

Driven by a particularly active and powerful pro-RCT movement, hundreds of RCTs have been conducted, mobilizing millions of dollars and leading to hundreds of publications in prestigious scientific journals (Bédécarrats Guérin and Roubaud 2019; Ravallion 2019). This success has been followed by a number of criticisms regarding the universal scope of the method and its applications (see for the most vocal ones: Heckman 1991; Rodrik 2008; Ravallion 2009 and 2019 ; Barrett and Carter, 2010 ; Deaton, 2010 ; Harrison, 2011 ; Deaton and Cartwright, 2018) including our own contributions (Bédécarrats Guérin and Roubaud 2013 et 2019). The main interrogations relate to the internal and above all external validity of the methodology, with the consequence of its ability to draw general lessons in development policy (*what works and what doesn't work*), a central claim of the movement (Ogden 2017); more marginally, some papers have called into question the innovative aspect of the method from a historical perspective (Oakley 2000 ; Jamison 2017 ; Labrousse 2017), while others have focused on the political economy of this success (Jatteau 2016 ; Bédécarrats Guérin and Roubaud 2019). However, and despite the international renown of some of these critical voices, the pro-RCT movement continued without blinking its victorious road to impose this new standard, to the point that it is not illegitimate to claim that it seems to have won, at least temporarily, the battle of ideas in this area. The attribution in 2019 of the Nobel Prize (in economics) awarded to Esther Duflo, Abijit Barnerjee and Michael Kremer, leading members of the RCT movement, constitute the acme of RCTs in development consecration.

Along with the spread of RCTs, the rise of microcredit marked a second major milestone in development policies for poverty reduction in recent decades (Cling, Razafindrakoto and Roubaud 2003). It took off in the 1990s and reached its zenith in the early 2000s with the launch of the UN International Year of Microcredit (2005) and the award of the Nobel Peace Prize to the Grameen Bank and to its founder, Mohammad Yunus (2006). These two developments are actually closely interlinked: microcredit was one of the flagship topics, an emblematic subject, to be evaluated by random experiments in development.

This paper presents a detailed examination of RCTs on microcredit in development drawing on a wide range of analytical tools used in statistics, political economy, sociology and development anthropology. Its main focus is the special issue (hereafter, the Special Issue) published in 2015 in a major economics journal – the *American Economic Journal: Applied Economics* (AEJ:AE). This

Special Issue brings together six RCTs on microcredit, and the papers are prefaced by a general introduction (hereafter, the General Introduction) drawing broad conclusions. The Special Issue has had a great impact in both academic and professional circles, and tends to be seen as the definitive conclusion on the (limited) impacts of microcredit. But is it really?

We discuss this Special Issue from two angles: 1) top-down with a test on a specific case (microcredit) of the general criticisms made of RCTs, especially those developed by the authors in a previous article (Bédécarrats, Guérin and Roubaud 2019); and 2) bottom-up with a study of the implementation of RCTs on the ground. We take as a starting point our replication of one of the six RCTs discussed in the Special Issue: the RCT conducted in rural Morocco (Bédécarrats, Guérin, Morvant-Roux, et al. 2019a, 2019b), which plays a central role in the Special Issue's 'economy'. We then expand the focus from the Moroccan case to take a more general angle by identifying the invariants that hold in other RCTs and ascertaining each RCT's particularities. More broadly, the main question we ask in this paper is, "*What lessons can be learned from RCTs on microcredit and how can their worldwide success be explained when they are not robust*?"

The rest of this paper is organized as follows. After summarizing the main features of the six experiments, the second part presents their main results and situates the Special Issue in the general context of the weight and role of microcredit in the RCT industry. The third part takes a comparative view to identify the main technical criticisms that can be made of this corpus of experiments, in terms of both their internal and external validity, as well as the ethical concerns raised. Moving beyond the method and the quantitative results, the fourth part analyses the interpretations proposed by the authors (particularly in the General Introduction), and their underlying theory of change. In conclusion, we propose an interpretation of the hiatus outlined above – a far-reaching success despite major shortcomings – and we draw more general lessons from our work.

II. The RCT on microcredit: a sinking flagship product?

Microcredit is one of the main services provided by microfinance, one of the sectors the most frequently evaluated by RCTs. An illustration of this importance can be found on the RCT online repository managed by J-PAL (a global research centre promoting this method for poverty reduction and the leading provider and promoter of RCTs)**.** In 2010, this repository displayed 233 RCTs, of which 32% were labelled as "microfinance" (Bédécarrats, 2012). JPAL since then reorganized its evaluation labelling with broader categories, and it currently posts 287 "finance" RCTs out of its 978 RCTs.[6] Finance is J-PAL's foremost sector of interest, ahead of Education (233), and Political Economy and Governance (216). Although microfinance is just a subset of 'Finance' RCTs, J-PAL is a major provider of impact evaluations on the subject. The mid-2000s saw a boom in the number of RCTs on microfinance and the RCT industry as a whole (Bédécarrats, Guérin, and Roubaud 2019; and Ravallion, 2019). Since then, the number of microfinance RCTs has dropped sharply while RCTs in general have continued to grow (Figure 1). There is no easy way to count the number of RCTs conducted worldwide. Our estimates are based on 3IE's online impact evaluation repository rounded out by Bédécarrats (2012) and J-PAL's online evaluation repository.[7] Figure 1A illustrates that the impact of microfinance has

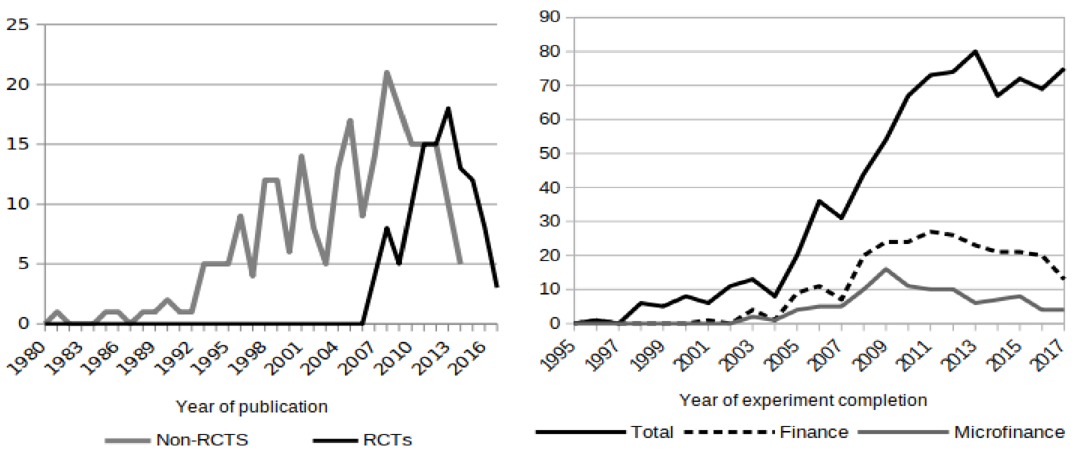
[6] Source: The Abdul Lateef Jameel Poverty Action Lab website: www.povertyactionlab.org/evaluations, visited on 13/10/2019.

[7] 3IE's online impact evaluation repository forms the main catalogue of results of impact evaluations on development interventions (https://www.3ieimpact.org/evidence-hub/impact-evaluation-repository, last accessed for the authors' update on 13 October 2019). 3IE tends to underreport non-experimental evaluations and its inventory work appears to have dropped off in recent years, as references decrease from 2015 onwards. We have rounded out 3IE's data with the impact evaluations listed in Bédécarrats (2012) and references included in J-PAL's evaluation repository. References have been matched to avoid double counting the same evaluations. Figure 1B is based on the references listed in J-PAL's online evaluation repository (https://www.povertyactionlab.org/evaluations, accessed for the last update on 18

long been a disputed issue, generating numerous non-experimental impact evaluations. Despite the fact that experimental methods provide theoretically stronger quantitative empirical evidence, non-experimental studies furnish a wealth of relevant evidence. There has also been a sharp increase in experimental evaluations, coinciding with a sharp decrease in non-experimental evaluations, although these trends might be marginally exaggerated by omissions of the most recent studies in the registries we used. Figure 1B also shows that microfinance was a prominent theme for the *randomista* movement up to 2013, but that interest has since waned. The fall in the second half of the 2010s following the peak in the first half is intriguing: is it due to a trend shift or is it because there is not much left to say about this overstudied issue? This is one point we will address in the following.

Figure 1: RCTs on microfinance

1A: **1B:**



Source: Authors, based on: 3IE evaluation repository (2019), J-PAL evaluation repository and Bédécarrats (2012) for Panel 1A; and J-PAL online evaluation repository for Panel 1B.

It was at the height of RCTs in microfinance that a 2015 special issue was published in the *American Economic Journal: Applied Economics* (AEJ:AE) featuring six RCTs on microcredit (Banerjee, Karlan, et al. 2015). This special issue is seen by leading RCT movement figures as the decisive contribution to settle a long-standing debate on the subject (Ogden 2017), both in academia and among donors and policymakers. It quickly attracted massive coverage, as seen from the 3,607 citations of its articles in other scientific publications.[8] In a move to promote its use to inform policy-making, J-PAL and IPA published a policy briefcase that took stock of the special issue and drew general conclusions for microcredit worldwide (J-PAL & IPA Policy Bulletin 2015). Some researchers even mused that it might be the "last word on microcredit" (Sandefur 2015).

Looking more carefully at the academic impact of the AEJ:AE Special Issue, the result is impressive. *Google Scholar* (accessed 13/10/2019) lists the General Introduction alone as having been cited 527 times. A great performance, although way behind the paper by Banerjee et al. (2015a) on the *Spandana* microcredit programme in India (1,813 citations). The other five papers have also performed very well: 320 citations for Angelucci et al. (2015) on *Compartamos Banco* in Mexico, 298 for Crépon et al. (2015) on Al Amana in rural Morocco, 225 for Attanasio et

October 2019). 'Finance' in the key is the label assigned by J-PAL to the registered evaluation. The authors assigned the 'Microfinance' label after reviewing the summaries of all the evaluations registered as 'Finance' on J-PAL's website. The dates in Figure 1B correspond to the year in which the experiment was completed, while the dates in Figure 1A stand for the year in which the experiment results were published.

[8] Source: Google Scholar citation indexes on the articles featured in this special issue, see corresponding webpage, visited on 13/10/2019.

al. (2015) on Mongolia, 214 for Augsburg et al. (2015) on Bosnia, and 210 for Tarozzi et al. (2015) on Ethiopia. By way of comparison, the count for Pitt and Khandker (1998), quoted by Roodman and Morduch (2014) as the all-time most cited empirical article on an individual microcredit project, stands at 1,956 citations more than twenty years after its publication.

In addition to direct citations, the Special Issue's impact is cascaded through quotations of citations (like any article), but also through systematic reviews or meta-analyses, which build mostly on the Special Issue as their main body of evidence (Brody et al. 2015; Buera et al. 2015; Chernozhukov et al. 2018; Demirguc-Kunt et al. 2017; Meager 2019). Special mention can be made of the article published in the prestigious *Science* review in 2015 (Banerjee, Duflo, Goldberg, et al. (2015), cited 484 times).[9] This article extensively discusses the Special Issue, highlighting the comparative merits of a different approach ("graduation" programmes).

Lastly, the results of the Special Issue have circulated widely beyond academic circles to the world of microfinance practitioners (J-PAL & IPA Policy Bulletin 2015). CGAP, which plays a leading role in disseminating good practices in the microcredit sector, commented on it even before its release (Cull et al. 2014). For many practitioners (whom one of us meets regularly in conferences and in the field), the results of the Special Issue are now conventional wisdom.

Ultimately, whether judged on the basis of the number of RCTs conducted or the dissemination of results, microfinance, and microcredit impact evaluations in particular, do appear to be the flagship products of the franchise created by the *randomistas*[10] based on the RCT method, and the Special Issue the outstanding prototype for this movement.

A focus on the design of the AEJ:AE Special Issue

The Special Issue features six articles on six microcredit RCTs conducted by six affiliated J-PAL teams in six different countries (Bosnia & Herzegovina, Ethiopia, India, Mexico, Mongolia and Morocco) at around about the same time (from 2006 to 2012). It is preceded by a General Introduction that draws general lessons from this collective experience. The Special Issue draws its strength from a downstream harmonization process organized by the journal in preparation for its publication.[11] A common analysis plan was drawn up to facilitate comparisons. As far as possible, the impact of microcredit was estimated using the same econometric methodology for a set of common outcomes, themselves calculated the same way. This was the first time that such a pooling effort had been made on this scale. It represents a decisive advantage when it comes to generalization.

Not only does the Special Issue appear decisive in terms of results, but it also marks a 'good practices' shift by RCT proponents. Hence the issue seeks to address a number of limitations. For the first time, the issue as a whole, and the General Introduction in particular, provide elements of response to five types of recurrent criticisms of the pro-RCT movement (Bédécarrats, Guérin and Roubaud 2019): a theoretical model is developed in response to the agnostic empiricism criticism of RCTs; a cost-benefit analysis is proposed to answer the question of effectiveness, to move beyond mere causal impact; the issues of take-up rate, estimator accuracy and treatment heterogeneity are acknowledged and discussed; contextual diversity is addressed by a range of settings, products and institutions covered by the six papers, enabling the Special Issue's editors to claim their sample is "*fairly representative of the microcredit industry/movement worldwide*" (Banerjee, Karlan and Zinman 2015: 2); and, lastly, the Special Issue professes to make available

[9] This is not the first time that *Science* has opened its columns to RCTs on microcredit (Karlan & Zinman 2011).

[10] We call *randomistas* those RCT proponents who are convinced that RCTs are the only way to rigorously assess impact in evaluation, and that they are superior to other methodologies in all cases.

[11] "*Drawing lessons across the six studies has been greatly facilitated by the efforts of the six research teams and the editor, Esther Duflo, to make the papers readily comparable,*" (Banerjee, Karlan and Zinman 2015:2).

the original databases in response to the complaint about replicability and in order to facilitate meta-analyses.

Let us briefly describe the six RCTs. Despite an upstream harmonization process (data processing and analysis), the experiments differ significantly in their protocols. The types of microcredit products, microfinance institutions (MFIs hereafter), unity of randomization procedures, and so on vary from one RCT to another. The authors interpret this diversity based on the assumption that the similarity of results across this wide range of environments is a guarantee of their robustness, and therefore evidences the generic properties of microcredit impacts; a way of addressing the recurrent criticism of RCTs as lacking external validity.

Table 1: Main characteristics of the six RCTs

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Interest rate (APR)	22%	12%	24%	110%	27%	14%
Liability	Individual	Group	Group	Group	Both	Group
Average loan/household income	9%	118%	22%	6%	43%	21%
Sex of potential clients	Both	Both	Female	Female	Female	Both
Loan eligibility (among other)	Strong collateral, repayment capacity, creditworthiness …	Poverty status, business plan …	18-59 years old, proof of residence home ownership...	18-60 years old, valid ID card, proof of address ….	Assets<$869 Profit<$174/month	18-70 years old ID card, non-livestock agricultural activity …
Area coverage (urban/rural)	Both	Rural	Urban	Both	Rural	Rural
Area coverage (regions/cities)	14 (nationwide)	2 (Western)	1 (City)	4 (NC Sonora)	5 (North)	11 (nationwide)
Unit of randomization	individual	Association	Neighbourhood	Neighbourhood & village	Village	Village
Final Sampling Unit	Risky and unreliable applicant …	Random households	Household with >=1 woman >=3 years in the area…	Has a business or would like one …	Interested in obtaining a loan …	Household deemed likely borrowers …
Sample size (endline)	995	6,263	6,862	16,560	964	5,551

Source: authors, based on Banerjee, Karlan and Zinman (2015) (Tables 1 and 2).

The General Introduction gives a detailed presentation of the main features of the six RCTs, summarized in **Table 1**. The MFIs vary in size, with some being commercial while others are not. We find all kinds of products: joint liability and individual loans, weekly and monthly repayments, an annual interest rate varying from 12% to 110% (on average), and the (average) loan amount ranging from 6% to 118% of monthly income. Half of the microcredit programmes target women. In terms of geographic areas, one is exclusively urban (India), three are exclusively rural (Ethiopia, Mongolia and Morocco) and the remaining two cover both types of area. One point of note is that, in all cases, the client eligibility criteria are ad hoc: they depend on both the internal rules of each MFI and on the parameters of each RCT. As a result, the target populations are highly specific (if not unique), undermining the possibilities for inference and extrapolation to larger populations; we will come back to this point in the third part.

A focus on the AEJ:AE Special Issue: main results

The General Introduction draws seven major lessons from the exercise. In the first place, low take-up is a constant in all the studies except Bosnia, leading to the conclusion that microcredit cannot be the universal panacea for lifting the poor out of poverty. An unfortunate consequence of the low take-up is that it poses a problem of statistical power and a challenge for the RCT identification strategy. However, the General Introduction puts forward the Moroccan, Indian and Mexican RCTs to provide new elements to address these shortcomings (take-up prediction and sampling strategy). Second, and tying in with the previous point, it is particularly difficult to predict the take-up rate, and no study has managed entirely satisfactorily to do so. Third, and probably the main conclusion, access to microcredit is not transformative either for microenterprise performance or for household living conditions – including social well-being and women's empowerment – at least on average. The only robust finding for consumption is a decrease in 'discretionary spending', defined by the authors as "temptation goods, recreation/entertainment/celebrations" (Banerjee, Karlan and Zinman 2015: 13). Fourth, only firm investment is stimulated by microcredit, showing that it cultivates micro-entrepreneurs' intentions to develop their business. Fifth, other modest, albeit potentially important effects are pointed up: freedom of choice in particular. Sixth, although microcredit is not transformative, it does not have any catastrophic effects either, which places proponents and opponents of microcredit on a level pegging. Lastly, the seventh lesson relates to the presumption of heterogeneity of microcredit impact, which could be positive (even transformative) for some (the upper tier), and negative for others. This brings us back to the issue of statistical power, the sample sizes required to properly estimate impacts and the representativeness of the targeted populations. **Table 2**, based on the General Introduction and the J-PAL and IPA Policy Bulletin (2015), summarizes the results obtained by the six RTCs for the main outcomes monitored.

Table 2: Main results of the six RCTs

	Bosnia and Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Business ownership	Positive	n.s.	n.s.	n.s.	Positive	n.s.
Business revenue	n.s.	n.s.	n.s.	Positive	n.s.	Positive
Business assets	Positive	-	Positive	-	Positive	Positive
Business investment	n.s.	n.s.	Positive	Positive	-	Positive
Business profits	-	-	-	-	-	Positive
Household income	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Household consumption	n.s.	Negative	-	Negative	Positive	-.
Household consumption of temptation goods	Negative	-	Negative	Negative	n.s.	Negative
Social well-being	n.s.	n.s.	n.s.	Positive	-	n.s.
Women's empowerment	-	n.s.	-	Positive	-	-

Source: authors, based on J-PAL and IPA Policy Bulletin (2015); Banerjee, Karlan and Zinman (2015)
Note: n.s. (not significant at 10%); - (no data).

In conclusion, the Special Issue is considered by many, starting with the authors themselves (Ogden 2017), as the most comprehensive summation on the impact of microcredit. Its general conclusions have scarcely been questioned since its publication in 2015 (Dahal and Fiala 2018; Wydick 2016). In a way, it freezes the state of the art on the causal impacts of microcredit and its role for development and poverty eradication. For AEJ:AE's editors, and subsequent papers elaborating on the six RCTs, the Special Issue does even more than this. It is praised for pushing back the frontiers of scientific knowledge, both on microcredit and on the RCT method. Three papers, posterior to the Special Issue and directly following up on the same set of RCTs, are good illustrations of this. Meager's (2019) article, published again in AEJ:AE, confirms that it is still considered the must on microcredit. This article takes the six RCTs in the Special Issue (plus an RCT in Philippines; Karlan andZinman 2011) to re-estimate the general impact on the main variables and answer the question of external validity using an innovative method (a Bayesian Hierarchical Analysis). Then there is Chernozhukov et al. (2018), who apply a double machine learning method to study heterogeneity in this data set. A third example is Banerjee et al. (2019), published as this paper was being written. The paper draws on a third-round survey for the *Spandana* Indian RCT. While responding to some of the criticisms of RCTs (by addressing heterogeneous treatment, lengthening the time span and developing a theoretical model), the paper largely refers to and takes stock of the Special Issue, presented as the seat of knowledge on microcredit to date. This paper may not be the last in the series. In the same vein, Crépon et al. (2015) also announce in the conclusion to their paper a third-round survey for the Moroccan RCT to assess the long-term impact of microcredit.[12]

III. Validity and scope of the Special Issue: a critical assessment

In the literature, RCTs are appraised from two main angles: external and internal validity. External validity is pivotal when it comes to scaling up, informing and designing public policies on a broader scale (national or regional) and to testing a theory. Internal validity is usually taken for granted with RCTs, and seen as their major strongpoint over other methods. While this property may be true in theory, implementation constraints in the field can call these ideal conditions into question, a point hitherto overlooked.

Internal validity

Assessing the internal validity of RCTs calls for a probe into the making, and tinkering, of RCTs in the field. We performed this demanding exercise on the Moroccan study (Crépon et al. 2015). We present below the main results of the two companion papers we produced from this review (Bédécarrats, Guérin, Morvant-Roux, et al. 2019a, 2019b)

The emblematic case of the Moroccan RCT

From 2006 to 2010, a research team from J-PAL conducted an RCT in rural Morocco to measure the impact of microcredit provided by Al Amana, then the Moroccan market's leading MFI, in the midst of a phase of expansion.

We replicated Crépon et al.'s paper and identified a number of issues that challenge their conclusions (Bédécarrats, Guérin, Morvant-Roux, et al. 2019a). We argue that they used inconsistent trimming procedures and thresholds, and that their results depend heavily on how their data was trimmed. 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.

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, e.g. on the household head, gender and spoken language. 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. Inclusion of tractors and reapers in asset appraisal increases the sample's average value of agricultural assets per household by 470% (from 1,377 Moroccan Dirham to 5,111 Moroccan Dirham). The identified coding errors altered some 80% of the observations.

Inconsistencies in credit measures warrant particular attention, as they are essential to characterize the treatment evaluated by this experiment. Crépon et al. (2015) append administrative data to the survey data, reporting the former's given microcredit take-up of 17% rather than the latter's 11%. They contend that the Moroccan population underreport borrowing because of religious shame. However, we argue that this is implausible as the inconsistencies between sources go way beyond differences in averages. A total of 195 of the

[12] "*We are currently following up with the households, now that a much longer time period has elapsed, to check if the investment in business assets paid off in the longer run,*" (Crépon et al. 2015: 148).

435 reported clients said they had never borrowed from the MFI. However, a 'credit shame' explanation for these households would imply a 'credit pride' explanation for the 152 households that reported having a loan from the MFI even though they did not appear on its registers. According to the survey data collected on the panel sample, access to credit remained stable in the treatment group between baseline and endline, while it was decreasing in the control group (there was a major crisis in Moroccan microfinance from 2008 to 2010). Our results challenge the very meaning of this RCT: what was tested appears to have been 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.

We also found sampling errors. For example, the sex and age composition for 20% of the households interviewed at baseline and reportedly re-interviewed at endline differs to such an extent that it is implausible that the same units were re-interviewed in these cases. In addition, we found that Crépon et al.'s sample characteristics differed in substantial ways from the population's characteristics. The number of household members grew from 5.17 to 6.13 between the baseline and endline surveys. The national census, however, reported that Moroccan rural households had an average of 6.03 members in 2004 and 5.35 members in 2014. Such discrepancies raise questions about the sample's representativeness, and hence undermine the external validity of this study.

The authors produced a reply to our replication, entitled *Rejoinder*, rejecting most of the errors we documented (Crépon et al. 2019). They referred to our analysis, but they do not appear to have replicated or closely analysed its statistical content and we argue that their rejoinder thereby contains numerous factual errors and omissions. We published a review of their main arguments in response to our replication (Bédécarrats, Guérin, Morvant-Roux, et al. 2019c). We found that all the coding, measurement and sampling errors documented in our replication still hold.

Distortion of the protocol: product and sampling tweaking

Our second paper sought to explain how such inconsistencies could occur, using a qualitative field study specifically designed to round out the RCT (Morvant-Roux et al. 2014) and various data and documents, public and internal from the RCT's key stakeholders (Bédécarrats, Guérin, Morvant-Roux, et al. 2019b). The paper describes the entire study production chain, from sampling, data collection, data entry and recoding, estimates and interpretations to publication and dissemination of results. Far from ideal laboratory conditions[13], the analysis of the randomized protocol's implementation on the ground by the different players (each with their own motivations and constraints) finds a number of discrepancies compared with the theoretical protocol reported in the published article.

A major concern during the study was take-up, much lower than initially expected, which prompted a number of corrective measures. 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.[14] When these measures proved insufficient, the team tweaked the sampling method (modification of prediction models, and addition of new

[13] Field experiments such as RCTs are designed precisely to get out of the artificial world of laboratories. But too often randomists think that the protocol can be applied as it is, as in the laboratory, which is not the case.
[14] Changing the product for the sake of the RCT is also an external validity issue (as experimental conditions are not in line with how it functions in the "real world" (Peters et al. 2019).

households at endline, with a supposedly higher propensity to borrow). Villages with zero take-up were dropped.

Poor data quality and measurement errors

Data collection and entry were subcontracted to a consultancy firm specialized in engineering, but with no experience of statistical surveys. For the purpose of monitoring the RCT's design and implementation, the RCT's funder (AFD) appointed a team of economists and specialists in household surveys. The team reporting back on its field missions found serious data collection dysfunctions at an early stage. These included translation problems because interviewers did not speak Berber, a language spoken by a large part of the target population. Interviewers therefore made extensive use of impromptu translators, including local leaders, raising comprehension and response bias problems (social desirability and mistrust of government).

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 understand each other and the interviewers. These observations probably explain in part the abovementioned significant discrepancies between baseline and endline. 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 overlooked. Lacking time and supervision, some interviewers may simply have interviewed households available at the time of their visit. AFD's team made recommendations to improve the quality of the data collected, expressing concerns about the potential repercussions of these shortcomings on the experiment's results. They also raised the data entry issues. Although the J-PAL team responded, challenging the gravity of the problems and contending that they did not call into question the internal validity of the experiment, the next steering committee meeting decided that all questionnaires already entered were to be sent to the French National Statistical Office (INSEE) in Paris to be re-entered.

These different issues were omitted from the published article and point to shortcomings in the preparation, implementation and follow-up of field work.

Beyond the Moroccan RCT: a general assessment

It is not feasible to analyse the other five RCTs in the Special Issue in such detail, both for reasons of time and because the necessary raw data are available only for two of them (table 4). We therefore perform a partial exercise, namely a critical reading on the usual review summary terms, i.e. based on the published articles. Table 3 summarizes the internal validity problems as they can be assessed from the information available to us. Hardly any of these problems are addressed by the Special Issue, and even less so by the General Introduction. We discuss here the sampling error and measurement error issues in turn.

With regard to sampling, note that the papers generally do not provide the basic elements to be able to accurately describe and qualify the adopted sampling designs and selection plans (the standards for such descriptions are provided, for instance, in StatCan 2010 and Ardilly 2006). The authors focus their analyses on randomization and causal inference issues. First, the reference population is never clearly established. In most cases, it corresponds to eligible clients in the MFI's expansion areas, although it is not known how the latter are defined. This has unfortunate repercussions on the external validity of the RCTs (see below). Second, the adopted sampling plans fall into the general category of multi-stage stratified random sampling, with the exception of the RCTs in Bosnia and Herzegovina and Mongolia. Neither of these cases is randomly sampled: in Mongolia, the first 30 poor women in each selected village to state an interest in obtaining a loan were selected; in Bosnia, loan officers were asked to select potential clients who were not deemed eligible by the current MFI's standards. In all cases, these complex sampling designs, to use statistical terminology, either do not enable the confidence intervals associated with the estimated impact to be computed (the abovementioned two cases) or would call for particularly complex variance estimation calculations, which are not performed (except for estimating cluster-robust standard errors). The direct consequence of this gap is that the confidence intervals are probably underestimated and the impacts deemed significant, already small in number, should not be statistically different from zero.

Moreover, four of the six RCTs deviated from the experimental method's canonical protocol: random selection of a treatment and control group, a pre-treatment baseline survey (BL) and then panel monitoring based on a post-treatment endline survey (EL). In the Ethiopian case, the baseline and endline surveys were not on panels, but cross-sections (i.e. different individuals were surveyed). This makes it impossible to identify potential imbalances at baseline for the population for which impact is estimated at endline. In the Mexican, Moroccan and Indian cases, the field surveys could not be conducted as initially planned and threats to the experiment's success led to the initial protocol being readjusted along the way. In India, the baseline did not constitute a base panel for either of the two subsequent endlines,[15] raising the same problems as in the abovementioned case of Ethiopia. In Mexico, the baseline was aborted due to the poor quality of the data collected: 73% of the baseline households were not revisited at endline and 89% of the endline sample had not been not surveyed at baseline, so the majority of the households surveyed at endline were added at this stage. A similar strategy was adopted in Morocco. Low take-up by households identified as potential borrowers meant that new households were selected at endline that represented 26% of the endline sample. If we also take into account the attrition rates (available only for the panel protocols) ranging from 8% (Morocco) to 37% (Mexico), it is clear that none of the RCTs was conducted in keeping with the standards (non-random sampling of targeted households in Bosnia and Mongolia, and non- or failed-panels for the other four due to data collection issues or low take-up).

However, it is fundamentally important to verify sample balance at baseline. The studies vary a great deal in terms of the variables tested. Some tested surprisingly few variables compared with the wide range of data collected (Mexico). Others tested many more, but all differ as to which variables were tested. In some cases, most of the variables include at least some of the outcomes for which impact was measured at endline. In the case of Morocco however, the balance tests were applied only to specific subsets of the outcome variables evaluated at endline (e.g. sales for crop farming households, or livestock breeding households, instead of overall sales reported at endline). In our replication, we found large, significant imbalances in these outcomes. Households in the treatment group made 22% less sales and profits from self-employment than households in the control group (significant at the 5% level). They also invested 61% more (significant at the 5% level). In addition, there are imbalances at baseline with respect to a number of important variables, such as the surface area of owned land, access to basic services and women's empowerment. In addition to the variables tested, the calculation basis is also important. For example, the Mexican study limited its balance tests to 1,823 households surveyed at both baseline and endline. If we extend the same tests to all the households surveyed at baseline (6,786), as is the case with India and Morocco, we find significant differences in household income per adult in the previous month, especially those in an informal group.[16]

Even if baseline differences between treatment and control groups are not statistically significant, they can be very large. In Mongolia and Ethiopia, baseline balance tests found average differences often over 10% (and up to 50%), but not significant (not surprisingly given the small sample sizes). They are systematically interpreted for what appears to be convenience's sake (absence of imbalances and therefore success of the randomization process), while the opposite explanation is often given for the results: where coefficients are non-significant due to underpower, they are construed as being "economically meaningful".

[15] Banerjee et al. (2015a) conducted a first endline survey in 2007 and 2008. They re-interviewed the households of the first endline in a second endline survey in 2009 and 2010.

[16] Computations are available from the authors on request.

None of the papers discusses measurement error issues in depth. However, the literature emphasizes how difficult it is to obtain reliable measurements of many of the outcomes analysed, especially household consumption and microenterprise and agricultural production (Deaton 2000). Measurement errors merely get a mention in a footnote on potential memory bias in the Indian case, and in a discussion on under-reporting of borrowing in the Moroccan case, said to explain the differences between the administrative data and the surveys on this subject. Only the Ethiopian RCT reports major data quality concerns, and explicitly acknowledges that this issue affects internal validity. The Mexican RCT specifies that the baseline had to be interrupted and that its data could not be used because they were unreliable, without providing any details or indicating how more reliable data could have been collected at endline. Unfortunately, it is impossible to discuss data quality further from the articles alone. However, a detailed analysis of data consistency and the recoding conducted by researchers in the Moroccan case (Bédécarrats, Guérin, Morvant-Roux, et al. 2019a) shows that this problem altered the results. There is evidence to suggest that similar problems may exist in other cases. For instance, a preliminary analysis of the Mexican data finds that the age ranges do not match between surveys for 231 (12.7%) of the 1,823 women interviewed in principle at both baseline and endline.[17]

[17] Ibid.

Table 3: Internal validity of the six RCTs

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Population of interest	**Potential clients** initially rejected by the MFI as uncreditworthy	**Rural households in two ad-hoc areas**	**Likely borrowers** (women living in slums for more than three years with valid ID) in MFI expansion areas in Hyderabad	**Potential clients** (women owning or planning to create a business or intending to borrow) in MFI expansion areas in Central Sonora, Mexico	**Poor women**: (assets <$869 & profit<$174/month) **Signed up to get a loan**	**High borrowing propensity households in rural MFI extension areas**
Sample design, Randomization						
Sample design	Purposive individual sample	- Stratified (2 "zones") - 3 degrees (Admin units/Village/HH)	2 degrees (slums/HH)	2 degrees (village/HH)	- Stratified (5 provinces), - 2 degrees (village/HH) Northern Mongolia	2 degrees (village/HH).
Info on area selection #of areas (T, C)	Not applicable	Yes 353 Villages	Yes 104 (52, 52)	Yes 238 (120, 118)	No 25 (15, 10)	Yes 162 (81,81)
Discarded areas	Not applicable	No	Yes (16 slums)	Yes (12 areas)	No	Yes (not specified)
Info on selection of individuals	Yes, not random	Yes, random	Yes, random	Yes, random	Yes, not random (1st 30 to sign up)	Yes, random
Info on Randomization (T vs C)	Yes (individual level)	Yes (Village level)	Yes (Slum level)	Yes (Area level)	Yes (Village level)	Yes (village, level)
Sample size (full; control)	BL (1,196; 568) EL (994; 444)	BL (6,412; n.a.) EL (6,263; n.a.)	BL (2,800; 1,220) EL1 (6,863; 3,264) EL2 (6,142; 2,943)	BL (6,786; n.a.) EL (16,560; 8,298)	BL (710, 299) EL (610, 260)	BL (4,465; 2,266) EL (5,551; 2,810)
Attrition rate (BL->EL): %total, %control	Panel (17%; 22%)	No panel 2 cross-sections	BL->EL: no panel EL1->EL2 (11%; 10%)	Panel (37%; n.a.)	Panel (16%; 15%)	Panel (8%; 7%)
Respect of Experimental Design	Yes	No 22% areas misallocated (12% T not treated), 23% C treated)	No (16 areas dropped; BL unreliable)	No (BL aborted)	Yes	No (new HHs added at EL)
Balance tests at baseline						
Population included	Panel households only	Panel households only	All BL households	Panel households only	Panel households only	All BL households
Tested variables Include main study outcomes	27 Yes	35 Yes	33 Yes	14 No	48 Yes	43 No
Reported significant imbalances	No	No	No	Yes	Yes	Yes
Trimming	Results with and without 1% trimming for robustness checks	Results with and without trimming 8 obs for robustness checks	No	No	No	BL: trim highest values for 10.3% of obs. EL: trim 0.5% of obs.
Data quality (discussion in paper)	No	Yes, marginal (measurement errors)	Yes, marginal (possible recall errors)	Yes, marginal (missing outcome variables at EL)	No	No (except take-up admin vs survey)

Source: Authors based on AEJ:AE (2015). Notes: HH: households, BL: baseline survey, EL: endline survey, T vs C: treatment group versus control group, obs.: observations.

External validity

The question of RCTs' external validity is the most discussed in the literature. External validity is a key issue, especially since, in contrast to a lot of observational data, RCTs are conducted on a small scale and in non-representative locations as seen above. External validity is also at risk when sampling is selective, that is when a study focuses on specific sites and population categories. Then there is the implementers' bias: for instance, where the results obtained by an NGO do not replicate when the same intervention is delivered on a larger scale by a government (Bold et al. 2013, Vivalt 2017). However, the issue of the external validity of RCTs is rarely given serious consideration by the *randomistas*. Peters et al. (2018) conduct a systematic review of all (54) RCTs published in leading economic journals from 2009 to 2014 to assess the main threats to external validity (Hawthorn/Henry effects,[18] general equilibrium effects, specific sample problems and special care in treatment provision). Based on a set of objective indicators, albeit with lenient criteria, the paper finds that the majority of published RCTs do not discuss these hazards and many do not provide the necessary information to assess potential problems.

External validity also has to do with the relevance of the selected results. The focus on an 'average' impact and problems capturing the heterogeneity of impacts and their distribution form a major obstacle to the relevance of results (Ravallion 2009; Stern et al. 2012; Vivalt 2017). The restriction to a short-term impact (for reasons of cost and attrition) often means that mid-point indicators are studied, which can be very different from final outcomes (Boone et al. 2013), if not vice versa, since many project trajectories are not linear (Labrousse 2010; Woolcock 2013). Knock-on and general equilibrium effects are overlooked, albeit partially in the Moroccan RCT, despite there being any number of them (Acemoglu 2010; Deaton and Cartwright 2018; Ravallion 2009). The same holds true for the political consideration involved in programme replication, despite its being a key consideration for scale-up (Acemoglu 2010; Bold et al. 2013; Pritchett and Sandefur 2013). Last but not least, the *reasons* for the impact are disregarded: RCTs might be able to measure and test some intervention impacts and aspects, but they cannot analyse either their *mechanisms* or their underlying *processes*. Notwithstanding the method's limitations, the absence of theory prevents any form of understanding of the processes of change. Overcoming this limitation of the probabilistic theory of causality would call for a 'causal model' (Cartwright 2010), a coherent theory of change (Woolcock 2013), a structural approach (Acemoglu 2010) and evaluation of the intervention in context (Pritchett and Sandefur 2015; Ravallion 2009).

Table 4 summarizes the problems of external validity as they can be assessed from the information available to us. The usual RCT shortcomings hold here.

First, the sampling is selective: the experiment's selection criteria are ad hoc since they were conducted in MFI extension zones. As Wydick (2016) shows, the constraint of randomization (identifying virgin areas or populations) forced *randomistas* to choose 'marginal' areas and populations previously neglected by MFIs and therefore highly specific in relation to the 'normal' market. The unsuccessful bids by the Moroccan, Mexican and Indian studies to identify likely borrowers demonstrate that it is hard to characterize the microcredit target population. This rules out the possibility of extrapolation to a wider population as a legitimate action. A fortiori, the samples surveyed are not representative of anything aside from themselves: the households surveyed in the case of Bosnia and Mongolia, and the expansion areas (selected villages and neighbourhoods) in the case of the other four. Moreover, this property only exists in theory: the multiple failings of the survey protocols in the field mean that the expansion zones' theoretically representative samples are not representative in practice.

[18] These are behavioural biases induced by the experiment when subjects know they are taking part: biases on the treatment group (Hawthorne effect) or on the control group (John Henry effect). In the medical field, single or double blind (subjects and experimenters) RCTs are the usual way to control these biases (see Abramowicz and Szafarz, 2019).

If data cannot be extrapolated, comparison with other sources can be instructive to qualify respondent profiles. Official figures from representative surveys conducted by national statistical offices are a good benchmark to characterize a national or local context. Only two studies did so (Bosnia and Mongolia). In the other four studies, it is hard to get any idea of who was surveyed. As reported above, we performed this exercise for the Moroccan RCT. We have shown, among other results, that the average household size is atypical and tends to increase, while it decreases across the rest of the population over the same period. To take this assessment further, we use the typology of hazards to external validity established by Peters et al. (2018): Hawthorne and John Henry effects and general equilibrium effects (the others being addressed above). The papers do not discuss these hazards and many do not provide the necessary information to assess potential problems, except (partially) for Hawthorne effects (Bosnia and indirectly Mexico, see the discussion on ethics below) and general equilibrium and spillover effects (Morocco), despite the fact that they are at work in all cases.

Are these external and internal validity threats acknowledged by the authors? More broadly, what types of caveats do the papers mention? We report on them in Table 4. With the exception of the Moroccan RCT, the authors discuss a number of caveats. Almost all mention the lack of external validity given the lack of statistical power due to insufficient sample sizes. Heterogeneity to treatment is also widely acknowledged. The fact that the other RCTs return similar (but equally underpowered) results is considered as a source of robustness (see, for instance, Banerjee, Duflo, Glennerster, et al. 2015: 25). In addition, more specific caveats are quoted such as non-compliance with randomization design (Ethiopia) and selective attrition (India and Ethiopia), and measurement errors (Ethiopia). These observations tend to confirm the persistence of the Peters et al. (2018) results concerning the limited attention paid to external validity, to which we should add the internal validity problems raised above.

Lastly, ethical considerations warrant discussion, as this issue is of specific concern to RCTs in general (see Ravallion, 2019; Abramowicz and Szafarz, 2019). These considerations are absent from all articles when they should be stressed. The papers do not specify whether the informed consent of the participants was requested and obtained, with the exception of Angelucci et al. (2015). In addition, the information that they report to have imparted to the participants is partial: they specify, possibly to rule out suspicion of a Hawthorne effect, that they asked for an agreement to participate in a "*comprehensive socioeconomic research survey*". Yet they knowingly failed to mention that the survey was connected with Compartamos and especially that it was part of an experiment. A look at the available survey questionnaires (Bosnia and Morocco) shows that, in these two cases, respondents were not informed that they were participating in an experiment. The Bosnian RCT raises further ethical issues. This RCT consisted of granting credit to individuals initially rejected by the MFI's risk criteria, as done in South Africa and the Philippines (Karlan and Zinman 2009, 2011). This strategy placed the treated group at risk, at odds with the "do no harm" principle. The RCT confirms that marginal customers have significantly more repayment difficulties than regular customers, with a risk of over-indebtedness.[19] Considering the multitude of examples, more broadly for RCTs in general, it would appear that the creation of Institutional Review Boards in many academic institutions has done nothing, or at least insufficiently, to remedy the observed ethical lacunas (Barrett and Carter, 2020). Without going so far as to call for a "moratorium on experimentation" in the South (Hoffmann, 2020), the issue should be at least addressed in priority.

[19] "*All this suggests that the loan officers had good reason to classify our target population as marginal,*" (Augsburg et al. 2015: 201).

Table 4: External validity, acknowledged caveats and ethical concerns

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Population of interest	MFI expansion	MFIs expansion	MFI expansion (partial)	MFI expansion	MFI expansion	MFI expansion (partial)
Extrapolation to any superpopulation?	No	No	No	No	No	No
Potential threats discussion (in paper)**?**						
Hawthorne or John Henry	Yes	No	No	No	No	No
General equilibrium	No	No	No	No	No	Yes
Comparison with NSO data?	Yes	No	No	No	Yes	No
Other surveys/methods implemented?	No	No	No	No	Village surveys, qualitative interviews	No
If yes used?	-	-	-	-	No	-
Explicit caveats acknowledged?	**Yes** 1- No External Validity 2- Underpower 3- Potential H&JH effects	**Yes** 1- No External Validity 2- Underpower 3- No panel= Imbalance at BL, selective attrition, heterogeneous effect 4- No respect of experimental design 5- No Consumption 6- Measurement errors	**Yes** 1- Underpower 2- Non-Representative BL 3- Selective Attrition and migration 4- Contamination 5- ITT representative of "likely borrowers" only	**Yes** 1- No External Validity 2- Data quality 3- No BL 4- Heterogeneous treatment periods	**Yes** 1- No External Validity 2- Underpower 3- Presence of other MFIs 4- Attrition (possible imbalance) 5- Not robust at MHT	**Yes** 1- Small significant imbalances at baseline
Ethical concerns discussion						
Informed consent for experiment	No	No	No	No	No	No
Risk analysis and monitoring	No	No	No	No	No	No
Equipoise	No	No	No	No	No	No
Reproducibility						
Data available	Raw data	No	Aggregated data	Aggregated data	Raw data	Raw data
Detailed code available	Yes	No	Partially	Partially	Yes	Yes
Survey questionnaire available on AEJ	Yes	No	No	No	No	Yes

Source: Authors based on AEJ:AE (2015). Note: MFI: microfinance institution; NSO: National Statistical Office. ITT: Intention to treat; BL: baseline survey; EL: endline survey; H&JH: Hawthorne and John Henry, MHT: Multiple Hypothesis Test.

Now that we have discussed the issues of internal and external validity, we turn to the question of the impacts themselves. Even without considering the limitations outlined above, and sticking to the results proposed by the authors, the impacts are problematic. Table 5 provides an overview. First, take-up data are unreliable and often contradictory between survey and administrative sources. The Moroccan case shows that the inconsistencies go beyond differences in averages and under-reporting (see above on the discrepancies between administrative and survey data). On average, the experiments' impacts on credit take-up range from 8% to 50% when clusters were randomized to 98.5% in Bosnia where individuals were randomized.

Regarding the impacts on microcredit, low take-up has huge implications in terms of the significance of the coefficient. Dahal and Fiala (2018) replicate the six AEJ:AE RCTs. They find that each one is significantly underpowered due to the low take-up of the financial product offered. Even after pooling the data, the minimum detectible effect magnitudes are still very large: 230% for main outcomes under perfect compliance and 1,000% under actual compliance. They conclude that, "*The existing research on the impact of microfinance is generally underpowered to identify impacts reliably and suggests that we still know very little about the impact of microfinance.*" Although Banerjee, Duflo, Glennerster et al. (2015) acknowledge the problem of underpower in their introduction, Dahal and Fiala (2018) is the first paper to quantify how big of an issue it is. It confirms the previous study by McKenzie (2012), which estimates the necessary sample size at 15,000,000 to be able to secure the power to identify impact magnitudes of 10% in the Indian RCT.

Turning to the impacts on the selected outcomes, the presentation on this issue made by the authors of the General Introduction (see Table 2), which is supposed to summarize the consolidated results of the six RCTs, is misleading. An exhaustive count of estimated impacts on all variables considered in the six papers draws the following conclusions. No less than 298 impacts are estimated throughout the volume (excluding quantile estimates). Of this total, only 10 are significant at the 1% level, meaning that 97% of the possible retained effects are not significantly different from zero. Three RCTs have no significant impact at all (Bosnia: 0/47, Ethiopia: 0/37, and Mongolia: 0/41) and one has only one significant impact (India: 1/99). Even when the threshold is relaxed to 10% (a more lenient threshold than in usual practice), 81% of the effects are not significant. The Bosnian RCT is an extreme case in this respect with only three significant impacts at this threshold out of the 47 tested. Such proportions raise all the more doubt as all the articles mention a systematic problem of statistical underpower, which would explain the lack of impact. The sample sizes are not large enough to estimate the impacts given the low take-up, and this is indeed what we find. Moreover, 60% of the significant impacts (at 1%) come from the Moroccan RCT, whereas it represents just 12% of the total number of estimated impacts. This result confirms the central role played by this experiment in the Special Issue, above and beyond the praise it has attracted for its sampling strategy and spillover estimates. However, we have shown the doubtful nature of the results obtained by this RCT. This further reduces the number of significant impacts, which were already impressively low.

Symptomatically, the transition from academic papers' results to the General Introduction, and then to the synthesis in the Policy Bulletin (J-PAL and IPA 2015) proceeds, by successive approximations, to simplify and magnify the lessons, even to the point of displaying erroneous results. If we go back to the summary of the impacts presented in the Policy Bulletin (Table 2, p.11; see also our Table 2), out of the 48 impacts measured (8 outcomes and 6 countries), 16 are announced as significant (14 positive and 2 negative). That is essentially wide of the mark. First, the significance threshold chosen is 10%, which is a level of precision at the upper limit of that which is usually used. If we adopt a more demanding threshold closer to standard practices (i.e. 1%), none of the 16 impacts is significant.

A more detailed analysis of the 16 selected impacts finds many inconsistencies. For Bosnia and Herzegovina, the impacts on *Business ownership* and *Business inventory/Assets* are announced as positive. But the first is not significant at 10%. As for the second, what is significant at 10% is a dummy variable measuring whether the firm owns capital or not. The impact on the total value of *Assets* is negative (although not significant), so at best null. For Ethiopia, the only impact considered significant and negative is that on *Household spending/consumption*. Yet consumption was not measured in the survey. In India, the two positive impacts are on *Business inventory/Assets* and *Business inventory/costs*. Neither impact is robust: the first impact is positive in the second endline survey, but not significant in the first survey, and vice-versa for the impact on *Business inventory/costs*. In Mexico, two positive impacts are noted. While the impact holds for *Business Revenue*, no data allows for a measurement of *Investment* (the second outcome assumed to increase with the treatment). *Assets* are furthermore decreasing (effect significant at 5%). In Mongolia, three outcomes are expected to have positive effects. This conclusion holds for two of them: *Business ownership* and *Household consumption* (at 10%). However, although the composite index of *Assets* is positively impacted (at 10%), the effect is non-significant (and even negative) for the *Assets* value. In the case of Morocco, where four outcomes are considered positive, we would refer to this RCT's abovementioned reliability issues. The synthesis of the Policy Bulletin appears biased, or at best highly imprecise.

Given these shortcomings, the high but non-significant coefficients would have been the same even if the sample sizes had been sufficient. These results have two implications. First, they place a question mark over the general statement that microcredit is not 'transformative'. This may be so, but nobody has produced any reliable evidence on this question. Second, Dahal and Fiala (2018) conclude that, "*Researchers and policymakers actually know very little about the impact of microfinance*"[20]. This paradox in view of the amount of resources put into RCTs on microcredit is confirmed by Jonathan Morduch (2020), one of the best specialists of microcredit worldwide (*Why RCTs failed to answer the biggest questions about microcredit impact*).

Another finding for both external and internal validity is the fact that none of the replication studies (Dahal and Fiala 2018; Kingi et al. 2018; Meager 2019) pointed up the errors we documented in our Moroccan replication. This includes the most obvious such as the authors' statements about the total absence of contamination in the control groups, the inconsistent household counts before and after trimming, and the claim that no trimming was conducted at baseline. This underlines the shortcomings of "push-button replications" or replications that apply different econometric specifications to the same data without checking the reliability of the original data, codes or sampling.

[20] This point is acknowledged in a roundabout way by the editors of the Special Issue: "*The individual studies may lack strong evidence for transformative effects on the average borrower, but they also lack strong evidence against transformative effects,*" (Banerjee, Karlan and Zinman 2015: 3).

Table 5: Impact, references and publications

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Impacts						
MFI credit take-up						
Data source	Survey	Survey	Survey	Admin, survey	Survey	Admin, survey
Presence of other MFIs	Yes	Yes	Yes	Yes	Yes	Yes
Substitution/Crowding effect	n.a.	No	Yes (substitution)	Yes (crowding-in)	Yes (substitution)	Yes (substitution)
Impact	Positive (98.5%)	Positive (25%)	Positive (13%)	Positive 8% (survey), 11% (admin)	Positive (50%)	Positive Survey (9%), 17% (admin)
Outcomes (others than credit take-up)						
#	47	37	99	37	41	37
# of sign. Impact (at 1%; 10%)	0/47 (1%) 3/47 (10%)	0/37 (1%) 5/37 (10%)	1/99 (1%) 13/99 (10%)	3/37 (1%) 9/37 (10%)	0/41 (1%) 10/41 (10%)	6/37 (1%) 17/37 (10%)
References, publications						
# of references:	22	24	27	28	37	16
Of which RCT	5	11	11	20	10	8
Of which methodology/theory	4	4	3	4	18	6
Of which other microcredit methods	6	7	4	4	5	0
Others	7	2	9	0	4	2
# of papers in academic journals	1 (AEJ:AE)	2 (AEJ:AE; Demography)	1 (AEJ:AE)	1 (AEJ:AE)	1 (AEJ:AE)	1 (AEJ:AE)

Source: Authors based on AEJ:AE (2015).

Note: The high number of outcomes in the Indian RCT (99) is due to the fact that two endline surveys were performed. sign. Impact (at 1%; 10%): number of impact estimates associated with p-values significant at the 1% level and at the 10% level.

IV. Results: from statistical biases to interpretative biases

Section 3 explores the fabric of RCTs in the field and highlights the many weaknesses of RCTs on microcredit in terms of their internal and external validity issues. The stages of statistical data collection and econometric analysis are then followed by the stage of interpretation: "*The beauty of randomized evaluations is that the results are what they are: we compare the outcome in the treatment with the outcome in the control group, see whether they are different, and if so by how much,*" (Banerjee 2007: 115-16). An analysis of how *randomistas* transform their data into scientific statements would appear to challenge this so-called "beauty" of RCTs.

Taken in isolation, most of the six RCTs' econometric results are meaningless in themselves, let alone in the absence of contextual information. The authors, particularly in the General Introduction, make this interpretation in a highly specific context and at the cost of implicit, but strong assumptions borrowed from a behavioural theory of change. Anthropology and political economy frameworks would return very different conclusions. Our purpose is not to disqualify the process of interpretation, which is inherent in data analysis, but to point out that *randomistas*, contrary to what they claim, cannot escape it. Results are not "what they are", as Naila Kabeer also shows using qualitative tools to revisit a field studied by an RCT (Kabeer 2019).

Moreover, their interpretation is based on a "persuasive rhetoric" (Labrousse, 2019), which consists of making a clean sweep of previous research and extrapolating (here, we find the problem of external validity), while overriding specific issues that are essential to understand the impacts of microcredit, and which other methods have already addressed.

Making a clean sweep of previous research

Randomistas' results are often presented as unprecedented 'discoveries', whereas they are often only the replication of conclusions obtained from previous studies, primarily those obtained from non-experimental methods that are almost never cited (Labrousse 2010). The General Introduction is a good illustration of this. The results are presented as the first scientific evidence of the impacts of microcredit. "*The evidentiary base for anointing microcredit was quite thin,*" (Banerjee, Karlan and Zinman 2015: 1). Up to this point, available empirical evidence had been based on "*anecdotes, descriptive statistics or impact studies that are unable to distinguish causality from correlation,*" (ibid: 1-2). The authors claim to be part of "*the debates that took place in the 2000s and continue today,*" (ibid: 2) but these debates are actually taking place in a surprisingly cloistered world. Of the 18 references in the General Introduction, 12 (two-thirds) come from the authors themselves and 17 (94.4%) from J-PAL members. Only one article escapes this endogamic principle.

No non-randomized studies are cited. Looking at the six articles in the Special Issue, the article on Morocco is equally exclusive (only RCTs are mentioned). The others are less so, although variably as shown in Table 5. The Bosnia and Herzegovina study is the most pluralistic, with an RCT/non-RCT ratio of 0.8; this ratio ranges from 1.57 to 5 for the others.

In addition to the disregard for available non-RCT evidence, there is a tendency to extrapolate and pass over key issues. Without claiming to be exhaustive, but focusing on the points that we feel are key, we address in turn the issues of take-up, business creation and freedom of choice, social transfers and self-reliance, saving and the problem of over-indebtedness.

Take-up

Low take-up is certainly the most accomplished result of the Special Issue. Many practitioners, decision-makers and researchers, even today, still predict an unlimited market, confusing financial exclusion with demand for credit. Although this result is useful, its true significance is limited. First of all, it should be noted that this exercise is nothing new. Some studies have long warned of low demand (Johnson and Rogaly 1997; Servet 2006), including providing quantitative estimates (Hes and Poledňáková 2013; Khandker et al. 1998). Moreover, the take-up rates referred to here are difficult to compare and interpret given the diversity of protocols and randomization methods (see above). It is therefore difficult to assess the nature and significance of the target population, and consequently to draw operational conclusions.

Moreover, the RCTs say nothing about the reasons behind the low take-up: does it reflect an intrinsically low demand and low propensity to get into debt, and/or does it reflect the inadequacy of the supply, with the two explanations not being mutually exclusive? Only more detailed data could answer this question based on a detailed analysis of financial practices, as seen with financial diaries (Collins et al. 2009) and their social, moral and political meanings (see, for example, the qualitative analysis of the Moroccan context, overlooked by the authors of the Moroccan RCT; Morvant-Roux et al. 2014).

Microcredit, self-employment and freedom of choice

The six studies in the Special Issue tend to concur that there are limited impacts on the creation of new businesses (significant only in two cases), with the expansion of existing businesses being more frequent (four cases). Improved profitability is found in only one case (Morocco), but we have seen above the low internal validity of these results. Moreover, even when a business is started up or expanded, no impact on income growth is observed, either because profitability is low or because self-employment income is offset by a decrease in paid work elsewhere. The authors of the General Introduction thus claim to draw a novel conclusion on the impact of microcredit on entrepreneurship.

However, since the late 1980s, numerous empirical studies have been conducted to measure the impact of microcredit.[21] The systematic review by Duvendack et al. (2011), conducted when RCTs were just starting to be used, draws two conclusions. First, a large number of quantitative studies, both experimental (including RCTs) and observational, are subject to multiple biases.[22] Second, when results are valid, they reveal a limited and heterogeneous impact, something that Morduch also observed in the late 1990s in his pioneering article on the partly unfulfilled promises of microcredit (Morduch 1999). So the results of the Special Issue do not look so new after all. More importantly, given the complexity of the causal chains induced by microcredit (Duvendack et al. 2011) and the heterogeneity of effects and types of microcredit,[23] RCTs do not seem appropriate (Bernard et al. 2012). Ultimately, the *randomistas*' question – Does microcredit work or not? – is poorly placed. What is shown by rigorous studies (whether quantitative, qualitative or mixed) is that certain types of microcredit may be useful for certain categories of populations and in certain contexts, but not others (Bédécarrats 2012; Copestake et al. 2016). For instance, the work of Copestake et al. in Peru and Zambia (Copestake et al. 2001, 2005) and Bouquet et al. in Madagascar (Bouquet et al. 2007) shows in detail which categories of populations benefit from microcredit and why and, conversely, which categories see their

[21] Bédécarrats (2012) identified 154 impact studies, compared with 51 for Duvendack et al. (2011).

[22] Several replications of non-experimental studies long used as 'evidence' of the positive impact of microcredit have revealed numerous biases and an overestimation of impacts. See Duvendack & Palmer-Jones (2012); Roodman & Morduch (2014).

[23] We shall give the example of rural microcredit, which is widely represented in the Special Issue. Over and above the credit modalities, what are the credit needs (inputs, equipment, livestock, cash flow to finance the lean season, etc.); what type of agriculture are we talking about (cash or food crops, agriculture in dry or rainfed areas, intensive or extensive, family-based or professional, independent or contractual through integration into agro-business sectors or producers' cooperatives, etc.); and what is the nature of the rural economies (degree of monetarization, remoteness and quality of infrastructure, non-farm income opportunities)? Last but not least, what kind of MFIs are we talking about? Status (for-profit/not-for-profit) is one thing (specified in the Special Issue), but other key questions include mode of governance, degree of integration and adaptation to local realities, and capacity to design products adapted to local demand. In view of this diversity, it makes no sense to talk about "rural microcredit". On this diversity, see for example (Morvant-Roux 2009).

situation deteriorate, with direct operational conclusions on how to transform the services offered. Still in Madagascar and ten years before the Special Issue, our own impact evaluation of a local MFI based on a quasi-experimental approach suggests three main stylized features presented later as "discoveries" by the randomistas: the impact of microcredit is not "transformative"; the impacts are heterogeneous across the firm size distribution; and context matters: microcreedit is more beneficial in time of growth than in time of crisis (Gubert and Roubaud 2011).

Understanding the heterogeneity of impacts (and drawing operational conclusions) requires a different conception of causality mechanisms, not in terms of "difference-making" but in terms of "mechanism" and "process" (Shaffer 2015). Moreover, given the many externalities, focusing on individual impact is also restrictive. Very few studies have applied general equilibrium models to the case of microcredit at mesoscale (for one exception, see Mahjabeen 2008). Examinations of externalities have been carried out mainly by political economy analyses, considering that it is precisely the analysis of the embeddedness of MFIs in their social, cultural, political and economic environment and externalities that has a powerful effect on product uptake and hence impact (Copestake et al. 2016). Convincing and useful impact studies in rural areas have shown the key role of financial innovations anchored in local territories – capable of developing specific products designed locally (bridge loans, guarantee funds and leasing) and combining with other measures (cropping contracts, harvest warehouse, technical assistance, etc.) – in enabling small farmers to upgrade their participation in various value chains (Bastiaensen and Marchetti 2011; Bouquet et al. 2007), while often encountering threshold effects (Doligez 2002). Effects are sometimes questionable, such as when microcredit accelerates migration processes, as migration is necessary to repay microcredits (Bylander 2014; Morvant-Roux 2013). They are sometimes more political and cultural than economic in nature. For example in Egypt, the introduction of microcredit disrupts local values – understood broadly as what makes sense to people – and thereby the processes of recognition, identity and socialization (Elyachar 2006). In rural South India, the massive presence of MFIs in certain territories reconfigures local power relations and chains of patronage by feminizing them (Guérin and Kumar 2017). These results (and their related questions) are far removed from those of the *randomistas*. And yet, if we really want to understand what microcredit is changing in people's lives, it is precisely these kinds of broad questions that need to be asked.

In addition to these in-depth studies, which are systematically based on sound knowledge of local contexts over time, it is useful to mention other, lighter methods designed to quickly identify the characteristics of customers (and non-customers) and the way in which services are used, and to derive recommendations for improving the quality of supply, which remains the key recurrent question asked by microcredit providers.[24]

We shall now come back to the Special Issue. Not only do the authors not bring anything fundamentally new to the existing evidence, but their interpretation of the quantitative results is problematic. Microenterprise may reflect absence rather than expansion of choices. A large proportion of micro-entrepreneurs, condemned to self-employment for lack of paid employment, resemble more the self-exploitation analysed by Alexander Chayanov (1966 [1925]) than the Schumpeterian entrepreneur (Lautier 2004). The case of Mongolia is instructive in this regard. The RCT shows that access to group credit allows women to start new micro-enterprises, but for negative incomes, while their working time increases by more than a third (without any change in household time). These negative effects are mainly observed for

[24] Examples include the tools developed by AIMS (Assessing the Impact of Microenterprise Services) and Imp-act, which have been denigrated for their lack of a sophisticated quantitative method. These tools may have lost their relevance to 'prove' impact on a large scale, but they have nevertheless been very useful to 'improve' and diversify the microfinance service supply.

less-educated women (Attanasio et al. 2015: 105, note 21). The authors believe that profitability may improve once the credit is repaid (Attanasio et al. 2015: 115). Here, we find the problem of temporality, already highlighted as a strong limitation of the RCT (Bédécarrats, Guérin and Roubaud 2019; Labrousse 2010). These women may indeed have chosen to embark upon the entrepreneurial venture, and this may explain the improvement in consumption (results indicate more and healthier consumption). But what is the meaning of this 'choice' and, above all, what are its consequences if it then gives rise to increased responsibilities and possibly disengagement by other household members (and hence intra-family inequalities)? The quantitative data do not enable a conclusion to be drawn, and the authors of the RCT do not make any particular judgement. A robust interpretation would call for other types of data, quantitative or qualitative. The authors of the General Introduction, on the other hand, focus only on the 'freedom of choice' dimension, without mentioning the potentially negative effects of these 'choices' on women, especially the most disadvantaged.

Microcredit, social expenses, social transfers and self-reliance

While the effects in terms of business and income are inconclusive, the authors of the General Introduction observe what they describe as positive effects on two indicators: "non-essential expenditures", a sign of better discipline and management skills, and a decrease in "social transfers", a sign of greater autonomy. "Non-essential expenditures" include "temptation goods" and decreased in four countries (they were not measured in Ethiopia and the results were not significant in Mongolia): alcohol and cigarettes in Bosnia-Herzegovina; cigarettes, sweets and soda in Mexico; and alcohol, tobacco, betel leaves, gambling and food consumed outside the home in India. These expenditures also included festivals, with decreases observed in India and Morocco.

The authors put forward a number of explanations to explain this decrease in "temptation goods": repayment and investment constraints, better self-discipline, and more involvement of women in decision-making. The reduction in temptation goods is one of the major results of the Indian RCT, highlighted in the abstract. The study's authors take care to specify that it is the populations themselves who describe these goods as "temptation goods", in the sense that they would like to reduce them (Banerjee, Duflo, Glennerster, et al. 2015: 24). But for people to express this preference (an observation of vague origin, seeming to be more a matter of "anecdotes" whose use is highlighted by A. Labrousse, 2019) may well reflect that they have taken on board the moralizing discourses frequently given by development organizations (including MFIs), and this since the colonial period.[25]

Moving beyond the moralizing dimension of the *randomistas*' conclusions,[26] a detailed analysis of the meanings and role of these outlays could shed a different light. On the subject of alcohol, no one would dispute that excessive consumption is a public health concern. Yet if we really want to understand this type of consumption and devise courses of action, it is essential to recognize the social and political dimension of alcohol. Like many other temptation goods, and contrary to what behavioural economics suggests, it is not a good defined solely by its "immediate utility" (Banerjee and Mullainathan 2010). Alcohol can play a social role since it enables workers to endure physical work and access socialization spaces and therefore strategic information (bars are often strategic places to negotiate employment contracts and orders; Picherit 2018). Alcohol can play a political role when it gives workers the opportunity to make demands of employers and bosses that are more easily acceptable under the influence of drunkenness (Scott 1977).

[25] In India, for example, reports from British settlers and Christian missions in the early 19th century already mentioned the improvidence and prodigality of the poor (Cederlöf 1997; Hardiman 2000).

[26] The statements made by the *randomistas* are reminiscent of the Victorian morality of the European industrial revolution, legitimized by the arguments of neoclassical economists of the time. Faced with the extreme poverty of the working class world during the British Industrial Revolution, some lamented the poor's lack of self-reliance, lack of foresight and wasteful alcohol expenditure, and argued for financial education courses rather than wage increases (see for instance (Jevons 1883: 196-200; 205).

Above all, alcohol is frequently deliberately offered by employers and labour recruiters in order to build loyalty (Picherit 2018). To think that sacrifice or more self-control would be enough to fight these 'temptations' is therefore fallacious.

Similarly, catering expenses (meals and tea) outside the home are not solely "lucrative opportunities to save" (Banerjee and Duflo 2011: 170). Street restaurants and tea shops are eminently strategic places. In an informal opaque economy, structured by interpersonal relationships, these spaces enable traders to keep each other informed of the market situation, price trends, opportunities to be seized, possible sources of financing, risks of tax or police checks, etc. Small entrepreneurs cultivate exchange and mutual support links, whose role is often decisive for the survival of their business.

On the subject of expenditure on social and religious rituals, anthropology has long shown that 'social wealth' is an essential factor of success and protection (Guyer 1997) and that 'investing' in social relations can, in certain situations, be much more rational than trying to save money by cutting oneself off from one's surroundings (Narotzky and Besnier 2014). Beyond *randomistas*, the question of "community taxes" and their cost benefits in terms of protection has been the subject of various studies by development economists. But these studies rarely take into account the complexity of the financial channels to which these expenses give rise and their long-term nature. An analysis conducted in India of the correlation between festival expenditure and lunch invitations shows, for example, that these expenses act as safety nets (Rao 2001). Moreover, what is considered by economists as an expense is sometimes conceived as an entitlement or as savings, since it will give rise to a future counter-gift. Also in India, accounting for all debts and receivables generated over time by ceremonial spending, which families are well aware of because they calculate in these terms, shows that families' net financial wealth is radically different from that suggested by an analysis in terms of 'spending' (Guérin et al. 2019). This contradicts the short-term bias that *randomistas* often attribute to the poor (Banerjee and Duflo 2011: 183-204).

With regards to social transfers, of the eight estimates retained (which concern transfers from the family or the State), five are negative. This observation leads the editors of the Special Issue to conclude that "self-reliance" has improved, a factor that is judged in a positive light.[27] This interpretation is both risky – there is no reason to believe that the decrease in transfers from family and friends is seen as positive or a source of well-being by the people concerned themselves – and normative, as are previous interpretations. Here again, anthropology is valuable in elucidating the decisive role of social interdependencies, in terms of both material protection and identity. Looking past the *randomistas,* there are those in the development world – policymakers, practitioners and some researchers – who consider dependency both as a political problem (assistance is expensive) and as a moral problem (dependency is seen as being incompatible with individual freedom). However, in many contexts, being connected and dependent on others is both a mode of action and a deliberate strategy. Rather, people's agency translates into the ability to choose certain forms of dependency and interdependence.[28]

Ultimately, the General Introduction's conclusion on improving self-reliance, as well as that on "freedom of choice", is driven by specific interpretations of econometric results (if not extrapolations from the conclusions of some of the RCTs). These interpretations are underpinned by a singular conception of individual autonomy and freedom, and thereby their own theory of change, viewing people as isolated atoms, denying the multiple roles that social interdependencies play at different levels and implicitly considering these interdependencies as harmful. These two conclusions – "self-reliance" and "freedom of choice" – were nonetheless

[27] It should be noted, however, that this interpretation is that of the authors of the introduction, and not of the authors of the papers, who either do not comment on this result or underline its ambiguity. On Mongolia, the authors mention, for example *« Increased within-group financial discipline may come at the cost of disrupting informal credit and insurance systems based on kinship and other social ties"* (Attanasio et al. 2015: 114).
[28] For a general overview of how anthropology addresses this issue, see for example Ferguson (2015).

included in J-Pal and IPA's Policy Bulletin (2015), which was then widely disseminated by many blogs and discussion networks and seen as an indisputable asset of this research.

Microcredit and saving

An analysis in terms of interdependence also takes a different look at the results of RCTs on another microfinance service: savings. The Special Issue does not mention it, but the *randomistas* give it particular importance (Banerjee and Duflo 2011: 183-204; Karlan et al. 2014).

Why do poor people save little, even when they are offered innovative services and it is in their best interest to secure their meagre deposits? The reason why savings do not attract so much support is that the poor act as impulsive and impatient consumers, explain Banerjee and Duflo (2011), and because their time scale is one of immediacy. Insights from human psychology would explain this paradoxical result. The poor suffer from "procrastination" - putting off until tomorrow what can be done today - and lack "self-control". Imagine the savings that a small vegetable vendor could make if he drank two cups of tea a day less, write Banerjee and Duflo (2011: 171). Taking into account the interest paid on the money he has to borrow because of lost earnings, he would save 40 rupees a day... and the authors describe this little vegetable seller struggling with the temptations we all know from our impulsive behaviours, such as eating chocolate or smoking a cigarette.

Temptations and impulsive behaviours are not the prerogative of the poor, they tell us again, but it is more difficult to control oneself in stressful situations and it is for this type of population that the consequences are the most dramatic. The lack of confidence and optimism is also problematic: people who are confident that their aspirations will be fulfilled have good reason to save money, stop frivolous spending and invest in the future. On the other hand, people who think they have nothing to lose tend to make decisions that reflect their desperation. Optimism and hope can make all the difference, conclude Banerjee and Duflo (2011: 201).

In another example, from a study in the Philippines, the same authors mention the debts of small street vendors that microfinance promoters have tried to reduce. After 10 weeks, 40% of them were still debt-free: they were "patient enough" not to give in to the temptation of credit. But in the end, all of them fell back into debt and never got out of it, probably discouraged by a lack of self-discipline (Banerjee and Duflo, 2011: 202).

Behavioural analysis obscures the complexity and diversity of motivations as well as the structural conditions that restrict choice or narrow the time horizon for reasoning. Behavioural analysis is also completely blind to long-term projections. Yet precarious populations are far from being without them. They simply have other criteria for assessment and calculation. They have multiple savings practices (in kind, or in the form of "expenditure-investment": such as loans to others and ceremonial gifts, as mentioned earlier), which in turn relate to variable temporalities. Contrary to common biases, which are particularly marked in behavioural economics, the poor may have long-term prospects. People have their own conceptions of savings, as something that must "circulate": immobilisation, for example in a bank account, is often considered useless, or even experienced as dispossession. Beyond the individual perspective, there is a structural dimension to savings practices: tax incentives but also social norms (e.g. in India, where Dalits find it much harder to save in the form of real estate and are confined to gold) (Goedecke et al. 2018).

Here again, the lessons of economic anthropology are helpful. They invite us to situate monetary savings practices in the light of all the practices of storage, accumulation and circulation of value. These practices take many forms, and respond to equally multiple logics and constraints (Douglas & Isherwood 1980). The achievements of economic anthropology also lead us to reconsider certain postulates of economic analysis, especially those related to the unit of analysis and the scale of temporality. Individuals do not reason systematically as atomised individuals, but as members of a collective, which may be a household, a lineage or a clan. Not that individual reasoning and motivations should be excluded, but social interdependence, both for identity reasons and for material reasons due to the lack of social protection, is still prevalent. Ensuring the social reproduction of the group, both in its material and statutory dimensions, is often a predominant objective, inscribed in the long history of relations with ancestors and future generations, sometimes that of the Gods and the cosmic order.

While economic theory sees savings as a residual component of income - what remains after consumption - economic anthropology shows that it is often a practice influenced by cultural norms. Depending on cultural and social contexts and periods of history, hoarding is sometimes seen as a sign of wisdom and responsibility, sometimes as a sign of greed, selfishness and avarice (Douglas and Isherwood 1980). Both the propensity to save - the share of income spent on savings - and the forms of savings vary greatly across contexts and social groups. Like consumption, savings practices have an essential social and symbolic function. They are signs, codes, and social markers that manufacture identities and social relationships and are part of a process of constantly reinventing social categories (Guyer 1997).

If monetary hoarding is often low, as economists often say, it is a question of security (risk of theft, fire, flooding, etc.) and it would then suffice to offer secure services. However, limiting hoarding can be a deliberate choice aimed at protecting oneself against the demands of the entourage while at the same time "investing" one's own funds. Here we find the ambiguity of the social interdependencies mentioned above: people are constantly trying to comply with them while circumventing them. "Blocked money is useless," Tamil villagers often say. It is both a necessity - people are always in dire need of liquidity - and a means of maintaining social networks. In Senegal, it is often said that money "burns" because it circulates so quickly. As soon as it is received, any cash flow is either spent or re-injected into local circuits as an "investment" - women use this term - that can be recovered at any time in case of pressing need, with precise accounting practices that guarantee reciprocity. To the question "are you saving? "it is not uncommon for people to answer in all ingenuity that they lend... lending to others is then considered one way among others to save (Morvant-Roux 2009; Rutherford 2000). In many rural communities, any form of wealth is subject to loans if the owner does not have immediate use for it: coins and banknotes, but also jewellery (e.g. India), bricks (e.g. Argentina), cereals or other food products (e.g. Madagascar), livestock (e.g. Guinea, Morocco, Mexico or Tanzania), etc. This circulation of wealth is based on collective management of liquidity and mutualisation of surpluses. It makes it possible both to hide one's own wealth and to strengthen social ties (Shipton 2010). Most economic analyses ignore the existence of these financial circuits, which are all forms of savings since each loan/grant must be repaid or returned.

Behavioural assumptions also ignore the prevalence of non-monetary savings practices and the logic behind them. Banerjee and Duflo mention some informal savings practices, but state at the outset that they are complicated, impractical and a poor substitute for banking services (Banerjee and Duflo 2011:187). However, it is often more justified for poor people to save in kind, for example in the form of livestock, jewellery, grain or precious fabrics. Here again, the choice of savings vehicles is based on sophisticated calculations and combines several criteria such as security, liquidity or speculation. The cost of raising small livestock for resale in case of need or for anticipated expenditure is much more profitable than opening a bank account. Saving in kind is also a way of protecting against inflationary risks, or even speculating when the good in question is subject to regular increases in its price, such as gold or grain, for example. Identity and status issues also come into play, since goods used as savings often have an ostentatious function.

In many countries, particularly in Africa, livestock perform multiple functions. For the Fulbe pastoralists of Burkina Faso, livestock is an ostentatious asset, a source of reputation and prestige, but also a source of accumulation. It plays the role of precautionary savings: in an emergency, an animal can easily be sold on the nearby market. In a context where individual enrichment is reproved, the dispersion of herds and transhumance practices allow a certain discretion (Lont & Hospes 2004). In some Moroccan countryside, small livestock is a quasi-liquid asset and a large part of the local population, including small traders, manage their cash flow through their sheep herds, which are made and broken up according to hazards and opportunities (Morvant-Roux et al. 2014). In Tanzania, the central role of livestock is found for pastoralists, while for farmers it is cereals that play this role, notably rice (Lont and Hospes 2004). Livestock are not the only good perceived as savings. In many Asian countries (e.g. India, Sri Lanka, Indonesia and Malaysia), gold in the form of jewels is considered to be the main source of savings (Lont and Hospes 2004).

Microcredit and over-indebtedness

A major conclusion of the Special Issue is that microcredit is not the "debt trap" denounced by microcredit opponents. First of all, it should be noted that no scientific studies are mentioned in the General Introduction, as if the "debt trap" were anecdotal evidence. It is true that when a number of microcredit repayment crises erupted, the media made a big deal of it (in the same way as they had praised microcredit when it started). However, the press aside, there is a vast body of scientific literature dealing with household over-indebtedness in the Global South and the role played by microcredit, including in the countries covered by the Special Issue. A number of problems arise here.

The first concerns external validity, where extrapolation occurs without taking into account the singularity of the contexts studied and the fact that this is microcredit “on the margin” (Wydick 2016). The six RCTs focused on areas and populations that were supposed to be free of microcredit.[29] However, by definition, the problem of over-indebtedness is less acute than in areas and populations previously exposed to microcredit. It is therefore tautological that the "debt trap" does not appear. Yet over-indebtedness among some of the microcredit clients has been documented and sometimes measured in four of the countries studied.[30] The fact that the RCTs did not quantify it does not enable them to conclude that the debt trap does not exist. Contrary to what the authors of the General Introduction suggest, the available literature is not content to make do with “anecdotes". Scholars demonstrate (most often qualitatively) the role of microcredit based on a detailed analysis of its specific characteristics in relation to other sources of debt, in particular the rigidity of the repayment terms and low tolerance for non-payment.[31] In some contexts and MFIs, this zero tolerance takes the form of coercive enforcement procedures.[32] These scholars also propose a nuanced and contextualized analysis, highlighting the role of the global context (including stagnant and declining real incomes in the face of growing needs) as well as the ambivalent role of microcredit (for some borrowers, microcredit can be a way to repay informal debts and reduce over-indebtedness).[33] The causal link between microcredit and over-indebtedness may only concern a minority of microcredit clients (which brings us back to the issue of heterogeneity). But its repercussions (impoverishment, social exclusion, suicide, etc.) (Schicks 2013) are sufficiently tragic to warrant *randomistas* taking the phenomenon more seriously.

[29] As mentioned above, this virginity was in fact a decoy and all control populations actually had access to microcredit. However, the market was not saturated as it might have been elsewhere, so there was less of a risk of over-indebtedness.

[30] For Mexico, see Morvant-Roux (2013), Angulo Salazar (2013), Hummel (2013), Rozas (2014). For India, see Guérin et al. (2013), Joseph (2013), Taylor (2011), Prathap and Khaitan (2016). For Bosnia-Herzegovina, see Maurer and Pytkowska (2014); Opem and Goronja 2013; Bateman 2010). For Mongolia, see Javoy and Rozas (2013).

[31] In addition to the references in note 26, see (Schicks 2013; Schicks and Rosenberg 2011; Guérin, Morvant-Roux, and Villarreal 2013; Guérin, Labie, and Servet 2015).

[32] In India, for example, the prosecution of defaulters in the workplace or at home, public denunciations and insults, solicitation of relatives, physical threats, confiscation of property and administrative documents; in some cases, the most recalcitrant have been tied up in a public square or in direct sunlight (Arunachalam 2011; Servet 2011).

[33] As we finalize this paper (October 2019), the United Nations has just taken up this issue, commissioning a report on the subject. This would seem to suggest that the problem does exist. https://www.ohchr.org/EN/Issues/Development/IEDebt/Pages/ReportPrivateDebt.aspx

The second problem is the extrapolation from the six case studies by the introduction's authors. Even for areas and populations recently exposed to microcredit, over-indebtedness cannot be ruled out. The Bosnia and Herzegovina RCT was conducted in a context of a proven over-indebtedness crisis, which the authors mention as contextual data (Augsburg et al. 2015: 185). This RCT specifically concludes that the treatment group had repayment difficulties (Augsburg et al. 2015: 199-201), and that these repayment difficulties are a potential symptom of over-indebtedness.[34] The RCT does not enable a conclusion of either the existence of over-indebtedness or the role of microcredit. However, the existence of a "debt trap" cannot be excluded. In the Mongolian RCT, the authors take care to specify that their study does not measure over-indebtedness, but only repayment defaults, which are two distinct things.[35] The special issue's introduction makes no reference to these clarifications.

In Morocco, a qualitative study conducted by one of us at the same time as the RCT concluded that there was low propensity for debt in rural areas, for cultural and religious reasons (Morvant-Roux et al. 2014). This general observation, valid "on average", does not, however, exclude over-indebtedness problems among a fraction of the population. Given that Morocco also experienced a default crisis (which the authors do not mention, although it took place during the RCT), MFIs concentrate their supply on a minority of clients judged solvent and reliable. These clients are hence overexposed to microcredit, and some of whom do face over-indebtedness problems (Morvant-Roux and Roesch 2015).

Like Bosnia and Herzegovina, India has been hit by some major microcredit default crises: in Krishna District in Andhra Pradesh back in 2006, then in a small town in Karnataka in 2009, and in the entire state of Andhra Pradesh in 2010. Analyses of this crisis, both quantitative and qualitative, have highlighted the existence of an over-indebtedness problem for some of the clients. The over-indebtedness of poor populations, with or without microcredit, has also been documented outside of default crisis areas, including in urban areas. As already mentioned, the Indian RCT was conducted from 2005 to 2010 in marginal areas of Hyderabad newly exposed to microcredit (in the knowledge that the area was not completely virgin, see above). Yet how is it possible to extrapolate from this highly specific case study when there is a vast body of evidence demonstrating the existence of over-indebtedness? On this issue, the article by Banerjee et al. cites just one press article, "*Anecdotes about highly successful entrepreneurs or deeply indebted borrowers tell us nothing about the effect of microfinance on the average borrower, much less the effect of having access to it on the average household*," (Banerjee, Duflo, Glennerster, et al. 2015: 23). In view of the state of the art's alert over the level of over-indebtedness among poor Indian populations, and given the extreme specificity of the districts they study, is it not maybe their own study that should be qualified as anecdotal?

Finally, the question may be put as to whether the measurement of household debt was properly conducted. The collection of reliable debt data calls for a number of precautionary measures for the following reasons: debt taboo, exacerbated when MFIs claim to eradicate informal borrowing since this encourages clients to conceal their informal debts; diverse terminology used; and the range of debts that may be held by different family members without their necessarily sharing that information. Given the approximations observed at the other stages of data collection and analysis (see Section 3), it is not unreasonable to question the ability of the *randomistas* to design a questionnaire that can adequately capture household debt. However, it should be noted that this difficulty is not unique to the *randomistas*. Collecting reliable data on incomes in the Global South has taken decades of learning to adapt the statistical tools to contexts where households juggle different sources of income, including informal sources. The same work has yet to be done on debt, which remains poorly measured and often underestimated.

[34] Defaults can also be 'strategic' defaults expressing a refusal to repay, particularly in the context of a repayment crisis.
[35] Since some defaults can be strategic, good repayment rates can mask sacrifices made to honour debts, which the authors of the RCT in Mongolia acknowledge (Attanasio et al. 2015, footnote 25, p. 114).

V. Conclusion and discussion

Given the many limitations and shortcomings we have found with the method, applied here to microcredit, the question could be asked as to why RCTs have had such academic, media and political success. We have already explored the reasons for this contradiction (Bédécarrats, Guérin and Roubaud 2019) in a study of the political economy of what has now become a real industry (Ravallion, 2019). As with any industry, the impact evaluation market is where supply meets demand. We have explored these two elements in detail, showing that the demand is twin-engined, driven by both the donor community and the academic world, while the supply is largely shaped by a brand of scientific businesses and entrepreneurs who appear to have created a new business model designed to build a monopoly and a rent position on the evaluation impact market. Further illustrations of this would-be domination strategy are turned up when exploring how the data have been produced and analysed, as we have done here. In addition to making a clean sweep of the past (see above), three other strategies appear to be key: disengagement from a "data culture", ignoring criticism (up to a certain point) and sidestepping certain rules of scientific ethics.

Disengagement from a data culture

The many data collection and data entry errors observed in the Moroccan RCT would appear to suggest a certain 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) excused the researchers from the need for the know-how required for the first stage (collection of good quality data). To what extent does this concern apply to other RCTs? Unfortunately, that question remains open for the moment, since only full replications would be able to provide the answer. What is clear however, is that *randomistas* tend to disregard the debates regarding data collection (as they do the issue of ethics; Abramowicz and Szafarz, 2019). In most quantitative empirical research protocols, there is a division of labour between data collectors and analysts: the former are statisticians, the latter are economists (econometricians or thematicians). With few exceptions (Deaton 1997; Grosh and Glewwe 2000), few people can occupy both ends of the spectrum. These are fully-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 statisticians have less social prestige than economists (Desrosières 2013). Given the skills involved and the way academic journals work, all efforts are concentrated upstream on designing a "smart" randomization process, and downstream on econometric estimations of the impacts with a view to publishing papers in top-ranking reviews.

The disconnect between researchers and the field is another illustration of the data culture. This disconnect is particularly acute at J-PAL. Its hierarchical organization makes for 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 arguably not adequately trained (Jatteau 2018). This division of labour is a practice 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. Moreover, teams are required to adhere to precise protocols to validate the rigour of the experiments conducted. This cannot be not the case here given the dozens of RCTs in which the most prominent RCT leaders are involved (Bédécarrats, Guérin and Roubaud 2019). This disconnect has been exacerbated by J-PAL's exceptionally rapid expansion, as mentioned above.

This growth, combined with highly centralized governance, 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 separate RCT (and deepens the disconnect with the field). In February 2019, Esther Duflo had 64 RCTs to her name, equal to just over four new RCTs a year. Dean Karlan, however, is by far the most prolific with 100 trials (and 42 ongoing; January 2017).

So how much can they really personally put into each of the RCT results they sign? In fact, the signature of a top *randomista* researcher appears to be more of a seal to facilitate publication in a top-ranking journal, as part of a global *randomista* strategy, than a guarantee of research quality.

Ignoring the critics

Whereas *randomistas* have built a universal narrative on the impact of microcredit based on this Special Issue (and subsequent publications), other players have drawn different conclusions from these same studies (see also Kabeer 2019). Here again, the Moroccan RCT is a typical illustration. As early as 2009, while the endline was still in progress, the RCT's funder started publicly sharing its feedback on RCTs based on the Moroccan RCT and another study conducted in Cambodia at the same time. The conclusions were clear: they highlighted the challenges faced by the method to produce rigorous impact evaluations given the multiple breaches of protocol that the funder's research team had partially identified (problem of representativeness and product change) and the time constraints that compelled a focus on the short term. Although the findings of the funder's research team have been publicly presented and published on numerous occasions (Bernard et al. 2012), they have gone unheeded by the RCT team (Bédécarrats, Guérin, Morvant-Roux, et al. 2019b).

Our own experience with the Moroccan RCT, although illustrative, is a good example of what might be a strategy to ignore the critics, up to a point. In the course of our critical research on RCTs in development, we have invited some of the most vocal RCT proponents to engage in a scientific debate (controversy) on many occasions (dedicated sessions at international conferences). To date, we have received no answer. We also invited ten of the most famous *randomistas* to take part in this collective book to balance out the voices on RCTs. They all declined. Directly on the subject of our critical review of the Moroccan RCT, we informed the authors of the completion and publication of our replication (Bédécarrats, Guérin, Morvant-Roux, et al. 2019a). At the same time, we drafted a *Comment* and suggested that AEJ:AE publish the piece with an *Answer to the Comment* from the authors, as is common practice in many journals. AEJ:AE turned down the offer on the basis that the journal does not publish comments. Lastly, when our paper was picked up by coverage in vocal blogs and the press, Crépon et al. (2019) produced a (51-page) *Rejoinder* using sophisticated analyses to argue that their original results were robust: double post lasso procedure, Benjamini-Hochberg False discovery rate correction of multiple testing, the Bayesian hierarchical model and machine learning analysis, among others, concluding our replication was not scientific. They posted the *Rejoinder* on their website and enjoined us to post it on the DIAL website, which we duly did. They also informed the AFD hierarchy. IREE suggested both parties publish a short version of the *Rejoinder* with our answer (*Rebuttal of the Rebuttal*; Bédécarrats, Guérin, Morvant-Roux, et al. (2019c). In view of the totally contradictory conclusions of the two pieces, we suggested seeking a third party assessment that would decide on whether to retract our replication (Bédécarrats, Guérin, Morvant-Roux, et al. 2019a) or the initial paper (Crépon et al. 2015), depending on the conclusion. Again, they declined the invitation. These episodes illustrate two characteristics of the *randomistas'* make-up. First, contrary to one of the main selling arguments for RCTs (the simplicity of the method, compared to the so-called 'black box' of alternative econometric methods), this type of RCT is extraordinarily complex. In their *Rejoinder,* they added complexity to the already complex randomization design (which is one of the three paradoxes we sought to explain in Bédécarrats, Guérin, Morvant-Roux, et al. (2019b)). Second, they sidestepped scientific standards by not providing their codes, turning down a peer review of their *Rejoinder*, and ultimately eluding a fair scientific controversy.

Circumventing scientific ethics

In addition to disregarding all things non-RCT, the *randomistas* have bypassed certain basic rules of scientific conduct. This problem appears to be growing in the scientific community as a whole (Heckman and Moktan 2018). Yet while it is not specific to J-PAL or the *randomista* community, it is particularly patent here. In the research world, knowledge validation is based on the "peer review" principle, that is a collective review by 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 journals' editorial boards. Editorial favouritism is a recognized and demonstrated process, particularly among economists (Fourcade et al. 2015). The Special Issue is illustrative in this regard. The issue's three scientific editors are members of J-PAL (Banerjee, Karlan and Zinman 2015). In addition to the General Introduction, each editor co-signed an article and two of them were members of the board of editors (Banerjee and Karlan). Esther Duflo is both the journal's editor (and founder) and co-author of two of the six articles. Given, in addition, that nearly half of the articles' authors (11 of the 25) are also members of J-PAL and four others are affiliated professors or PhD students with J-PAL, the journal strayed somewhat from the peer review principles supposed to govern scientific publication. This single example shows in cameo the extraordinary density of the links between RCT promoters identified by Jatteau (2016).

What remains of the special issue?

At the end of the day (or of our in-depth investigation), what have we learned from RCTs on microcredit in the development field? Going back to this paper's title, if microcredit is not a miracle, as defended by the Special Issue, what are RCTs on microcredit: *miracle or mirage*? Let us wrap up our results and provide key takeaways.

We will start by addressing the internal validity claims, the acclaimed strong points of RCTs. First, as acknowledged by *randomistas* themselves, there is a lack of strong evidence that microcredit is transformative, just as there is a lack of strong evidence that it is not (Banerjee, Karlan, et al. 2015). Given that RCTs are generally underpowered due to low take-up and compliance, we simply do not know. Second, and again acknowledged by the *randomistas*, heterogeneous effects may be the norm. Microcredit may be transformative for some and not for others (or worse, microcredit may be negatively transformative). Again, given the general underpower of RCTs due to low take-up and compliance, we simply do not know. Furthermore, we do not know why some may benefit from microcredit and some may not (or may suffer a "transformative" penalty). We have no idea through which channels microcredit might have an impact. Third, poor data quality and measurement errors may prompt reconsideration of some of the results that have hitherto been taken for granted. In this respect, the many problems we have identified with the Moroccan RCT need to be taken seriously. Maybe the Moroccan RCT is a one-off (the bad apple). But in this case, its conclusions should be definitively revoked. This would have two direct repercussions. The overall demonstration would be weakened. The "*fairly representative sample*" used to draw general conclusions would become "*less fairly representative*". Its good properties put forward in the issue to estimate spillover issues and predict take-up rate, and its sampling strategies to address the issue of low compliance and underpower would evaporate. Maybe it is not a one-off (although we presume that other RCTs could not perform as poorly), in which case we have a structural problem here. The only way to know would be to conduct full replications, such as ours. We strongly advocate this avenue of research. Fourth, we have shown that many interpretations of the impact of microcredit, underlying the theory of change, are biased, while some obvious impacts (or causes of low take-up) are not even considered. Additionally, other generic concerns remain such as general equilibrium effects, macro policies, etc. (both are internal and external validity concerns).

Second, external validity has never been the RCTs' strong point. Our assessment does nothing to change this view. The usual criticisms, not worth quoting again here, still hold. The Special Issue's novelty is that it considers different RCTs on microcredit taken together in tandem. However, the accumulation of individual cases does not solve the problem. What is gained from diversifying geographic, but hyper-specific contexts, is lost from increasing the heterogeneity of treatment, implementers, and so on. One type of product may work in one context and not in another. Changes to products and allocation schemes do not tie in with "real world" conditions.

Lastly, ethical issues remain largely unaddressed despite major departures from good practices in the medical field and even social RCTs in developed countries.

Taking all that into account, what is left? To paraphrase Banerjee and Duflo (2011) as quoted by Agnès Labrousse (2019), we can follow up, nearly ten years and dozens of RCTs on microcredit later, with, "*Unfortunately, [...] ~~until~~ even very recently, there ~~was~~ is in fact very little evidence, either way, on these questions. What ~~CGAP~~ randomistas call~~s~~ evidence turns out to be case studies [...]*." Although it is not clear what is left at this stage, what is not left is the huge amount of money and resources spent, some which were withheld from other alternatives and uses. Is it worth spending millions of dollars in return for one single academic paper for each RCT (Table 5)? Wouldn't it be more useful for the same sums to be used to fund a developing country's public statistical system to collect a huge amount of representative observational data in the long run? Although RCT proponents have acknowledged some of the methodological shortcomings discussed in this paper, their answer to resolve them is, "More RCTs!" Yet if RCTs have not delivered on their promises, or at least the promises that *randomistas* have been selling the world these past two decades, then it would be just as legitimate to say, "No more RCTs!"

We may come across as extreme. Yet the *randomista* tidal wave has been so powerful (as seen from the way they have swept aside the past by (apparently) ignoring all non-experimental studies) that a small push back in the other direction would do no harm to rebalance the state of the art. 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. Rather than "No more RCTs," our advice is actually, "No more standalone RCTs." While RCTs are likely to remain appropriate and legitimate for certain precisely circumscribed policies, they should still be conducted by the book. Furthermore, they are never self-sufficient. It is both necessary and possible to use other methods without compromising scientific rigour. 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 a requirement for all development issues, projects and policies not suited to RCTs, and microcredit with its relatively closely targeted interventions is a good example of this given the low take-up and complexity of its effects. 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, 2019)**.**

List of references

Abramowicz, M., & Szafarz, A. (2019). 'Ethics of RCTs: Should Economists Care about Equipoise?'. in Bédécarrats F., Guérin I and F. Roubaud (Eds), *Randomized Control Trials in Development: A Critical Perspective*, Chapter 10, Oxford: Oxford University Press (forthcoming).

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

Angelucci, M., Karlan, D., & Zinman, J. (2015). 'Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco', *American Economic Journal: Applied Economics*, 7/1: 151–82.

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

Angulo Salazar, L. (2013). 'The Social Costs of Microfinance and Over-indebtedness for Women'. Guérin I., Morvant-Roux S., & Villarreal M. (eds) *Microfinance, Debt and Over-indebtedness. Juggling with money*, pp. 232–52. Routledge: London.

Arunachalam, R. S. (2011). *The journey of Indian micro-finance: lessons for the future*. Chennai: Aapti Publications.

Attanasio, O., Augsburg, B., De Haas, R., Fitzsimons, E., & Harmgart, H. (2015). 'The impacts of microfinance: Evidence from joint-liability lending in Mongolia', *American Economic Journal: Applied Economics*, 7/1: 90–122.

Augsburg, B., De Haas, R., Harmgart, H., & Meghir, C. (2015). 'The impacts of microcredit: Evidence from Bosnia and Herzegovina', *American Economic Journal: Applied Economics*, 7/1: 183–203.

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

Banerjee, A., Breza, E., Duflo, E., & Kinnan, C. (2019). *Can Microfinance Unlock a Poverty Trap for Some Entrepreneurs?* National Bureau of Economic Research.

Banerjee, A., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs.

Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). 'The miracle of microfinance? Evidence from a randomized evaluation', *American Economic Journal: Applied Economics*, 7/1: 22–53.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., et al. (2015). 'A multifaceted program causes lasting progress for the very poor: Evidence from six countries', *Science*, 348/6236: 1–16.

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

Banerjee, A., & Mullainathan, S. (2010). 'The Shape of Temptation: Implications for the Economic Lives of the Poor', *NBER Working Paper No. 15973*.

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

——. (2020). 'Finding our balance? Revisiting the randomization revolution in development economics ten years further on'. *World development*, 127 104789.

Bastiaensen, J., & Marchetti, P. (2011). 'Rural Microfinancr and Agricultural Value Chains: Strategies and Perspectives of the Fondo de Desarrollo Local in Nicaragua'. Armandariz B. & Labie M. (eds) *The Handbook of Microfinance*, pp. 461–95. World Scientific Publishing,: London/Singapore.

Bateman, M. (2010). *Why doesn't microfinance work? The destructive rise of local neoliberalism*. London: Zed Books.

Bédécarrats, F. (2012). 'L'impact de la microfinance: un enjeu politique au prisme de ses controverses scientifiques', *Mondes en développement*, 2: 127–142.

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

——. (2019b). 'Lies, damned lies, and RCT : une expérience de J-PAL sur le microcrédit rural au Maroc'. *Working Paper 2019-04*. DIAL: Paris.

Bédécarrats, F., Guérin, I., Morvant-Roux, S., & Roubaud, F. (2019c). *Verifying the internal validity of a flagship RCT: A review of Crépon, Devoto, Duflo and Pariente. Rebutting the Rebuttal* (DIAL Working Paper 2019-07B). DIAL: Paris. Retrieved from <[https://www.iree.eu/publications/publications-in-iree/estimating-microcredit-impact-with-low-take-up-contamination-and-inconsistent-data-a-replication-study-of-crepon-devoto-duflo-and-pariente-american-economic-journal-applied-economics-2015/]>

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

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*, 4/2: 314–327.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). 'Scaling up what works: Experimental evidence on external validity in Kenyan education', *Center for Global Development Working Paper*, 321.

Boone, P., Eble, A., & Elbourne, D. (2013). *Risk and Evidence of Bias in Randomized Controlled Trials in Economics*. Centre for Economic Performance, LSE.

Bouquet, E., Wampfler, B., Ralison, É., & Roesch, M. (2007). 'Trajectoires de crédit et vulnérabilité des ménages ruraux: le cas des Cecam de Madagascar', *Autrepart*, 4: 157–172.
Brody, C., De Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P., & Dworkin, S. L. (2015). 'Economic self-help group programs for improving women's empowerment: a systematic review', *Campbell Systematic Reviews*, 11/1: 1–182.
Buera, F. J., Kaboski, J. P., & Shin, Y. (2015). 'Entrepreneurship and financial frictions: A macrodevelopment perspective', *economics*, 7/1: 409–436.
Bylander, M. (2014). 'Borrowing across borders: Migration and microcredit in rural Cambodia', *Development and Change*, 45/2: 284–307.
Cartwright, N. (2010). 'What are randomised controlled trials good for?', *Philosophical studies*, 147/1: 59.
Cartwright, N. (2007). 'Are RCTs the Gold Standard?', *BioSocieties*, 2(1): 11-20.
Cederlöf, G. (1997). *Bonds lost: Subordination, conflict and mobilisation in rural south India c. 1900-1970*. New-Delhi: Manohar.
Chayanov, A. V. (1966). *The theory of peasant economy*. Homewood, IL: Richard Irwin for the American Economic Association.
Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). *Generic machine learning inference on heterogenous treatment effects in randomized experiments*. National Bureau of Economic Research.
Cling, J.-P., Razafindrakoto, M., & Roubaud, F. (2003). *New international poverty reduction strategies*. London and New York: Routledge.
Collins, D., Morduch, J., Rutherford, S., & Ruthven, O. (2009). *Portfolios of the poor: how the world's poor live on $2 a day*. Princeton: Princeton University Press.
Copestake, J., Bhalotra, S., & Johnson, S. (2001). 'Assessing the Impact of Microcredit: A Zambian Case Study', *The Journal of Development Studies*, 37/4: 81–100. DOI: 10.1080/00220380412331322051
Copestake, J., Dawson, P., Fanning, J.-P., McKay, A., & Wright-Revolledo, K. (2005). 'Monitoring the diversity of the poverty outreach and impact of microfinance: A comparison of methods using data from Peru', *Development Policy Review*, 23/6: 703–723.
Copestake, J., Johnson, S., Cabello, M., Goodwin-Groen, R., Gravesteijn, R., Humberstone, J., Nino-Zarazua, M., et al. (2016). 'Towards a plural history of microfinance', *Canadian Journal of Development Studies / Revue canadienne d'études du développement*, 37/3: 279–97. DOI: 10.1080/02255189.2016.1197102
Crépon, B., Devoto, F., Duflo, E., & Parienté, W. (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: 123–50.
Crépon, B., Devoto, F., Duflo, E., & Pariente, W. (2019). '"Verifying the internal validity of a flagship RCT: A review of Crépon, Devoto, Duflo and Parienté": A rejoinder'. *Working Paper 2019-07A*. DIAL: Paris.
Cull, R., Ehrbeck, T., & Holle, N. (2014). *Financial Inclusion and Development: Recent Impact Evidence* (CGAP, focus note n°92.).
Dahal, M., & Fiala, N. (2018). *What do we know about the impact of microfinance? The problems of power and precision*. Ruhr Economic Papers.
Deaton, A. (2010). 'Instruments, Randomization and Learning about Development', *Journal of Economic Literature* 48(2): 424–55.
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.
Demirguc-Kunt, A., Klapper, L., & Singer, D. (2017). *Financial inclusion and inclusive growth: A review of recent empirical evidence*. The World Bank.
Desrosières, A. (2013). *Gouverner par les nombres: L'argument statistique II*. Presses des Mines via OpenEdition.

Doligez, F. (2002). ‘Microfinance et dynamiques économiques: quels effets après dix ans d’innovations financières?’, *Revue tiers monde*, 783–808.

Douglas, M., & Isherwood, B. (1980). *The world of goods*. Harmondsworth: Penguin Books.

Duvendack, M., & Palmer-Jones, R. (2012). ‘High noon for microfinance impact evaluations: re-investigating the evidence from Bangladesh’, *The Journal of Development Studies*, 48/12: 1864–1880.

Duvendack, M., Palmer-Jones, R., Copestake, J. G., Hooper, L., Loke, Y., & Rao, N. (2011). *What is the evidence of the impact of microfinance on the well-being of poor people?*. EPPI-Centre, Social Science Research Unit, Institute of Education.

Elyachar, J. (2006). *Markets of Dispossession: NGOs, Economic Development, and the State in Cairo*. Durham, NC: Duke University Press.

Ferguson, J. (2015). *Give a man a fish: Reflections on the new politics of distribution*. Durham and Duke: Duke University Press.

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

Goedecke, J., Guérin, I., D’espallier, B., & Venkatasubramanian, G. (2018). ‘Why do financial inclusion policies fail in mobilizing savings from the poor? Lessons from rural South India’, *Development Policy Review*, 36: O201–O219.

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

Gubert, F., & Roubaud, F. (2011). *The impact of microfinance loans on small informal enterprises in Madagascar: A panel data analysis*. World Bank.

Guérin, I., & Kumar, S. (2017). ‘Market, freedom and the illusions of microcredit. patronage, caste, class and patriarchy in Rural South India’, *The Journal of Development Studies*, 53/5: 741–754.

Guérin, I., Labie, M., & Servet, J.-M. (Eds). (2015). *The crises of microcredit*. London: Zed Book.

Guérin, I., Morvant-Roux, S., & Villarreal, M. (Eds). (2013). *Microfinance, debt and over-indebtedness: juggling with money*. London and New-York: Routledge.

Guérin, I., Roesch, M., Venkatasubramanian, G., & Kumar, S. (2013). ‘The social meaning of over-indebtedness and creditworthiness in the context of poor rural South Indian households (Tamil Nadu)’. Guérin I., Morvant-Roux S., & Villarreal M. (eds) *Microfinance, Debt and Over-Indebtedness. Juggling with Money*, pp. 125–150. Routledge: London and New-York.

Guérin, I., Venkatasubramanian, G., & Kumar, S. (2019). ‘Rethinking saving: Indian ceremonial gifts as relational and reproductive saving: Journal of Cultural Economy’:, *Journal of Cultural Economy*, 0/0. DOI: https://doi.org/10.1080/17530350.2019.1583594

Guyer, J. I. (1997). ‘Endowments and Assets: The Anthropology of Wealth and the Economics of Intrahousehold Allocation’. Haddad J., Hoddinott J., & Alderman H. (eds) *Intrahousehold Resource Allocation in Developing Countries*, pp. 112–29. The Johns Hopkins University Press: Baltimore.

Hardiman, D. (2000). *Feeding the Baniya: peasants and Usurers in Western India.* Oxford University Press.

Harrison, G. (2011). ‘Randomization and its Discontents’, *Journal of African Economies*, 20(4): 626–652.

Heckman, J.J. (1991;) ‘Randomization and Social Policy Evaluation’? NBER Technical Working Paper No. 107. Cambridge, MA: National Bureau of Economic Research.

Heckman, J. J., & Moktan, S. (2018). ‘Publishing and promotion in economics: The tyranny of the top five’. *NBER Working Paper No. 25093*, National Bureau of Economic Research. Cambridge, MA.

Hes, T., & Poledňáková, A. (2013). ‘Correction of the claim for microfinance market of 1.5 billion clients’, *International Letters of Social and Humanistic Sciences*, 2/1: 18–31.

Hoffmann, N. (2020). ‘Involuntary experiments in former colonies: The case for a moratorium’, *World development*. 127 104805.

Hummel, A. (2013). 'The Commercialization of Microcredits and Local Consumerism : Examples of Over-indebtedness from Indigenous Mexico'. Guérin I., Morvant-Roux S., & Villarreal M. (eds) *Microfinance, Debt and Over-indebtedness. Juggling with money*, pp. 253–71. Routledge: London.

Jamison, J.C. (2017.) 'The Entry of Randomized Assignment into Social Sciences'? Policy Research Working Paper No. 8062. Washington, DC: World Bank.

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

——. (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].

Javoy, E., & Rozas, D. (2013). 'Estimating levels of credit market saturation'. Guérin I., Morvant-Roux S., & Villarreal M. (eds) *Microfinance, Debt and Over-indebtedness. Juggling with money*. Routledge: London.

Jevons, W. S. (1883). *Methods of Social Reform*. London: Macmillan.

Johnson, S., & Rogaly, B. (1997). *Microfinance and Poverty Reduction*. London: Oxfam.

Joseph, N. (2013). 'Mortgaging Used Saree-skirts, Spear-heading Resistance: Narratives from the Microfinance Repayment Standoff in Ramanagaram, India, 2008-2010'. Guérin I., Morvant-Roux S., & Villarreal M. (eds) *Microfinance, Debt and Over-indebtedness. Juggling with money*, pp. 272–94. Routledge: London.

J-PAL, & IPA Policy Bulletin. (2015). *Where Credit Is Due*. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab and Innovations for Poverty Action.

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.

Karlan, D., Ratan, A., & Zinman, J. (2014). 'Savings by and for the Poor: A Research Review and Agenda', *Review of Income and Wealth*, 60: 36–78.

Karlan, D., & Zinman, J. (2009). 'Expanding credit access: Using randomized supply decisions to estimate the impacts', *The Review of Financial Studies*, 23/1: 433–464.

——. (2011). 'Microcredit in theory and practice: Using randomized credit scoring for impact evaluation', *Science*, 332/6035: 1278–1284.

Khandker, S. R., Samad, H. A., & Khan, Z. H. (1998). 'Income and employment effects of micro-credit programmes: Village-level evidence from Bangladesh', *The Journal of Development Studies*, 35/2: 96–124.

Kingi, H., Vilhuber, L., Herbert, S., & Stanchi, F. (2018). 'The Reproducibility of Economics Research: A Case Study'. Presented at the BITSS Annual Meeting, Berkeley.

Labrousse, A. (2017). 'Learning from Randomized Controlled Experiments. The Narrative of Scientificity, Practical Complications, Historical Experience'. *Books and Ideas*, Online: http://www.booksandideas.net/Learning-from-Randomized-Controlled-Experiments.html.

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.

Labrousse, A. (2017). 'Learning from Randomized Controlled Experiments. The Narrative of Scientificity, Practical Complications, Historical Experience'. *Books and Ideas*, Online: http://www.booksandideas.net/Learning-from-Randomized-Controlled-Experiments.html.

——. (2019). 'The rethorics of Poor Economics', in Bédécarrats F., Guérin I and F. Roubaud (Eds), *Randomized Control Trials in Development: A Critical Perspective*, Chapter 8, Oxford: Oxford University Press (forthcoming).

Lautier, B. (2004). *L'économie informelle dans le tiers-monde*. Paris: La Découverte.

Lont, H., & Hospes, O. (Eds). (2004). *Livelihood and Microfinance. Anthropological and Sociological Perspectives on Savings and Debt*. Delft: Eburon Academic Publishers.

Mahjabeen, R. (2008). 'Microfinancing in Bangladesh: Impact on households, consumption and welfare', *Journal of Policy Modeling*, 30/6: 1083–92.

Maurer, K., & Pytkowska, J. (2014). *Indebtedness of Microcredit Clients in Bosnia and Herzegovina*. Frankfurt: European Fund for Southeast Europe. Retrieved October 11, 2019, from <https://www.findevgateway.org/library/indebtedness-microcredit-clients-bosnia-and-herzegovina>

McKenzie, D. (2012). 'Beyond baseline and follow-up: The case for more T in experiments', *Journal of development Economics*, 99/2: 210–221.

Meager, R. (2019). 'Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments', *American Economic Journal: Applied Economics*, 11/1: 57–91.

Morduch, J. (1999). 'The Microfinance Promise', *Journal of Economic Literature*, 37/4: 1569–614. DOI: 10.1257/jel.37.4.1569.

——. (2020). 'Why RCTs failed to answer the biggest questions about microcredit impact'. *World development*, 127 104818.

Morvant-Roux, S. (2009). 'Accès au microcrédit et continuité des dynamiques d'endettement au Mexique: Combiner anthropologie économique et économétrie', *Revue Tiers Monde*, 1: 109–130.

Morvant-Roux, S. (Ed.). (2009). *Exclusion et liens financiers: microfinance pour l'agriculture des pays du Sud*. Paris: Economica.

——. (2013). 'International migration and over-indebtedness in rural Mexico'. Guérin I., Morvant-Roux S., & Villarreal M. (eds) *Microfinance, Debt and Over-Indebtedness: Juggling with Money*, pp. 170–92. Routledge: London and New-York.

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.

Morvant-Roux, S., & Roesch, M. (2015). 'The Social Credibility of Microcredit in Morocco after the Default Crisis'. Guérin I., Labie M., & Servet J.-M. (eds) *The Crises of Microcredit*, pp. 113–30. Zed Book: London.

Narotzky, S., & Besnier, N. (2014). 'Crisis, Value, and Hope: Rethinking the Economy: An Introduction to Supplement 9', *Current Anthropology*, 55/S9: S4–16. DOI: 10.1086/676327

Oakley, A. (2000). 'A Historical Perspective on the Use of Randomized Trials in Social Science Settings', *Crime & Delinquency* 46(3): 315–29.

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

Opem, L. C., & Goronja, N. (2013). *Responsible finance : reducing over-indebtedness for Bosnia and Herzegovina's microfinance borrowers*. Washington, D.C: International Finance Corporation.

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

Picherit, D. (2018). 'Rural youth and circulating labour in south India: The tortuous paths towards respect for Madigas', *Journal of Agrarian Change*, 18/1: 178–195.

Pitt, M. M., & Khandker, S. R. (1998). 'The impact of group-based credit programs on poor households in Bangladesh: Does the gender of participants matter?', *Journal of political economy*, 106/5: 958–996.

Prathap, V., & Khaitan, R. (2016). *When is microcredit unsuitable? Guidelines Using Primary Evidence from Low-Income Households in India* (IFF Working Paper Series No. WP-2016-01). Chennai: IFMR Finance Foundation.

Pritchett, L., & Sandefur, J. (2013). 'Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix', *Center for Global Development Working Paper*, 336.

——. (2015). 'Learning from experiments when context matters', *American Economic Review*, 105/5: 471–75.

Rao, V. (2001). 'Celebrations as social investments: Festival expenditures, unit price variation and social status in rural India', *Journal of Development Studies*, 38/1: 71–97.

Ravallion, M. (2009). 'Should the randomistas rule?', *The Economists' Voice*, 6/2.

——. (2019). ‘Should the Randomistas (Continue to) Rule?’. in Bédécarrats F., Guérin I and F. Roubaud (Eds), *Randomized Control Trials in Development: A Critical Perspective*, Chapter 1, Oxford: Oxford University Press (forthcoming).

Rodrik, D. (2008). ‘The New Development Economics: We Shall Experiment, but how Shall We Learn?’, Paper presented at the Brookings Development Conference, The Brookings Institution, Washington, DC (29–30 May).

Roodman, D., & Morduch, J. (2014). ‘The impact of microcredit on the poor in Bangladesh: Revisiting the evidence’, *Journal of Development Studies*, 50/4: 583–604.

Rutherford, S. (2000). *The Poor and their Money*. New Delhi ; New York: Oxford University Press.

Rozas, D. (2014). ‘Microfinance in Mexico: beyond the brink’’. European Microfinance Platform.

Sandefur, J. (2015). ‘The Final Word on Microcredit?’. *CGDev Blog*. Retrieved from <https://www.cgdev.org/blog/final-word-microcredit>

Schicks, J. (2013). ‘The Definition and Causes of Microfinance Over-Indebtedness: A Customer Protection Point of View’, *Oxford Development Studies*, 41/sup1: S95–116. DOI: 10.1080/13600818.2013.778237

Schicks, J., & Rosenberg, R. (2011). ‘Too Much Microcredit? A Survey of the Evidence on Over-Indebtedness’, *CGAP Occasional Paper*, 19.

Shipton, P. M. (2010). *Credit Between Cultures: Credit Between Cultures: Farmers, Financiers, and Misunderstanding in Africa*. Yale University Press.

Scott, J. C. (1977). *The moral economy of the peasant: Rebellion and subsistence in Southeast Asia*. Yale University Press.

Servet, J.-M. (2006). *Banquiers aux pieds nus*. Paris: Odile Jacob.

——. (2011). ‘La crise du microcrédit en Andhra Pradesh (Inde)’, *Revue Tiers Monde*, 3: 43–59.

Shaffer, P. (2015). ‘Two concepts of causation: Implications for poverty’, *Development and Change*, 46/1: 148–166.

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

Tarozzi, A., Desai, J., & Johnson, K. (2015). ‘The impacts of microcredit: Evidence from Ethiopia’, *American Economic Journal: Applied Economics*, 7/1: 54–89.

Taylor, M. (2011). ‘“Freedom from poverty is not for free”: rural development and the microfinance crisis in Andhra Pradesh, India’, *Journal of Agrarian Change*, 11/4: 484–504.

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

Woolcock, M. (2013). ‘Using case studies to explore the external validity of ‘complex’development interventions’, *Evaluation*, 19/3: 229–248.

Wydick, B. (2016). ‘Microfinance on the margin: why recent impact studies may understate average treatment effects’, *Journal of Development Effectiveness*, 8/2: 257–265.