DOCUMENT DE TRAVAIL DT/2019-07B

Verifying the internal validity of a flagship RCT: A review of Crépon, Devoto, Duflo and Pariente. Rebutting the Rebuttal

Florent BEDECARRATS

Isabelle GUERIN

Solène MORVANT-ROUX

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

Verifying the internal validity of a flagship RCT: A review of Crépon, Devoto, Duflo and Pariente. Rebutting the Rebuttal

Florent Bédécarrats Division des Evaluations de l'AFD , 75012 Paris, France bedecarratsf@afd.fr	Isabelle Guérin IRD, UMR CESSMA, 75013 Paris, France UMR CESSMA, 75013 Paris, France isabelle.guerin@ird.fr
Solène Morvant-Roux Graduate School of Social Sciences UNIGE-G3S, University of Geneva, Swisserland Solene.Morvant@unige.ch	François Roubaud IRD, UMR DIAL, 75010 Paris PSL, Université Paris-Dauphine, LEDa, UMR DIAL, 75016 Paris, France roubaud@dial.prd.fr

UMR DIAL Working Paper

July 2019

Abstract

We reply to CDDP's response to our replication of their published article in AEJ:AE. They reject most of the errors we documented in our replication paper. We provide a detailed answer to each objection they raise. We find that almost all of their rebuttals are driven by mistakes on their part. Once all these mistakes in CDDP's answer have been rectified, we find that all the coding, measurement and sampling errors documented in our replication still hold. All that remains then of the rejoinder is CDDP's argument that the issues we raised are not relevant because they do not substantially modify their impact estimates, and the use made by CDDP of additional sophisticated econometric tests to argue that their original results are robust. We disagree, as we find that correcting the rectifiable errors we identified does indeed show that the impact on assets and profits is not significant, and that the main results are to be found in increasing turnover from self-employment, which is trivial and generates very different conclusions to the original paper's findings. CDDP also omit to mention that the core conclusion of our replication was that, irrespective of the revised impact estimations, these results must be considered as lacking validity due to the massive inconsistencies found in the data, the substantial imbalances at baseline, the flaws in the experiment's integrity and the signs of probable contamination by other utility-related interventions. We are unable at this stage to assess the validity of the double post lasso procedure, the Benjamini-Hochberg False discovery rate correction of multiple testing, or the machine learning analysis put forward by CDDP, as they have not disclosed the related statistical scripts. Yet we fail to understand how even the most sophisticated methods could solve the "garbage in-garbage out" issue characteristic of this study. At this stage, we can only say that we have a very different notion of what underpins the internal validity of empirical research. We encourage CDDP to submit their answer to a peer-reviewed journal for a third-party appraisal of this debate.

Keywords: RCT, Microcredit, J-PAL, Replication, Morocco, Internal validity, Data quality

JEL Code : C18 C83 C93 G21 016 055

Introduction

Crépon, Devoto, Duflo and Pariente (hereafter referred to as CDDP) carried out a randomised control trial (RCT) in rural Morocco showing that microcredit had substantial, significant impacts on the assets, outputs, expenses and profits of self-employment activities (Crépon et al., 2015). We replicated this RCT, translating their analysis into a different statistical language (R) to fully analyse and reproduce every detail of the procedure they applied to the data. Although we were able to reproduce the exact same results, we demonstrated that these were due to a series of coding errors, measurement errors and sampling errors. We studied the sensitivity of CDDP's published results to these errors and published our analysis in a peer-reviewed journal, which is the only journal dedicated to replication studies in economics (Bédécarrats et al., 2019a). CDDP produced an answer, entitled "rejoinder", rejecting most of the errors we documented in our replication (Crépon et al., 2019: other references will soon be available as CDDP announced they would publish it also on their institutional websites). CDDP refer to our analysis, but they did not replicate or closely analyse its statistical content and their rejoinder contains numerous factual errors and omissions. We offered to detail these issues so that they could correct them before onlining their document, but they declined. We consequently publish here a review of CDDP's main arguments in response to our replication. We used the headings "Replication:" for the key arguments in our replication paper, "CDDP:" for the key arguments used by CDDP to rebut our arguments in their rejoinder; and "BGMR:" for our responses to CDDP's arguments.

Results rely primarily on the trimming procedure and threshold used

Replication: CDDP applied different trimming procedures. At baseline, they removed the highest values of 24 variables for 459 households (10.3% of the baseline sample). At endline, they completely removed 27 observations (0.5% of the baseline sample), which presented the maximum distance to a normalized distribution for 22 variables. Only the latter procedure was reported in CDDP's paper and they claim that no further trimming was performed on the data.

> CDDP: "*We do not trim from the analysis sample any household based on their baseline value*."

BGMR: This is wrong. As documented in the Replication's Table 5, observations were not fully removed based on the baseline data, but values were removed, which is the definition of trimming (Heckman and Leamer, 2009, p. 5443).

> CDDP: "*The assertion that we did not trim in the same way at baseline and endline is misleading. In the outcome regressions, observations are only trimmed on the basis of their endline value."*

BGMR: We did not say otherwise. We stated that CDDP included as controls in their outcome regression the variables reported as imbalanced at baseline. The regression CDDP use to test balance at baseline is based on heavily trimmed variables. CDDP would have reported large imbalances at baseline for other key variables if they had trimmed the baseline data in the same way as they trimmed the endline data.

Replication: Impact estimates vary substantially depending on the trimming threshold.

> CDDP: "*Effects are clearly decreasing for assets with trimming at 2, 3 and 5% and for profits at 2 and 3% (there is a small jump again at 5% for profits).* (…) *It is a logical implication of this result that the effect would become smaller and smaller and eventually vanish when trimming more and more of the data*."

BGMR: Of all the different thresholds we tested, CDDP only report on the farthest from the threshold they retained at endline (0.5%, i.e. removing 27 observations). They omit to mention that we presented in the same table results with thresholds much closer to theirs (0.3% and 0.7%).

> CDDP: "*It is certainly impossible to reject equality in the estimates BGMR report with various trimming thresholds and what we report (granted, the standard errors are large). The results look quite similar across all rows and columns except for zero percent trimming for profit*."

BGMR: No, they do not look quite similar. A 0.3% threshold (removing 16 observations instead of 27) cancels out the significant impact on profits and produces only marginally significant impacts on assets, outputs and expenses, and no impact on investments. The impact on profits is also non-significant at the 0.7% threshold (removing 38 observations instead of 27), but the impacts on assets, sales and expenses appear as highly significant, even though a marginally significant negative impact on investments defies logical interpretation as it directly contradicts the positive impact on assets. The only consistent results are to be found in increasing turnover from self-employment, which is trivial and generates different conclusions to CDDP's published findings.

> CDDP: "*It turns out that the result that the point estimate would be sensitive to removing large values was already in our paper (although, surprisingly, BGMR do not refer to it). We report quantile treatment effects (see Figure 1 of the original paper), and changes in the cumulative distribution of compliers (Figure 2). Both show that quantile treatment effects are large at the top of the distribution but zero below the 75th percentile.*"

BGMR: CDDP's quantile regression focused on the top 10% and 25% of the distribution, while we show that moving the threshold by just 0.2% (only a dozen observations) would have produced substantially different headline results. 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. Hence, CDDP's published results do rely primarily on the unusual trimming procedures they used.

Imbalances at baseline and impacts on implausible outcomes

Replication: We found substantial and significant imbalances at baseline for a number of important variables, including the outcome variables of this RCT.

> CDDP: "*We chose to include [in our balance checks] a parsimonious set of well measured variables to introduce at baseline that are representative of the main dimensions of the analysis (…) Table 6 of BGMR could appear to be cherry-picking the outcomes that are unbalanced, or on which there is an effect without a structured procedure for selecting outcomes.*"

BGMR: Testing balance for the main RCT outcomes is standard practice for RCTs (Bruhn and McKenzie, 2009). In their Toolkit for RCTs in development, Duflo et al. (2008) write, "*Information on covariates should therefore be collected in the baseline surveys. A special case of a covariate of interest is the pre-treatment value of the outcome. (…)*." We wonder whether the fact that CDDP chose variables closely related to their outcomes of interest, but not their outcomes of interest, is not cherry picking. The other variables (language, access to land, access to water and electricity, and migration) stem from a qualitative study conducted at the same time as the RCT to enrich CDDP's analysis, but which they chose not to take into account (Guérin et al., 2011; Morvant-Roux et al., 2014). We do not understand how CDDP could dispute the fact that these are essential variables considering the scope of their analysis.

CDDP: *"Once again, the text of BGMR is not consistent with their tables. In fact, the results in the two last columns of Table 6 are extremely similar to the results in Crépon et al. (2015). The effect on sales, for example [only increases by 8%]."*

BGMR: CDDP's argument is misleading as they only report the variation in the outcome variable displaying the smallest variation, i.e. sales. When controlling for imbalances at baseline, the estimate points increase by 47% for assets and by 24% for expenses. They drop by 24% for profits and the impact on profits no longer appears to be significant. This is not "*extremely similar".*

CDDP: *"Researchers (…) include covariates for two reasons. The first one is to get consistent estimates. In case there are imbalances on variables which can affect the outcome variable of interest, we want to control for them. The second reason is to increase the power of the experiment (the ability of the experiment to detect an effect when there is an effect). The main risk is, however, specification search. Why introduce this set of controls and not another one? Athey and Imbens (2017) recommend in general a simple treatment control comparison that does not introduce any control."*

BGMR: This is not what Athey and Imbens (2017) recommend. They recommend checking for imbalances in covariates for three reasons: i) to see if imbalances appeared by chance; ii) if the sampled population was different from the randomized population *"to assess how big the imbalances are that resulted from the sample selection"*; and iii) *"if there is some distance between the agencies carrying out the original randomization and the researcher analyzing the data [to] check on the validity of the randomization"*. Reasons ii) and iii) clearly apply here due to the problematic sampling used by CDDP (see below) and the numerous coordination issues during RCT implementation documented in Bédécarrats et al. (2019b). Athey and Imbens (2017) write, "*If the randomization was compromised, adjusting for covariate differences may remove biases*." They advise controlling only for categorical variables in experiment regression specifications, because other variables require fulfilling additional statistical assumptions that are difficult to verify. If large imbalances at baseline cannot be corrected with categorical variables, they recommend aborting the experiment and re-randomizing it.

Measurement and coding errors in treatment (credit) measures

Replication: There were substantial inconsistencies between survey and administrative data. For example, the 'client' variable CDDP used to instrument the regression presented in CDDP's Table 9 (impact of borrowing) identified 435 households as clients, yet 241 of these said they had not borrowed from Al Amana. Another 152 households self-reported having a loan from Al Amana, but were not listed as borrowers in Al Amana's records.

CDDP: "*There are several inaccuracies and misunderstandings in the BGMR statement quoted above. First, the dummy variable "client" is not used to instrument anything in our analysis. It is an endogenous variable which is instrumented by the random assignment variable, when we compute the Local Average Treatment Effects (LATE)*. "

BGMR: As CDDP quoted in their rejoinder, we wrote "*used to instrument the regression*" and not "*used as the instrument in the regression*", so there is no mistake in our original phrasing.

CDDP: "*Second, this variable has absolutely no incidence on the balancing checks at baseline. In those checks we simply compare the average characteristics between individuals in villages assigned to receive Al Amana's intervention and those who were not (the borrowing variable plays absolutely no role in the baseline checks)*."

BGMR: CDDP misread the sentence to which they refer. It does not relate to the variable client fetched from administrative data at endline, but to "*the inaccuracy in borrowers' identification*". This concerns CDDP's errors when accounting for loan access at baseline, as they omitted to account for loans from other MFIs than Al Amana or for outstanding loans in the previous 12 months that matured before the baseline survey (see below).

CDDP: "*Third, this variable does not play any role in the ITT (Tables 2 to 7 in Crépon et al., 2015). ITT estimates measure the impact of Al Amana's presence in the village. In the core of the Crépon et al. (2015) analysis, the variable "client" plays absolutely no role."*

BGMR: This variable does play a role in the ITT: it is the first ITT result they present (Table 2, top left corner). This is used by CDDP to show that the experiment did indeed result in substantial and significant take-up, which renders plausible the impact found on other outcomes. As with the previous paragraphs, this failure to reliably identify which households took credit is first and foremost an indicator of the survey's data quality, inconsistencies in household identification and possible flaws in the experiment's integrity.

Replication: The count of total borrowing at baseline omitted credits from MFIs other than Al Amana. This same count included solely outstanding loans at the time of the survey, instead of all the loans that had been outstanding in the previous 12 months, as specified in the variable's definition and as stated in the published paper and computed at endline. CDDP use the count of total borrowing at baseline as a control variable in their outcome regression at endline.

CDDP: "*The first statement is wrong. Loans from other MFIs are included in the balance table (Table 1 of Crépon et al., 2015). 'Loans from other formal institutions' include both loans from other MFIs and from formal institutions other than MFIs*."

BGMR: CDDP seem to misunderstand the error here. They refer to a variable 'aloans_oformal2' that includes loans from other MFIs. But it is another variable, named 'aloans_oformal' (without a '2' at the end), which omits loans from other MFIs, that they include in the count of total borrowing at baseline ('borrowed_total_bl') used as a control variable for their outcome regressions. So CDDP do indeed omit loans from other MFIs when controlling for access to credit at baseline in their outcome regression.

CDDP: "*The second statement is correct: there was indeed an error in the construction of the variable 'had an outstanding loan'. Loans from other MFIs are indeed not taken into account. We revise this variable in Table 5 of this document. The percentage of control group households that have an outstanding loan from any source at baseline is now at 26.8% instead of the original average of 25.7%. The balance between the treatment and the control group is not affected, as shown in Table5. Obviously, this will make no difference*."

BGMR: On the contrary, it makes significant amount of difference, as shown in Replication Table 11. CDDP reported 6.0% access to formal credit other than Al Amana in the control group at baseline. The corrected level is 7.2% access to other formal credit sources (+20%), compared to 10.1% in the treatment group, with a significant difference at the 1% level. CDDP reported 6.8% access to informal

credit at baseline for the control group. The corrected level is 8.5% (+25%), compared to 10.4% in the treatment group, also with a significant difference at the 1% confidence level.

> CDDP: [BGRM's statement that total access to credit at baseline is used by CDDP as control variables, so correcting this variable modifies the measured impact results] "*does not make sense: It is important to note, as already mentioned, that the reduced-form specification of Crépon et al. (2015) does not include baseline levels of the dependent variables*."

BGMR: The reduced form specification of Crépon et al. (2015) does include as an independent variable total access to credit at baseline (named 'borrowed_total_bl'). As re-established above, CDDP omitted to include in this variable loans that had matured in the previous 12 months and they also omitted loans from MFIs other than Al Amana. Hence, correcting the independent variable used by CDDP in their reduced form specification does modify the impact estimates.

> CDDP: "*Finally, BGMR also affirm that changing the measure of the baseline covariate on access to credit significantly affects estimated effects (…) This statement is wrong and comes from a coding error in BGMR: the difference in effect does not come from the control variable but from using a different sample*."

BGMR: This is a mistake on the part of CDDP, who did not reproduce our results. As specified in our replication and unambiguously computed in our code (lines 2840 to 2883),[1] the corrected results quoted by CDDP were obtained using the exact same specifications as CDDP, correcting only the baseline level of credit access to include loans that matured in the previous 12 months and loans from other MFIs. There is no different specification and the corrected result to which CDDP refer (see replication section 5.1.3) used the exact same sample as CDDP used for their regression, which are the exact same 4,934 households reported by CDDP in their tables 2 to 7.

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

> CDDP: "*We agree with the BGMR claim that 'other loans' at endline have been aggregated with loans from a utility company (so this variable should have been labelled 'utility and other credit', not just 'utility') (...). This should not affect the results, and as we show below, it does not*."

BGMR: The substantial recoding of an outcome variable does affect the result of the regression on this outcome. As explained in the (longer) working paper version of our replication, rectifying the faulty utility credit reclassification "also alters the computing of the average treatment effect on access to utility credit at endline. This was estimated as 0.017 (0.017) in CDDP's Table 2, which is small and insignificant. Conversely, when preventing unjustified reclassification, it becomes 0.037** (0.016), which is larger and significant," (Bédécarrats et al., 2018, p. 19). CDDP do not provide any illustration that correcting the faulty recodification pointed out in our replication does not affect the results.

[1] Full replication code is available at http://dx.doi.org/10.15456/iree.2019071.090421.

CDDP: "*However, the classification that they operate in the baseline is incorrect." There was no answer option in the questionnaire for 'utility credit', but only 'other credit' and the complementary information is missing for most answers and it would be inappropriate to recode all credits without this information as 'other'. The only option we consider technically correct at baseline is to define a unique variable that aggregates both utility company loans and other loans*."

BGMR: We disagree. Regrouping these sources as an undefined category would have been less incorrect than labelling all loans as 'utility credit'. This is why, when reanalysing credit access after correcting CDDP's faulty recoding, we renamed this 'other' modality as 'None of the above or not specified', which corresponds to the information available in the data. However, we show that 3.9% of control group respondents at baseline specified that the 'other' source of credit they had access to was utility credit, compared to 6.1% in the treatment group. This indicates an imbalance at baseline that might have contaminated the experiment's results.

CDDP: "*Most importantly, the claim that reclassifying utility loans into utility and other loans at endline changes the effect of microcredit access on the probability to have up-taken a utility loan is entirely incorrect, and comes from the same error in the code that we discussed previously.* (…) *This claim (on which the authors insist so much that they mention it in the subtitle of the paper they published) once again comes from an analysis performed on a restricted sample comprised only of households surveyed both at baseline and at endline. The effect of 0.037 on utility loans is obtained on this restricted sample of 3,525 households (see column 3 of Panel B, Table 7).*"

BGMR: With respect to the incidence of the faulty credit accounting at baseline, CDDP have not carefully analysed our code and have assumed that these corrected results were obtained using a different sample to the one they used in their Table 2. This is not the case. The results reported here were produced using the same sample as CDDP, with the same 4,934 households on which they ran their regression reported on Table 2, only correcting for miscoded credit access.

CDDP: "*Their claim indicates 'contamination' in the study. As an aside, this is a puzzling comment. Even if it were true, why could microcredit not have a causal effect on home improvement and hence potentially on utility loans? Many evaluations find that the first order impact of a cash transfer is to buy a roof or improve the home. In Morocco we know that people value access to water enormously and are ready to borrow for it (Devoto et al., 2012). But as it turns out, this result is incorrect, and comes from the same error as above.*"

BGMR: Once again, CDDP are mistaken. The fact is that utility credit is the main source of access to credit in the experiment's villages, far ahead of microcredit. It is also the source of credit that varies the most between baseline and endline (way ahead of microcredit). Access to utility credit appears as significantly imbalanced at both baseline and endline. Access to water, sanitation and electricity is also heavily imbalanced at baseline, and varies significantly differently between baseline and endline. As we mentioned in our text, "*It is unclear whether this significant increase in access to utility credit in treatment villages is an unexpected impact of increased Al Amana credit or contamination by a co-intervention. In any case, further analysis would be required to disentangle the impact of microcredit and the impact of utility credit in this context.*"

Measurement and coding errors in outcome measures and controls

Replication: 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 from 1,377 Moroccan Dirham to 5,111 Moroccan Dirham.

> CDDP: [CDDP confirm that they appraised assets, sales and consumption using a pricing method that produced unreliable prices. They also confirm that they deliberately removed tractors and reapers at endline because the prices estimated for these assets were clearly irrelevant]: "*The information we have is so poor and the unit price so large that there is a substantial risk of introducing more noise than anything else in the regression and thus limits our ability to detect an impact if there is one.(...) In other words, BGMR introduce a huge amount of noise in the estimation by adding information coming from a limited number of households, and they obtain a similar point estimates with an enormous confidence interval.*" [CDDP then build classes of asset items, apply a Benjamini-Hochberg False discovery rate correction of multiple testing and test for the assumption of joint nullity of impacts on all the coefficients. They conclude that having included tractors and reapers would have delivered the same result and that these assets only increase coefficients and standard errors.]

BGMR: These assets were supposed to be included in the assets count according to CDDP's published article. CDDP should have disclosed their removal at endline in their published article, although this would have raised concerns among reviewers and readers, as agricultural mechanization is a crucial component of rural development (Pingali et al., 1988). CDDP argue that the items they removed have no meaningful effect on the results, except to add noise to genuine impacts. This contradicts their rejoinder where, rejecting the inconsistencies found in their trimming method, they write, "*In many cases, we are interested in "outliers", and we may be much more interested in how a distribution is affected than in the average effect of a distribution*". Bear also in mind that we documented 14 coding errors in CDDP's do-file that affect 3,866 of the 4,934 observations (78.35%) used by CDDP for their ATE estimation of self-employment activities. These errors interact to alter the results. It is not enough to assess the incidence of each standalone error. It is their combined impact that needs to be assessed. This is precisely what we did in our replication (Table 13) and, in their rejoinder, CDDP do not propose any results simultaneously incorporating corrections to the different errors documented in our replication.

> CDDP: "*BGMR rerun the regressions of Crépon et al. (2015), also including other controls in their Table 13*."

BGMR: This is not so. As specified in the Table 13 footnote, our corrected regressions include the same control variables as CDDP. This can be checked in our code (lines 3661 to 3679) [1].

[1] Full replication code is available at http://dx.doi.org/10.15456/iree.2019071.090421.

Sampling errors

Replication: Households were sampled based on their answers to a short preparatory survey. But 50.5% of the households surveyed at baseline displayed considerable differences compared to the data collected by the preparatory survey for the exact same variables.

> CDDP: "*The preparatory survey included a single question where the total number of members of the household was asked, while the baseline survey included a whole module where each household member was listed and the condition of residence of each member was verified. It is thus not surprising that the two pieces of information would differ*."

BGMR: As specified in the replication text and in our Table 14, we did not take into account small variations in the number of household members between the preparatory survey and baseline. We only considered differences where the number of members varied by more than 30% and at least two people. 14% of the households included at baseline presented these large inconsistencies in size compared to the preparatory survey. In addition, we also examined whether the information was consistent between the preparatory and baseline surveys for three dummy variables used by CDDP as criteria to include households in the sample: land ownership, tree ownership and receiving a pension. 24% of the households surveyed at baseline reported contradictory information for each and every one of these three criteria. Either the data collected in the surveys is unreliable, or these households are not the same.

Replication: The sex and age composition of 20.5% of the households interviewed at baseline and supposedly re-interviewed at endline differs to such an extent that it is not plausible that the same units were re-interviewed in these cases.

> CDDP: "*There could nevertheless be true changes in household composition, or different reporting. That is certainly standard in every panel data collection (RCT or not)*."

BGMR: Here again, we adopted an extremely lenient criterion to assess the compatibility in household sex and age composition between baseline and endline. We also provided the first records that did not match in our replication's Appendix 4, and we challenge anyone to come up with plausible narratives to show that these households are the same.

Replication: The borrowing propensity score used as the sampling criterion at baseline totally fails to predict borrowing.

> CDDP: "*So, the household borrowing score computed ex-ante is not the 'cornerstone of our identification strategy'. In fact, it is not used at any point in the analysis! It was just used to construct sampling probabilities. The source of identification is the randomization of villages. The IV estimates only use the 'treatment' dummy as instrument for the 'client' variable, and are just rescaling the results in Tables 2-7*."

BGMR: This is misquoted. What we say is, "*The cornerstone of this RCT protocol and the corresponding article's identification strategy is the household propensity to borrow, which was evaluated by scores*." What we documented is that CDDP used four different scores to assess propensity to borrow and that these scores were totally inconsistent with one another and that they failed to assess propensity to borrow. Propensity to borrow is indeed the cornerstone of CDDP's analysis: they titled their paper in reference to it and they mention it as a unique feature of the paper.

Replication: The borrowing propensity score used as the sampling criterion at baseline is at odds with the revised borrowing propensity scores used as sampling criteria to add new households at endline.

> CDDP: *We do not have "high" and "very high" propensity groups in our analysis but simply a group of households who are more likely to borrow, which is sampled in exactly the same way in treatment and in control villages*.

BGMR: This is not true. CDDP themselves state in their original paper that they selected at baseline households with a high propensity to borrow and that, "*At endline, we added a third group that had an even higher propensity to borrow, by reestimating the take-up equation in the whole sample, and using the initial census (available for all households) to construct a new score,*" (p.126). Note that there is a contradiction here, because these households deemed to have a higher propensity at endline had initially been classified as low propensity at baseline. We show in our replication that the score used to select high propensity households totally failed to predict borrowing. The scores used to add households with a higher propensity at endline were only loosely correlated with actual borrowing and were orthogonal to the score used at baseline.

> CDDP: "*Finally, what BGMR do with the boxplot is very difficult to understand. The bottom line appears to be that the second and third scores do not select the same households as the first score would have (in other words, the people who we classified as likely to borrow with the second score would not have been selected as a likely borrower in the first score). BGMR seem to have re-discovered a fact that we were very aware of and we cite repeatedly in the paper: predicting ex-ante who will borrow is very difficult.*"

BGMR: Yes, CDDP understood the boxplot: the score used to select households with high borrowing propensity at baseline is orthogonal to the two scores used to select households with a higher propensity at endline, and the latter are only loosely correlated with one another.

Replication: CDDP used an astonishingly complex method to weight their survey observations to estimate microcredit externalities and compute instrumental variable estimates. They claim that these weights reflect households' inverse sampling probability, but this is not the case. Sampling probability was determined by the subsequent, contradictory propensity scores computed by CDDP and the weighting procedure does not capture it.

> CDDP: [Referring to the description of the weighting method in BGRM's replication] *"This is another inaccurate statement. We winsorize, we do not censor. People with a sampling probability below 0.1 receive a weight of 10 and not 0."*

BGMR: As CDDP comment in their own code, what they do is censoring. The specific type of censoring procedure they apply is not winsorizing. Winsorizing consists in replacing an extreme value by the next most extreme value (Yale and Forsythe, 1976). The censoring procedure used by CDDP is truncating, that is replacing all the values exceeding an arbitrary threshold with the value of this arbitrary threshold. See Dixon (1960) for the definition of the different censoring procedures. As CDDP rightly rectify, the replacement value is indeed 10 and not 0. Note that the aspect CDDP refute in their answer is a mention we made in the description of their censoring procedure, but it is not the issue we called into question in this section. What we discuss in this section is that the weights do not reflect the household selection probabilities as claimed in CDDP's original study. CDDP assert that, "*Our procedure leads to weighted results which are representative of village-level impacts at the level of a village,*" but do not provide any argument to refute our argument that this is not the case.

External validity

Replication: Our re-analysis focused on this experiment's lack of internal validity, but a number of the identified issues also raise concerns about its external validity. The average number of household members grew from 5.17 to 6.13 between the baseline and endline surveys. According to the national census, Moroccan rural households had an average of 6.03 members in 2004 and 5.35 members in 2014, displaying a decreasing pattern contrary to the experiment's observations. Compared to the Moroccan rural population, the study sample also has significantly fewer households headed by women and its measured consumption is 59% lower on average.

> CDDP: "*We are perfectly aware that consumption estimates on our sample may be different than the ones from a representative sample of the population in rural Morocco. We selected villages in remote rural areas at the periphery of Al Amana's branches and it is quite natural that households' characteristics in those villages differ from the characteristics of households living close to the branch and usually near the center of the rural district. We have never claimed that our sample is representative of rural areas of Morocco*."

BGMR: This is worth stressing, because the synthesis paper of this special issue claims that, "*All told, the six settings represented in this volume strike us as fairly representative of the distribution of lenders, loan terms, borrowers, and markets that comprise the microcredit world,*" (Banerjee et al., 2015, p. 7). However, in their rejoinder, CDDP omit to provide an answer to the most troubling discrepancy highlighted in this respect in our replication, which is household size growth that exhibits demographic features at odds with general trends observed in this region. This demonstrates that the definition of the target population changed across survey waves, which jeopardizes any statistical inference made from this population (Biemer and Lyberg, 2003, p. 6).

Conclusion

CDDP did not attempt to reproduce our replication results. They instead copy-pasted some fragments and completed them with what they assumed were the origins of the discrepancies highlighted in our published paper. This led them to make the number of factual mistakes we document above. We show in detail that all the coding, measurement and sampling errors documented in our replication still hold. In addition to their objections that we rebut, CDDP referred to a series of sophisticated analyses to argue that their original results are robust: double post lasso procedure, Benjamini-Hochberg False discovery rate correction of multiple testing, the Bayesian hierarchical model and machine learning analysis, among others. We are unable to assess the validity of these additional tests, as CDDP have not provided the related statistical scripts. Our assessment shows that we are facing a typical garbage in-garbage out principle, which states that no statistical procedure will yield reliable results if the data used for them is unreliable and if the survey practices used do not enable the magnitude of such inaccuracies to be assessed. At this stage, we can only acknowledge that we have a very different notion of what underpins the internal validity of empirical research.

We encourage them to rerun our replication to gain an accurate understanding of the errors we documented and to submit their answer to a scientific journal for a peer review of its reliability. It would be useful if CDDP could also disclose the statistical script they used to obtain their results so that these too can be checked (note that the data, code and results of our replication are available on the IREE website). A review by impartial third parties qualified not only in econometrics, but also

in statistics and rural economics would doubtless return an outside opinion and settle this debate. An interesting possibility for just such a review would be the International Journal for Re-Views in Empirical Economics (IREE), in which we published our replication, since it is the only journal dedicated to replication studies in economics. Such a peer review would determine whether, despite the numerous errors and misleading re-interpretations contained in their rejoinder, their arguments do actually disprove the arguments raised in our replication study. If, as CDDP claim, our replication is ultimately assessed as "*non-scientific*" by a qualified, independent panel, then we would be happy to accept that we are wrong and ask IREE to retract our replication article. Yet if we are proved right, which we are even more convinced of following CDDP's response, the logical upshot would be for AEJ-AE to retract the original CDDP paper. This verification would also be useful to assess the feasibility of continuing this same experiment in Morocco, as the conclusion of the initial CDDP article announced that a new survey wave would be carried out on the same RCT sample in the future.

References

Athey, S., Imbens, G.W., 2017. The econometrics of randomized experiments, in: Handbook of Economic Field Experiments. Elsevier, pp. 73–140.

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

Bédécarrats, F., Guérin, I., Morvant-Roux, S., Roubaud, F., 2019a. Estimating microcredit impact with low take-up, contamination and inconsistent data. A replication study of Crépon, Devoto, Duflo, and Pariente (American Economic Journal: Applied Economics, 2015). International Journal for Re-Views in Empirical Economics 3. https://doi.org/10.18718/81781.12

Bédécarrats, F., Guérin, I., Morvant-Roux, S., Roubaud, F., 2019b. Lies, damned lies, and RCT: A J-PAL RCT on rural microcredit in Morocco, Working Paper#2019-04. DIAL, Paris.

Bédécarrats, F., Guérin, I., Morvant-Roux, S., Roubaud, F., 2018. Estimating microcredit impact with low take-up, contamination and inconsistent data. A replication study of Crépon, Devoto, Duflo, and Pariente (American Economic Journal: Applied Economics, 2015), Working Paper#2018-12. DIAL, Paris.

Biemer, P.P., Lyberg, L.E., 2003. Introduction to Survey Quality. John Wiley & Sons.

Bruhn, M., McKenzie, D., 2009. In pursuit of balance: Randomization in practice in development field experiments. American economic journal: applied economics 1, 200–232.

Crépon, B., Devoto, F., Duflo, E., Parienté, W., 2019. “Verifying the internal validity of a flagship RCT: A review of Crépon, Devoto, Duglo and Parienté: a rejoinder", Working Paper#2019-07. DIAL, Paris.

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, 123–50.

Dixon, W.J., 1960. Simplified Estimation from Censored Normal Samples. The Annals of Mathematical Statistics 31, 385–391.

Duflo, E., Glennerster, R., Kremer, M., 2008. Using Randomization in Development Economics Research: A Toolkit, in: Schultz, P., Strauss, J. (Eds.), Handbook of Development Economics. North Holland, Amsterdam and New York.

Guérin, I., Moisseron, J.-Y., Roesch, M., Ould-Ahmed, P., Morvant-Roux, S., 2011. Analysis of the Determinants of the Demand for Financial Services in Rural Morocco, Ex Post - Analyses d’impact. AFD.

Heckman, J.J., Leamer, E., 2009. Handbook of Econometrics. Elsevier.

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. https://doi.org/10.1016/j.worlddev.2013.03.002

Pingali, P., Bigot, Y., Binswanger, H.P., 1988. Agricultural mechanization and the evolution of farming systems in sub-Saharan Africa (No. 10219). The World Bank.

Yale, C., Forsythe, A.B., 1976. Winsorized Regression. Technometrics 18, 291–300. https://doi.org/10.2307/1268738