"Verifying the internal validity of a flagship RCT: A review of Crépon, Devoto, Duflo and Parienté": A rejoinder[1]

Bruno Crépon, Florencia Devoto, Esther Duflo and William Parienté

June 2019

In a recent paper, Bédecarrats, Guerin, Morvan-Roux and Roubaud (2019) re-analyze the data from a randomized controlled trial of the impact of the program of Al Amana, a microcredit organization in Morocco, which we published in 2015 (Crépon, Devoto, Duflo, and Parienté, 2015). They make a number of strong claims about the validity of our approach, and the lack of robustness of our results to alternative assumptions. In this paper, we argue that their own quantitative results in fact demonstrate the robustness of our initial analysis. We question several of their suggested modifications to our analysis. Finally, in the spirit of serious scholarship, we provide our own candid re-analysis of our data, with more robust methods. Overall, our careful analysis of their paper, their code, and our own data suggest that the results we presented in our 2015 paper are in fact quite robust, contrary to what is claimed in their paper. In particular, we find strong support for large impacts on top quantiles of business profits, assets, and sales, no effect at the lower quantiles, a result that is robust to the method used for inference. Due to fat upper tail of the distribution of profit, the average treatment effect on profit is noisier than standard inference methods imply. Impacts on average treatment effects for the other business variables remain significant no matter the method used for inference. We continue to find off-settings effect on labor supply and no impact on average per capita consumption.

[1] A version of the working paper we respond to here (Bédécarrats et al., 2018) was published in *International Journal for Reviews in Empirical Economics* with a different name: Bédécarrats F., Guérin I., Morvant-Roux S., and Roubaud F. (2019). "Estimating microcredit impact with low take-up, high contamination and inconsistent data. A review of Crépon, Devoto, Duflo and Pariente (American Economic Journal: Applied Economics, 2015)", *International Journal for Reviews in Empirical Economics.* We did not have access to this paper until Mid-March, so this comment is based on the working paper.

A. Introduction

In a recent paper, Bédecarrats, Guerin, Morvan-Roux and Roubaud (2019) [henceforth, BGMR] re-analyze the data from a randomized controlled trial of the impact of the program of Al Amana, a microcredit organization in Morocco, which we published in 2015 (Crépon, Devoto, Duflo, and Parienté, 2015). Their work starts from our raw data files and codes, which are publicly available.

We are delighted that the data we posted is being used. Data and codes from experiments can be used to identify errors, check robustness to different plausible assumptions, and go deeper in the analysis. Our data has been downloaded 881 times, and it is used, for example, in Meager (2018, 2019), Giordano et al. (2016) and Chernozhukov et al. (2019).

We are also generally sympathetic with the kind of effort that BGMR undertake: as they point out, although conceptually straightforward, in practice, RCTs are complex projects undertaken over many years, and they involve many decisions and many lines of code. Errors are certainly possible. The editorial process at journals is not set up to detect them: the goal of this process is to judge the interest and validity of the method and the project as presented, not to find errors or question every single decision. This is why J-PAL has started offering a "replication service" for its researchers in which a graduate student attempts to replicate a complete paper from scratch, and can identify any error, omission, or questionable assumption. Given that the data has been used and re-analyzed repeatedly (Rachael Meager, for example, entirely reconstructed the analysis in each of her papers), we were a bit surprised by the claim of the paper that our entire analysis is flawed, but this of course could not be ruled out a priori. BGMR contacted us at the beginning of their project, and we welcomed their effort at that time.

We are genuinely impressed with the effort that BGMR put into their project. They started by fully replicating our code in a different coding language, looked for every possible error and omission, and probed the robustness of the results to different control variables and different samples.

We are, unfortunately, much less impressed by the final document they produced.

First and foremost, the text of the paper is entirely at odds with the evidence it presents. The text, abstract, introduction and conclusion make it sound like BGMR found very different results when re-analyzing our data and using different strategies. But the statement that the results are different from what we find directly contradicts their own findings: despite their extensive re-working of the data, BGMR's key findings are essentially unchanged from what we present. What they show is in fact that the original results are robust to different variable definitions, different control variables and different samples. This is evidenced in Figure 1, which compares BGMR's various estimates for the key variables with what is reported in the original paper, and shows how similar they are. In particular, the specifications in the purple bars, which use their favorite samples and control variables are virtually identical to our original point estimates. Occasionally, they even admit it in their text, but it is not the impression that someone would have if they read the paper rapidly.

Second, the only change that makes a difference in BGMR is trimming: when they trim a large number of households (3% to 5% of the households depending on the variables) some of the results on profits, sales and assets become small and insignificant. This, in fact, is directly consistent with the evidence we present on Quantile Treatment Effects, which shows that any effect on the averages for business variables are driven from the higher quantiles. This is explicitly discussed in the paper: we even find negative QTE at the bottom for profits. This is also consistent with all of the other studies on microfinance, and it is studied in detail in Meager (2018). We explicitly mention in the paper that the average effects on business profits come from the higher quantiles. While we are quite sure that the results they find after trimming are correct, we do not see these results as a limitation of the data or our strategy[2]. With variables where the action is in the tail, trimming a large number of observations does not make sense. One needs to use techniques that are not sensitive to outliers (like QTE) which is precisely what our original paper did.

Third, and related, the BGMR re-analysis unfortunately contains a large number of errors and misunderstandings. Many are conceptual misunderstandings of what RCTs are and how to analyze them. The trimming question is one of those. In another example, BGMR insists on the fact that the data should be analyzed with a "consistent sample" (with baseline and endline observations and stable family composition, if we understand correctly). This makes no sense for an RCT, especially one where 1,400 new observations were added at endline to increase power. In several places, BGMR appear to be confused about the notions of Intention to treat (ITT) and Treatment on the Treated (TOT). There are also several coding errors that invalidate some of their headline results. For example, their claim that our experiment has been "contaminated" by utility loans results from a mistake in their code (and a misinterpretation of this incorrect result). Finally, there appears to be some confusion between changed estimates and changed standard errors: BGMR often describe a change due to increased standard errors (e.g. from cutting the sample by 30%…) as if the point estimate changed. Cutting the sample would mechanically increase standard errors. If the point estimates are unchanged, it is not a sign of a lack of robustness of the original finding. It also contains many proposals for analyzing the data in different ways, which we do not consider neither vastly superior nor vastly inferior (although we generally prefer ours). The key point here is that we should be worried if the estimates were different under these assumptions, but they really are not.

Fourth, BGMR mostly perform ad-hoc re-analysis of our data, questioning some of our choices with little justification, and trying out other specifications instead. Somewhat confusingly, they don't systematically report the results of the robustness on the same set of core variables that are presented in Crépon et al. (2015). Only the business results are systematically presented, and for the rest they come up with their own choice of variables. While the work is extensive, it is thus not particularly systematic.

[2] The procedure we describe in the paper trims a very small number of households (0.5%) that seem to exhibit highly implausible values over a range of variables, and therefore seem to have compromised data. BGMR finds that trimming no households at all also affects the estimate of profits, though none of the other variables are affected. We still think those households should have been removed, but in any case, the most interesting aspect of the profit results is the quantiles, which is robust to outliers.

BGMR lost an occasion to perform a serious new look at the data using new methods that were not available or not widely used when we published the original paper. Some of these methods would have better addressed the issues they are concerned about (such as the choice of control variables, and fat tails for some of the variables that make the OLS specification inappropriate). We take advantage of going back to the data to perform these additional analyses on the core variables: in particular, we choose control variables with double LASSO, and we perform tests of treatment effects using randomization inference.

Our conclusions remain largely unchanged, with the one exception that the randomization inference test rejects the sharp null for average profits, which was only marginally significant in the original paper (10%). This is not particularly surprising, given that this variable exhibits a lot of kurtosis, and thus the standard tests we were using in the original paper don't have great properties in a small sample.

Meager (2018) re-analyses our data with more robust techniques to estimate QTE. She finds positive and significant treatment effects at quantiles 55, 65, 75, and 85 with no pooling (raw data). With partial pooling (including information on the other sites) she still finds significant treatment effects at quantiles 65, 75, 85 and, like us, a negative and significant treatment effect at quantile 5.

Overall, with this caveat on the average profit results, our careful reading of the document and our re-analysis of the re-analysis only reinforce our confidence in the original conclusions of the paper. Our sense is that it should also have persuaded BGMR of the same thing if they were not so intent on finding something different.

Overall, BGMR is a somewhat disheartening paper to read. The tone is polemical and often questions our motives or seriousness. The combination of misunderstanding, false assertions, and specification searching, culminating in the blatant mischaracterization of the paper's own results, detracts from the authors' own impressive amount of work. We hope that future re-analyses of this paper, or other papers, follow a more scientific and systematic approach, aimed to truly probe a paper's robustness, rather than to make some point no matter what the evidence says.

The remainder of this paper proceeds as follows. In section B, we summarize the key analyses of BGMR and their findings. We also summarize our objections to some of their criticisms of our approach. Finally, we briefly present our key results using "state of the art" methodology. This executive summary should be sufficient for a reader not interested in the details of each argument. In section C, we provide a point by point rebuttal of the arguments made by BGMR and discuss our additional analyses in more detail.

B. Executive summary

BGMR is very long and makes a large number of points. We took care to understand and probe the entire code, so our own response is also long and detailed. To help the reader sort through the main points, we provide an executive summary with our main takeaways in this section.

B.1 BGMR's re-analysis in fact demonstrates the robustness of our results to the variations they implement: their various corrected estimates are very similar to the original estimates.

BGMR suggest a large number of changes to our analysis. They proudly report that their corrections to the coding of the variables affect "3,866 of the 4,934 observations (78.35%) used by CDDP (Table 3)". They also suggest using a different sample, using a different set of control variables, and trimming differently. As we will explain below, we stand by our original analysis, and disagree with essentially all of these modifications. Nevertheless, exploring the robustness to these changes is interesting: it shows that no single specification choice, observation, or sample selection drives the results we reported.

Figure 1 shows the reduced form estimates of the "treatment village" dummies for the self-employment variables (the only ones that are systematically presented in BGMR and also present in our paper). We show the estimates in our original paper (first bar of each chart) and the estimates in the analyses BGMR present in Tables 5-18.[3] What is striking is how similar the results are. They are certainly all statistically indistinguishable from each other. Given the standard errors, this is perhaps not such an interesting result. More importantly, with the exception of the average effect on profits with no trimming at all (which goes to zero) and in BGMR Table 13 (which is twice the size), even the point estimates are quite similar. The results on average profits were only significant at the 10% level, and as we show in the paper and highlight in the abstract, very heterogeneous.

We are not sure what BGMR consider their headline corrected results, but assuming those are the estimates with revised sample and control variables (presented in purple in Figure 1), they are virtually identical to ours.

B.2 BGMR's headline conclusion that "trimming matters" for the business variables is a re-statement of one of the key results of the paper: the effect we observe on business variables are driven by the higher quantiles, and there are no effects (or possibly even negative effects for some variables) at lower quantiles

Figure 2A reproduces the results of Table 5 of BGMR for assets and profits (the same pattern can be found for business expenses and sales): the more observations that are removed, the lower the effects tend to be on the rest of the sample. 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).

[3] We omit from these graphs the analysis which trim 2%-5% of the households, which we discuss below.

Figure 2B, reproduced from our paper, shows why this is the case. We show Quantile Treatment Effects (the difference in the quantile of the distribution between treatment and control). The QTE are sharply increasing with the quantiles. For profits, they are even negative at lower quantiles. 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.

Figure 2C reproduces results from Meager (2018) for profits only. Meager re-calculated QTE for several quantiles in our data set, and then she re-computed QTE informed by the results of other randomized evaluations ("partial pooling"). The high point estimates on the tail of the distribution of profits are still there with partial pooling, but the 95% uncertainty interval includes zero. This is the main finding of Meager's paper: There is so much kurtosis in the profit variable that it is very hard to come up with generalizable conclusions.

Given the BGMR quote of Deaton and Cartwright, that *"Trimming of outliers would fix the statistical problem, but only at the price of destroying the economic problem; for example, in healthcare, it is precisely the few outliers that make or break a programme",* we find it completely baffling that the paper highlights the sensitivity to trimming as a particularly big problem.

B.3 Most of BGMR's proposed "corrections" or objections to our analysis are incorrect. There is a large number of conceptual errors and misunderstandings in the paper. Here is a partial list (a full discussion is in section C below)

(1) BGMR claim that trimming was "inconsistent at baseline and endline". This is incorrect. We trim only at endline (for the endline results). We do not trim from the analysis sample any household based on their baseline value, and we only control for variables that have no outliers by construction, because they are dummies or discrete variables (number of adults, household head age, does animal husbandry, does other non-agricultural activity, had an outstanding loan over the past 12 months, household spouse responded to the survey, and has other household member). None of these variables were trimmed or truncated.

(2) BGMR claim that there are large imbalances in some baseline variables we don't report, and that this is a sign that our design or our data was somehow compromised. They evidently checked every variable one at a time and, naturally, they find some differences. But when you run 100 regressions, 5 will reject the null of no difference with a 5% test. This does not invalidate the design or the data. When we look at all of the variables taken together, there is no evidence of more imbalances than would be expected by random chance.

(3) Similarly, BGMR seem to believe that finding "effects" on variables that should not be affected by microcredit (such as language spoken at home) is a fatal flaw in the paper. But with many potential outcome variables, some will always turn out significant. This data mining does not make sense. Perhaps they mean to imply that our results are also due to mining the data for outcomes where we see an effect. But this paper appeared in a group issue of a journal, where all

the papers reported the reduced form results using exactly the same set of variables, imposed ex-ante to all papers. This constrains the analysis and leaves no room to data mine.

(4) BGMR seem to misunderstand the role of the credit take-up variable. One representative example of such confusion is "*This 'client' variable was also used to instrument the regression presented in CDDP Table 9. Therefore, the inaccuracy in borrowers' identification highlighted in this section has an incidence on the tests applied to check sample balance at baseline, on the estimation of the average treatment effect and on the estimation of the local average treatment effect*". This sentence is a string of logical and conceptual errors. [the client variable is not an instrument; it has no incidence on the baseline; it plays no role in the estimation of the ITT –presumably what they mean by average treatment effect--]

(5) BGMR claim that the fact that controlling for baseline access to credit changes the result comes from a coding error (they did not code the variable for the new households added at endline, resulting in missing value for anyone not surveyed at baseline). In fact, the change in the control variables makes no difference when we use the entire sample.

(6) BGMR claim that the fact that reclassifying utility loans at endline changes the effect of microcredit access on the probability to have up-taken a utility loan comes from the same coding error as above.

(7) BGMR believe that the correct way to analyze the RCT is to run the regression on a "consistent sample" which includes households with a mostly stable composition (in terms of size and members' gender and age) between the original survey and endline. This is simply incorrect, since the randomization insures comparability between treatment and control. This removes a large number of households, and selects the sample based on recall accuracy and stability in household composition.

(8) BGMR compare baseline and endline households without taking into account the fact that the endline sample is bigger, and was selected with different sampling weights. Thus, the households interviewed at baseline and endline have no reason to be directly comparable.

(9) BGMR don't understand our sampling and weighting strategy. They claim that our failure to predict probability of borrowing very accurately invalidates the design. It does not. A more precise prediction would have led to a more precise set of estimates, but the estimates based on our main sample are still internally valid. Adding the "low probability households" without weighting them (as they do) leads to consistent estimates (OLS is blue) but the interpretation is a bit strange since it mixes different types of household. Reweighting, as we do, is more sensible.

(9) BGMR often use words for the wrong purposes. For example "resampling" is a statistical strategy to perform bootstrap. It does not describe the choice of a particular sub-sample of the data.

B.4 The write up of the paper is disingenuous, needlessly controversial, and non-scientific

The paper is replete with statements that question our motives, our honesty, or our seriousness. It misrepresents its own findings in numerous places. It presents itself as a re-analysis, although it ignores most of our results. A companion paper (in French) which builds on this analysis to make more general points about RCTs (Bédécarrats et al., 2019) is even more misleading and polemical.

There is no need to comb this entire paper for such statements, but we provide below a commented abstract.

We replicate a flagship randomized controlled trial carried out in rural Morocco that showed substantial and significant impacts of microcredit on the assets, the outputs, the expenses and the profits of self-employment activities.

The original results rely primarily on trimming, which is the exclusion of observations with the highest values on some variables.

That is a mischaracterization of our paper and the authors' own results: of all the variables they report themselves, only the results on profits are affected by trimming at the top. They don't report anything on the other variables (consumption, labor supply, etc.), but the results on those variables are unaffected.

However, the applied trimming procedures are inconsistent between the baseline and the endline.

This is false. Households were only trimmed based on endline variables. In total, 27 households (0.5% of the sample) were removed. No baseline variables included in the specification are truncated, winsorized, or trimmed.

Using specifications identical to the original paper reveals large and significant imbalances at the baseline and at the endline impacts implausible outcomes, like household head gender, language or education. This calls into question the reliability of the data and the integrity of the experiment protocol.

The authors' own results are inconsistent with "large and significant imbalances at baseline". They find some differences on very few variables (as we do and report), as one would expect when running a regression on a very large number of variables. Similarly, with hundreds of variables, one would expect to see some random differences between treatment and control for some variables at endline, even if there is in fact no effect. This absolutely does not call into question the reliability of the data or the integrity of the protocol.

We find a series of coding, measurement and sampling errors.

The authors in fact identify very few coding, measurement or sampling errors. The small number of errors they found affect a very small number of observations and a very small number of variables. Correcting them makes no difference at all. What the authors find is a series of strategies and assumptions they don't agree with, and propose alternative strategies and assumptions. This is an entirely different thing.

Correcting the identified errors leads to different results.

As shown above, this statement is contradicted by the authors' own results.

After rectifying identified errors, we still find substantial imbalances at baseline and implausible impacts at the endline.

As shown above, this is irrelevant.

Our re-analysis focused on the lack of internal validity of this experiment, but several of the identified issues also raise concerns about its external validity.

This claim reflects the authors' misunderstanding of the weighing strategy that is used in our paper to estimate the effect of microcredit representative of our sample of villages.

The entire BGMR paper is written in this format. This is not the appropriate way to conduct or report a re-analysis. The objective should be to get a better understanding of what is in the data, not to obscure things.

B.5 Supplementary analysis confirms most of our initial results and confirms that there is a pattern of significant effects, but robust inference lessens our confidence in the effects on average business profit.

Despite the fact that the current re-analysis is deeply flawed, we see value in the effort to probe our results. We therefore re-analyzed our data using some of the methodologies that have since then emerged as "state of the art", and address some of the questions that BGMR ask.

(1) Inference

One thing that is clear from the original paper and was re-discovered by BGMR is that the distribution of the business variables has a lot of kurtosis. This makes the basic OLS estimates of average treatment effects not very interesting. It also potentially invalidates inference: the t-test we are using may not be normally distributed in finite samples either.

Fortunately, there are now methods to deal with this. Rank sum tests and permutation tests can be used to test various hypotheses of interest. Rank sum tests compare the mean ranks of the observations in the test and control groups. One advantage they have is that they do not depend on observations with extreme values. Permutation tests have the advantage of that they rely on the exact distribution of the test statistic and do not rely on large sample approximations of this distribution.[4]

[4] They are resampling methods that consist of re-assigning villages, within pairs, to a fake treatment and control status and to recompute the parameter based on this fake assignment. Doing that a very large number of times allows us to identify what the distribution of the true estimated parameter would be under the assumption that there is no effect. The true estimate can then be compared to this distribution, rather than to the distribution based on large sample approximation. The validity of large sample approximation depends primarily on sample size. However, what a large sample means also depends on the shape of the distribution of the outcome variable considered. If it is highly skewed and has long tails, an increased number of observations is necessary for the approximation to be valid. This is something which is not necessary with permutation tests.

Moreover, Young (2019) suggests running joint tests of significance based on permutation tests for all the key variables of interest in the paper, to avoid cherry-picking results.

Table 1 presents the results.

Starting with the mean test, for each individual variable, the permutation test rejects the null for assets, sales, and hours worked outside. It does not reject the null for consumption, and number of hours worked in self-employment, which is also what we had originally found. It does not reject the null at the 5 percent level for income from daily labor, and business expenses, where we had found significant differences before. For profits, the p-value goes from 7% to 22%. This highlights the role of the kurtosis in those variables.

The test for the decile treatment effect has the same results, except that it also rejects the null for profits. The rank sum test has the same results as the mean test, except that it also rejects the null for business expenses. **All three tests reject the null of no effect for all variables taken jointly.**

This analysis highlights the implication of the very fat tail of business variables for inference, and hence the need to look beyond average treatment effects, and also to look at the significance of variables jointly.

(2) Choice of control variables.

In Tables 2 and Table 8 we re-analyze our main outcomes, following the method of Belloni et al. (2014) to systematically choose control variables. Both tables include 94 variables to choose from. Access to credit is included in the choice set since this was highlighted by BGMR.
The procedure uses LASSO to pick the relevant control variables. We include the baseline variables that are sufficiently correlated with treatment (after imposing the LASSO penalty) and the variables that are sufficiently correlated with control (after imposing the LASSO penalty). There are two results. First, no variable is chosen in the first step (this suggests that the treatment is in fact balanced, contrary to what is claimed by BGMR). Second, including the variables picked by LASSO in the second step makes no difference at all in the results.

(3) Disaggregated asset variables

One valid point made by BGMR is that the valuation of assets in the original study is a bit arbitrary. It is based on a very small sample of households that actually sold or purchased any asset. This causes problems with very expensive assets (like tractors and reapers) which led us to ultimately remove them.

To address this issue, we report a new analysis where we simply look at the data asset by asset, and perform a joint test of significance across all assets. This sidesteps the issue of aggregation. We reject the null of no effect for most assets and jointly reject the null (with the two large assets we ignored

in our original analysis, tractors and reapers), which confirms the original finding that there is a treatment effect on assets.

C. Detailed technical responses to BGMR

C.1. Trimming procedures

BGMR question the trimming procedure at baseline and endline of our paper, and affirm that changing the trimming threshold changes the main results of the paper.

They start their criticism with a citation from Deaton and Cartwright (2016): "*When there are outlying individual treatment effects, the estimate depends on whether the outliers are assigned to treatments or controls, causing massive reductions in the effective sample size. Trimming of outliers would fix the statistical problem, but only at the price of destroying the economic problem; for example, in healthcare, it is precisely the few outliers that make or break a programme.*"

In their section 3.1 - *Different trimming procedures were applied at baseline and at endline*, BGMR criticize the fact that we do not apply the same trimming procedure at baseline and at endline. They point to the fact that the trimming procedure at baseline was done variable by variable. They also criticize the decision to not trim based on the hours of work variable.

In their section 3.2 - *Variation in impact estimates depending on trimming threshold,* BGMR then criticize our trimming procedure. They argue it is not valid and they show different trimming procedures in their Table 5 results. They conclude section 3 with this statement: *In sum, CDDP trimmed 459 observations (10.3%) at baseline, removing only the most extreme values on those observations, while at endline they trimmed 27 observations (0.5%) differently by removing them entirely. The fact that the final results vary substantially depending on the number of removed observations could mean that there are data quality issues.*

Comments:

While BGMR cite Deaton and Cartwright (2016) to justify their exercise, they seem to misunderstand the main point that is being made: 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. In their headline results (that the results depend on trimming), BGMR rediscover a result that is already in our paper (Figure 1), and has been analyzed in detailed by Meager (2018): all the average results on assets, profits, and expenses in our paper are driven by the very top of the distribution.

C.1.1. About different trimming procedures at endline and baseline

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. Baseline data is used to build the following set of covariates: the number of household members, number of adults, household head age, does animal husbandry, does other non-agricultural activity, had an outstanding loan over the past 12 months, household spouse responded to the survey, and has other household member. **None of these variables were trimmed or truncated at baseline.**

Trimming was applied on the continuous baseline variables only for the purpose of the balance test (Table 1 of the original paper).

C.1.2. Trimming procedure at endline

Removing "outliers" is a delicate exercise. On the one hand, the underlying variables like profit, assets, and expenses, have distributions with very fat tails (Meager, 2018) and removing the high values would throw out most to the relevant observations, precisely as Deaton and Cartwright (2016) emphasize. As Meager (2018) writes, "The heavy tails (extreme kurtosis) in the household business outcomes has both methodological and economic implications. Ordinary least squares regressions such as those performed in the original randomized controlled trials are likely to perform poorly compared to quantile regression techniques or parametric modelling of the tail (Koenker and Basset 1978, Koenker and Hallock 2001). More substantively, heavy tails suggest that in these populations, certain individual households account for large percentages of the total business activity. It may be challenging to understand the economies of developing countries if we trim or winsorize the most productive households who make up a large percentage of total economic activity."

On the other hand, those outcomes are also hard to measure well, and there are clear measurement errors (see De Mel et al., 2009).

To balance these objectives, the approach we use aims to catch a very small number of households where the data collection was likely compromised without any "guessing" that could lead to cherry-picking, and to keep the most observations we can, even those with very high values. Our procedure is to parameterize so that only 0.5% of households are removed at endline. The trimming procedure for the endline is as follows: considering a set of key outcome variables, we compute the ratio of the observations to the quantile of order 90%. Observations are then ranked by this ratio, and 0.5% of households with the largest ratio values are removed[5]. The advantage of this procedure is that the sample remains exactly the same across all regressions, and a very small number of households (27)

[5] BGMR note that we do not include the variable "*number of hours worked"*. That was because this variable does not have any outlier: The ratios between the values of the variables included in the trimming and the 90th percentile of the 27 observations selected to be trimmed range from 58.38 to 21.35. The ratio between the maximum value observed for the variable "number of hours worked" and the value corresponding to the 90th percentile of the same variable equals 3.13, which is well below the threshold used to select the observations to be trimmed.

are removed, most of which have very high values for a number of variables. For example, removing 1% of observations variable by variable on say 20 variables could lead to having 20% of households affected in one way or another by the trimming procedure[6].

In Table 5, BGMR apply the trimming procedure used by Crépon et al. (2015) with different thresholds including 0%, 0.5%, 1%, ..., up to 5%. They claim in the text that removing observations at the top of the distribution changes the results.

But the numbers reported in the table do not support this interpretation: first, across the entire table, it is certainly impossible to reject equality in the estimates they report with various trimming thresholds and what we report (granted, the standard errors are large). Second, even looking at point estimates, the results look quite similar across all rows and columns except for zero percent trimming for profit (this row is driven by very few observations, at most 27, which we removed with some purpose) and with a trimming of 3% or 5% of the observations.

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. Not that QTE are by definition robust to outliers, and well suited for looking at these kinds of variables: This result is found in all the different microcredit studies published in the AEJ applied special edition (see Meager, 2019). Given this heterogeneity, it is not surprising that trimming a large proportion at the top of the distribution of our sample makes the treatment effect smaller. BGMR fall victim to exactly what Deaton and Cartwright were warning us against: they get rid of the interesting results by attempting to "clean" the data.

C.1.3. Supplementary analysis

Meager (2018) investigates the robustness and the generalizability of the quantile treatment effects of all the studies published in the AEJ: Applied special issue. In particular, she applies Bayesian methods to compute QTE with partial pooling (that is, taking into account the results of the other studies). The partial pooling results are reproduced in Figure 2, Panel C. She finds that, for Morocco in particular, the results of significant effect in the higher quantiles (except for the 95th) remain significant, even with partial pooling.

We also provide supplementary analysis to test the robustness of the main results in our paper given the presence of fat tails in the dependent variables. Traditionally, most tests are conducted under the

[6] As pointed out in BGMR, we start from 5,551 households and remove 0.5% of them through our trimming procedure, which makes 27 observations out of 5,524. There is indeed a typo in footnote 7 of our paper (page 130). The right number of observations for the entire sample is in fact 5,524 as shown in Panel B of Table 8 where we present treatment effect estimates for the whole sample. Since in footnote 7 we only refer to use data collected at endline, our statement that no further trimming is done in the (endline) data is accurate. Other tables use 4,934 households identified, using data from a preparatory survey, as the most likely to take loans.

assumption that the distribution of t-statistics is normal. This is true but only asymptotically. In the presence of distributions with heavy tails, asymptotic might imply a very large number of observations well above the size of the sample at hand, and Meager shows that the standard t-test can be misleading.

Since 2015, new methods have been developed and some existing methods have been revived to improve transparency in data analysis (see for example Imbens and Rubin, 2015.) In this supplementary analysis, we implement two of these methods - Rank tests and Permutation tests.

Rank tests are only based on the ranks of observations in the treatment and control groups and thus are not subject to potential outliers. The idea of the test is very simple: it consists of ranking observations in the whole sample and comparing the average rank of observations in the treatment and control groups. As they are only based on the rank, these tests are robust to outliers[7]. The null hypothesis is that potential outcomes have the same distribution. Note that these tests are not powerful enough to test the hypothesis of any change in the distribution of potential outcomes. There are patterns of individual treatment effects that they fail to detect (for example, they would not detect a situation in which half the sample have a very negative impact and half the sample a very large impact).

We also run permutation tests. The null hypothesis in permutation tests is a sharp null: the treatment affects nobody. The advantage of permutation tests is that they provide exact p-values under this null hypothesis. The idea of a permutation test is to randomly generate a pseudo-treatment variable, following the design, and to run the estimation as if this pseudo variable is the true treatment variable. Replicating this operation a very large number of times (5,000) we can compute the proportion of pseudo-treatments that generate an estimate with an absolute value that is above the absolute value of the initial coefficient. If this share is smaller than the level chosen for the test (5%) then the hypothesis is rejected. Those tests are not affected by outliers because if there is an outlier, it will be equally likely to be assigned to treatment and to control in pseudo assignments. For these reasons, randomization inference tests have become popular in the analysis of RCTs (see Athey and Imbens, 2018, and Young, 2019).

We present the results of our analysis in Table 1. We show the p-values corresponding to each test for the main variables of our paper. In the first panel, we reproduce the original p-value, based on the standard assumption of asymptotic normality. In the second panel of the table, we provide in the first row the p-values of the permutation tests in which we consider the mean of each variable. It simply corresponds to the average over the pairs of the difference between treatment and control villages (it is thus analogue to the first panel of the table, but uses a permutation test)[8]. The second row provides the p-values derived from the permutation tests but considering the sum of the absolute values of the differences between deciles for the considered variables. It is useful to consider

[7] There are two version of this test that can be used: the Wilcoxon test, which does not account for stratification at the pair-of-villages level, and the Van Elteren test, which takes it into account. Both are suitable but they do not have the same power.
[8] The test simply consists of randomly re-assigning villages to a pseudo treatment and a pseudo control, and recomputing the estimated parameter. We simply count the proportion of times this pseudo estimate is above our estimated parameter with the true assignment (in absolute value).

quantiles to detect potential distributional/heterogeneous treatment effects (and they are also robust statistics). The last panel shows the p-values associated with rank tests. We implement the Van Elteren version of the test which accounts for stratification.

The results in Table 1 confirm the results of our paper, except that they highlight that the average effect on profits is quite imprecise (profits is where the tails are the fattest). For each variable (except for income from day labor) we consider there is at least one of the new tests that rejects the null, even if we multiply the lowest p-value by three (to correct for multiple testing following the Holmes procedure). There are variables such as assets or sales for which all three tests clearly reject the null hypothesis tested. There are other cases in which not all the tests detect an effect. That is, for example, the case with profits: only the test based on deciles rejects the null hypothesis. This reflects the fact that impacts on profits are negative for low quantiles and positive for large quantiles, something we noted in the paper.

The last column provides, for the two permutation tests and for the rank test, a test of the joint null across all the variables we consider. As can be seen, we reject the assumption in the three cases.

To conclude, we show in this section that
(1) trimming procedures on baseline outcomes have absolutely no influence on the estimation of treatment effects on endline outcomes.
(2) the "sensitivity to trimming" highlighted in BGMR is a re-discovery of a result in the original paper, which is that results on assets, expenses and profits are driven by what happens in the higher quantiles.
(3) supplementary robustness tests provided in this section, based on more advanced methods, largely confirm the initial results of Crépon et al. (2015), with the exception that the results on average profile are shown to be fragile.

C.2. Set of baseline covariates

C.2.1 Imbalance of baseline characteristics

BGMR do two things in this section. They first provide results of balance checks adding new variables compared to the set of variables used in Crépon et al. (2015). For example, they consider the spoken language, the access to utilities, etc. Second, they estimate impacts on the main outcome variables extending the set of control covariates included in the regression to include covariates that were imbalanced at baseline. Results are displayed in Table 6 of BGMR. They find some imbalances, most notably on sales and profits. BGMR claim that the introduction of the new set of covariates that they have selected changes the results: *Controlling for all the variables identified as imbalanced at baseline increases the magnitude and the significance of the estimated impacts on assets, sales and expenses.*

Comment:

We have several comments on this section and in a supplemental analysis we show an improved procedure to select control variables into the main regression.

C.2.1.1. *About the sensitivity of Crépon et al. (2015) results to a different set of control variables*

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, goes from 6,061 MAD (Moroccan Dhirams) in Crépon et al. (2015) to 6,518 MAD in Table 6 of BGMR, while standard errors are respectively 2,167 and 2,690 (note the increase in standard errors). This is the case across most variables: they are never statistically significant, and they are generally very much in the same ballpark.

C.2.1.2. *About balancing checks*

In Crépon et al. (2015) we chose to include a parsimonious set of well measured variables to introduce at baseline that are representative of the main dimensions of the analysis: household composition, access to credit, self-employment activities, risks and consumption. The fact that there are some random imbalances on some variable is not a sign of systematic issues with the design. By construction, in bivariate regression, some of the variables will appear to be unbalanced at baseline: for a test with a 5% level, we know that by construction in 5% of cases there will be an impact detected even if there is no real difference. 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.

C.2.1.3. *About specification choices*

Whether or not to include baseline covariates, and which covariates to include, is a difficult question. The most transparent specification rests on the simplest specification based on the design of the experiment. Researchers depart from this basic specification and 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.

In order to get rid of specification choices, Crépon et al. (2015) define a main specification that is constant across all regressions. In addition, the robustness of results to the set of control variables is checked in Table B7. In panel A of this table, treatment effects are estimated through a simple specification that only includes design variables (i.e. controlling by dummy variables for the pairs of villages but no baseline covariates), as proposed by Athey and Imbens. In Panel C, we estimate

another reduced-form specification with an extended set of covariates, including the dependent variable at baseline. As in BGMR's own Table 6, results show very small differences with the core specification.

Another issue with control variables, which is related to the issue of trimming that was discussed earlier, is that introducing continuous control variables with potentially very large values will lead to bias in small samples.

C.2.2. Supplementary analysis

To settle the discussion of control variables, in Table 2 we present new results in which covariates included in the regression have been selected by the double post lasso procedure presented in Belloni et al. (2014). This method allows us to pick control variables in the presence of a large set of potential control variables in a manner that is consistent, and does not lead to wrong estimates of the standard errors.

This procedure starts with a large number of potential covariates. We consider here a very large number of covariates, including socio-demographic variables, main outcomes measured at baseline, disaggregated loans at baseline, and several other variables such as language spoken by the head of the household and the dummy variables corresponding to the pairs of villages. This makes a total of 94 covariates, not counting for the village pair dummies.

For continuous baseline variables with fat tails (profits, assets, etc.), we also take on board recent results from the matching literature, which warns against controlling for variables with heavy tails because it can lead to bias in finite samples (Imbens, 2015). The recommendation is to include transformed versions of control variables. A standard transformation is to include indicator variables for quantiles.

The first step predicts the treatment variable based on all these covariates, using the LASSO. This can actually be seen as a balancing test and helps select unbalanced variables (this idea is discussed and developed in Ludwig et al., 2017). The regression is penalized because it includes many variables.

In the second step, we perform the same operation on the dependent variable. It helps select variables that are good predictors of the dependent variables. The main motivation is to get robust consistent estimates, but an additional gain of this step is to select variables that will increase the power. In the last step the treatment effect regression is run including both sets of control variables: those predicting the treatment and those predicting the dependent variable[9].

[9] The Stata command we use to implement this procedure is pdslasso. We enter the command:
pdslasso `var' treatment ($X _Ipaire_*) if samplemodel==1, partial(_Ipaire_*) cluster(demi_paire_n) lopt(prestd) noi;
Where $X contains all the potential covariates, _Ipaire_* includes the dummy variables for the pairs, samplemodel selects the sample of our main results which is in our main sample, partial() indicates that we did not penalize the dummy variables, cluster(demi_paire) indicates that the parameters of the lasso have to be computed using clustered standard errors, and

A first important result is that in the lasso step to predict the treatment variable **not a single variable was selected out of the 94 that we introduced**. This can be seen as a balancing test accounting for multiple testing. This is reassuring and suggests that BGMR's concerns about imbalances is misplaced.

Regression results are shown in Table 2. The results with the double lasso procedure and the results presented in Crépon et al. (2015) are very close.

To conclude, we have seen in this section that:

(1) BGMR go on a fishing expedition to find new variables that have some significant difference, and naturally find some

(2) However, including those variables makes no substantive difference to the findings, according to BGMR's own results

(3) A principled method to choose control variables reveals no systematic imbalances (which is actually better news than what we described in our main paper) and finds results that are unchanged from our original specification.

C.3. Discrepancies in the code and measurement errors addressed by BGMR

In this section, BGMR question the reliability of the administrative data used to identify the borrowing households in the sample in Crépon et al. (2015).

C.3.1 Discrepancies between administrative and survey data

BGMR claim in this section that measuring the credit variables at endline is a key step in the analysis of Crépon et al. (2015). They show that there are some discrepancies between the administrative data and the survey data concerning the identification of Al Amana's borrowers in treatment (and control) villages, and claim, without much evidence to back the claim, that the main explanation for the discrepancy between administrative and survey measures is related to an imperfect matching procedure.[10] They suggest that this discrepancy is a direct threat to the analysis.

BGMR then affirm (page 17): "*This 'client' variable was also used to instrument the regression presented in CDDP Table 9. Therefore, the inaccuracy in borrowers' identification highlighted in this section has an incidence on the tests applied to check sample balance at baseline, on the estimation of the average treatment effect and on the estimation of the local average treatment effect*".

the option lopt(prestd) is used to standardize covariates before introducing them in the procedure. Noi is a useful option that displays intermediate steps, especially the list of variables selected by the lasso procedure at each step.

[10] They support the claim by pointing to a paper advocating for frauds by loan officers in microfinance institutions, and one master thesis of a student on the case of Al Amana. We were unable to find the paper on the link provided in the paper.

Comment:

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). 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). 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. The main parameter of interest is the coefficient associated with the treatment variable (being equal to 1 for a village assigned randomly to receive Al Amana and 0 otherwise). Finally, it is correct that the estimation of Local Average Treatment Effects (LATE) depends on the variable "client". Note that there are a number of other issues in the LATE which are discussed in Crépon et al. (2015), including the possibility of externalities. The other papers in the AEJ applied issue usually do not present this specification at all, and we are careful to be very cautious when presenting these results.

BGMR are correct to point out that there are discrepancies between the recall variables and the administrative data. This is true in all the other microfinance studies as well. Typically, researchers consider the administrative data to be more reliable than the recall data. We have no reason to systematically doubt the matching procedure, so as other studies before us, we have chosen to use the administrative data. In fact, we paid close attention to the quality of the administrative data. Throughout the study, we had in place a reporting system that allowed us to closely follow loan disbursements in treatment villages. On a weekly basis, we received lists prepared by the loan officers that included information about disbursements made to clients that lived in treatment villages. We followed this strategy during the whole two years of the experiment. In addition to this, we verified that reported loans had indeed been disbursed through the Al Amana management information system[11].

That said, we are aware of some discrepancies between the administrative data and the recall data in our survey. As we mentioned, this is common in experiments. We start by describing the potential sources of these differences and we then discuss the reasons why it is very unlikely that these discrepancies represent a threat to our analysis (even to the LATE and externality regressions).

[11] Al Amana internal procedures establish that no loan can be disbursed without being recorded by loan officers in Al Amana management information system (MIS). Approvals from supervisors are made through the system as well as the provision of funds to the branches. The challenge for the study was that in Al Amana MIS there isn't a specific field where loan officers record the village where the client lives. The village is recorded as part of the address and it is not always accurately recorded. That's why we decided to put in place an ad hoc system that helped us identify loan disbursements in the villages of the study. We used maps produced by loan officers that covered the entire intervention area of the branch. Loan officers placed in this map all existing villages in the intervention area. During a training we conducted to explain the study research protocol, we went through the maps together to carefully identify treatment and control villages participating in the study. Based on this, loan officers prepared weekly reports including exclusively loans disbursed in the villages concerned by the study. As explained before, we then verified that these loans where indeed disbursed through Al Amana MIS.

As BGMR mention, this mismatch comes from two sources. First, there are some households that are considered clients in our administrative data but who did not declare to having borrowed from Al Amana. It is important to note that in our survey at endline we asked if the respondent had an active loan from Al Amana (still reimbursing it) or if they finished repaying a loan from Al Amana in the 12 months before the survey: a larger share of households in treatment villages declared to have an active loan from Al Amana than a matured loan (8% and 4% respectively). We can then look, using administrative data, at the period at which people took a loan in the treatment villages. Most clients in treatment villages enrolled right at the beginning of the experiment (70% did so during the first six months). Given that the survey took place on average 2 years after the baseline survey, it is thus likely that there is a recall bias for some of those households. There, the administrative data is surely more reliable. Second, there are some households (152) who declared borrowing from Al Amana and are not identified as such in our administrative data. This could also come from recall error, or there could be exclusion errors in the administrative database or in the matching. We use more recent data for an ongoing long term follow up to improve the matching, and are able to confirm that 65 of those are indeed clients, suggesting that there was indeed some exclusion error in our measure of being a client: we thus may under-estimate the number of Al Amana clients in treatment villages.

To quantify the extent of possible misclassifications, we construct new take-up variables based on different information and assumptions: first, we improve our matching procedure (as described above) and update the take-up variable from the administrative data. Second, we construct an upper bound estimate of the take-up data by combining all households that are identified as clients in our administrative data (improved method) or in our survey. Table 3 shows the take-up estimations according to the different definitions. We find that the take-up differential between treatment and control villages ranges from 16.7 percentage points using the initial definition (as in the paper), to 17.6 percentage points with the update on the matching procedure using the administrative data, and to 18.2 percentage points when both administrative (updated) data and survey data are combined.

We now turn our attention to the implications of this change on the estimation of externalities and its implications for the computation of the LATE. We re-estimate the score used to identify the households with the 30% largest probability to borrow and the households with the 30% lowest probability to borrow based on the different definitions of loan take-up. We report the results in Table 4. Results remain unchanged. As before, there is no sign of any externalities. Meanwhile, the LATE only changes with the scaling parameter in the denominator (ranging from 0.167 to 0.182).

C.3.2. Definition of the baseline covariate on access to credit

BGMR claim that Crépon et al. (2015) have omitted loans from other MFIs when checking for balance. They add, *"This result has an incidence on the impact evaluation results, as illustrated in the following section"*. They also claim (5.1.3) that Crépon et al. (2015) only consider outstanding loans at baseline and that we do not take into account loans that were outstanding during the last 12 months. They

affirm that *"the inconsistency between borrowing recall periods at baseline and at endline is problematic when it comes to evaluating the impact of growth in access to credit"*, and also that *"Total access to credit is used by CDDP as a control variable. The increase in their values after correcting the errors pointed out in 5.1.3 and 5.1.4 therefore modifies the measured impact results."*

Comment:

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[12].

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 Table 5. Obviously, this will make no difference.

We chose to consider only outstanding loans to build this baseline control variable since we consider current loans to be the most important, given that they are a measure of indebtedness at the time of the baseline survey[13].

The fourth statement 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. Therefore, the following comment in BGMR – "*The inconsistency between borrowing recall periods at baseline and at endline is problematic when it comes to evaluating the impact of growth in access to credit*" – is simply irrelevant. It would perhaps matter for a difference in differences specification that they appear to have in mind, but we don't see any reason to run such a specification.

Finally, BGMR also affirm that changing the measure of the baseline covariate on access to credit significantly affects estimated effects: *"Total access to credit is used by CDDP as a control variable, the increase in their values after correcting the errors pointed out in 5.1.3 and 5.1.4 therefore modifies the measured impact results. For instance, the average treatment effect on access to AAA credit was estimated in CDDP Table 2 at 0.09*** (0.01), while it gets to 0.069***(0.01) when correcting this error, which indicates an impact lower by 30% of the experiment on credit take-up. The average treatment*

[12] The mean of 0.060 (in the control group) is comprised of both 0.016 of households that declare to have an outstanding loan from an MFI different from Al Amana and 0.045 of households that declare to have an outstanding loan from formal institutions other than an MFI. The variable that corresponds to the line "Loans from other formal institutions" in Table 1 is called borrowed_oformal2, which is defined as 1 if a household member had either an outstanding loan from a formal institution other than an MFI or an outstanding loan from an MFI different from Al Amana. The first definition is: egen aloans_oformal2 = rowtotal (aloans_oformal aloans_oamc) (line 53 of Baseline do-file). We then define borrowed_oformal2 equal to 1 if aloans_oformal2 >=1.

[13] Notice, in addition, that while we find that outstanding loans are not perfectly balanced between treatment and control groups at baseline, that is not the case when considering the aggregate of both outstanding and matured loans in the past 12 months. Thus, had we followed the strategy of BGMR to include non-balanced variables as baseline household controls (which we do not adhere to), we would not have selected total loans as a variable to include as a control in the regressions.

effect on self-employment profits was also estimated in CDDP Table 1 as 2,005 (1,210), which is substantial and significant at the 10 percent level. Once corrected for the errors in total access to credit at baseline, the estimated treatment effect on profits becomes 1,454 (1,253), which is smaller and insignificant*".

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.

Although we have already shown in Section C.2 that estimated treatment effects of Crépon et al. (2015) are robust to different sets of baseline covariates, we analyze the specific implications of controlling by different definitions of baseline access to credit. Panel B of Table 6 shows that the results are not affected when loans from other MFIs are added to the vector of control variables. This is not surprising since households that had an outstanding loan with an MFI other than Al Amana were well balanced between treatment and control groups and represented only a minor portion of the baseline sample.

We then verify the robustness of the original results when controlling by a broader measure of baseline access to credit that also includes matured loans. Contrary to what is affirmed by BGMR, we find no difference in the estimated effects either (Panel C, Table 6). The lower estimates obtained in BGMR for Al Amana loan take-up and profits (from 0.090*** to 0.069*** and from 2,005* to 1,454, respectively) are not due to the use of a broader measure of access to credit as a baseline covariate, as claimed by the authors, but by estimating the effects on a smaller sample. The revised estimated effects in BGMR are based on a sample of 3,525 observations instead of 4,934 observations as in Crépon et al. (2015).

The smaller sample corresponds to households with a high-probability-to-borrow score surveyed both at baseline and endline. **This was not intentional. A mistake in their code on one of the control variables (credit at baseline) inadvertently changes to missing all of the endline observations of households not surveyed at baseline (around 1,400 observations). This is not comparable to our core results.** Later on, BGMR claim that this is the appropriate sample to use (we disagree, as we will describe below), but here, the point is that they claim that the results change because of a different set of control variables. That is simply not the case.

The findings exposed in this section thus confirm the results from the robustness tests performed in Section C.2.2. as well as those Crépon et al. (2015).

C.3.3. Utility loans

BGMR point to the fact that "other credits" were classified as loans from a utility company at baseline and at endline. They do some recalculation of amounts at baseline and at endline for the various types of credit and end up with the conclusion that there was a significant increase in loans from a utility company (when they are not aggregated with other loans) in treatment villages. Given this increase,

they claim that Crépon et al. (2015), instead of measuring the impact of access to microcredit, measure the joint impact of increased access to microcredit and increased access to utility credit. 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.

Comment:

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"). **We follow the BGMR strategy and reclassify these loans at endline, which leads to a change in the proportion of households that have access to a utility loan from 16.9% to 16.2% in the full sample (5,551 households).** This should not affect the results, and as we show below, it does not.

However, the classification that they operate in the baseline is incorrect. Our baseline survey instrument has a single response option called "other loans" where both utility and other loans were recorded, as opposed to our endline survey instrument where we included a specific pre-coded response for utility loans (in addition to "other loans"). The reclassification conducted by BGMR at baseline thus relies entirely on the description provided by respondents when they declared to have a loan recorded as "other" (i.e. from a source different than the ones pre-coded). The specification of other loans is unfortunately missing for most of the loans (71%) declared in this category, which implies that no reclassification can be conducted for this group of loans. BGMR assume that all loans for which the specification is missing are not from a utility company, which implies a strong assumption that is not discussed in their article and that is obviously inconsistent with the numbers we are getting at endline, where we asked separately about utility and other loans, and we find a tiny fraction of "other loan". The only option we consider technically correct at baseline is to define a unique variable that aggregates both utility company loans and other loans.

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. BGMR claim (p19) that "*this rectification also alters the computing of the average treatment effect on access to utility credit at endline. This was estimated at 0.017 (0.017) in CDDP Table 2, which is small and insignificant. Conversely, when preventing unjustified reclassification, it becomes 0.037**(0.016), which is large and significant [..] It is unclear whether this significant increase in access to utility credit at endline in treatment villages is an unexpected impact of increased AAA credit or contamination by a co-intervention".*

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). When we reconstruct the variable utility loans according to the definition of BGMR, but use the full sample of high-probability borrowers (as it should be), we find that the effect of Al Amana on access to utility loans is 0.016 (0.017) (see column 3 of Panel A, Table 7). This is a very small and non-significant effect, and almost identical to the original estimates of Crépon et al. (2015) (see column 1 of Panel A, Table 7).

C.3.4. Credit access at baseline and endline

BGMR then compare the sample of individuals surveyed at the endline (which they call "cross-section") and the sample of individuals surveyed at endline and baseline (which they call "panel"). They compare baseline and endline averages in their Table 11. They find they are very different and thus claim the data have to be analyzed in difference. When they look at the difference they say they find a different pattern of results.

Comment:

In Table 11 BGMR make several comparisons. The first panel called "cross section" mixes information on access to credit on cross-section households that were surveyed at endline only and information on access to credit on individuals surveyed both at baseline and endline. By definition, baseline data is only available for households surveyed at baseline. **But 1,400 households were added to the sample between baseline and endline! Thus, looking at the evolution between different household samples makes no sense (first panel).** This is particularly flawed since the households were not randomly selected at endline. Selection was based on the predicted probability to borrow based on the baseline census variables, in treatment and control group. So obviously these households are different. **Similarly, it makes no sense to compare the endline sample to the sample that was surveyed both at baseline and endline (second panel).**

Notice in addition that the table is an invitation to perform difference in differences analysis without doing it explicitly and also without providing standard errors. Thus, we cannot conclude much from the table. The only thing that can be taken from the table is that the restriction to having a baseline and an endline is strong.

In the first panel of Table 11, we can compare treatment and control at endline (and we find results in line with Table 2 Crépon et al., 2015). The second panel of Table 11 (panel data) and most of the subsequent tables intentionally (this time) restrict the analysis to households interviewed both at baseline and endline, and ignore the 1,400 households we added at endline. Exploiting the panel structure of the data would make sense in a non-experimental setting where access to microcredit would be correlated to unobserved, time invariant characteristics (which a first difference would eliminate). In such a case it would be crucial to strongly invest in building a representative set of households to follow at both dates. But such a strategy is unfounded with experimental data. The comparison between households assigned to treatment and control provides consistent estimates,

given randomization. The priority for an RCT is thus to have as informative an endline sample as possible. After completing the baseline and getting information on the take-up of the experiment, it became clear that the initial power calculations were too optimistic. We thus chose to add 1,400 new households at endline. By definition, these people were not surveyed at baseline. What BGMR find in the paper in this smaller sample are virtually identical point estimates, and slightly larger standard errors.

C.3.5. Supplementary analysis

As we already discussed, Table 10 in BGMR points to imbalances in baseline loans, and then their Table 11 implicitly recommends using difference in differences estimates in access to loans. We show here that when adding control variables to control for potential imbalances, there is no impact on the estimates of Crépon et al. (2015).

As already discussed, there is no need to adjust for covariates to get consistent estimates, and in fact the current recommendation is to use no control variables in RCTs. The only potential advantage in our case would be a gain in precision by reducing the residual variance. However, in practice the gains are usually small. A risk is that the choice of covariates leads to specification search and p-hacking. The best practice is to be very transparent on the specification estimated, to have the same specification from start to finish, and to introduce robustness checks with and without covariates.

The additional results we present here are based on the implementation of the procedure proposed in Belloni et al. (2014), that we already discussed in section 2.2. This procedure has the advantage of minimizing researchers' temptation to do specification search. We now implement the same procedure, starting from the same set of potential baseline variables, but restrict it to borrowing outcome variables. We show the results in Table 8. The first step in this procedure is, as before, to run a lasso regression to predict the treatment variables based on baseline covariates. As before, none of the baseline variable is selected by the procedure. This means that there is no variable to include in the regression. The second step is, variable by variable, to identify with a LASSO regression which are the predictors of the relevant variable that should be included. It is in this step that potential accuracy gains can be obtained. For example, the baseline variable of the outcome variable considered can be selected or it can be not selected. This choice is made by the algorithm and not the researcher, hence limiting specification search.

It is clear from the table that this makes no difference to the results. Interestingly, the baseline variables measuring borrowing activities are never selected as potential regressors to introduce in the first step. Most of the time a few variables are selected.

To summarize:

(1) The discrepancies between the administrative data and the recall data are to be expected, given both recall errors and potential administrative errors. We continue to believe that the

administrative data is more reliable than the recall data, but we show that using either of the variables or a combination of the two does not affect the results.

(2) Two of the headline results in this section (baseline credit access in the main specification, and access to utility as one of the outcomes) are due to a coding error. When this error is corrected, they make no difference.

(3) The implicit difference in differences proposed in Table 11 makes no sense since the sample is not the same in the baseline and the endline.

(4) Results on borrowing outcomes are unchanged when baseline levels of credit access are permitted to enter the set of control variables in a Belloni et al. (2014) doubly robust specification.

C.4. Discrepancies in the measurement of outcomes

C.4.1 Outcomes measures

In this section, BGMR criticize the way Crépon et al. (2015) have built some of the main outcome variables. The variables they consider are mainly assets and sales. To measure assets (we measure sales in a similar way), our survey had a detailed list of assets and we asked several questions to the head of household regarding each of these assets, including number of assets and unit prices when there was a transaction to record them. These prices are used to compute a value of each asset. We use the median price for all observed transactions in the sample. BGMR criticize the prices used to value assets in Crépon et al. (2015). They propose a new method to aggregate individual assets, basically considering the median of prices at baseline and endline that they consider as more robust. Another criticism is that Crépon et al. (2015) did not include tractors and reapers in the list of assets although these items are on the list that was recorded.

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

Comment:

The first comment is that using different prices makes essentially no difference to the conclusion (the profit point estimates is larger, but given the standard errors, not distinguishable).

In Crépon et al. (2015), the individual assets have been aggregated using current prices, and BGMR propose another way. Kling et al. (2007) propose yet another way based on standardization of each variable. What is the best aggregation method? BGMR's claim, that it is better to consider baseline and endline prices to value items, is not well documented or argued for. There might be cons to this view, especially if there is large volatility in prices.

We can think, however, somewhat more generally about the problem. We have a set of individual assets. Both our method and the method proposed by BGMR consist of finding weights to build an index and to run the regression on this index.

A first comment is that this will only be an index and the role of prices is mainly to build the weights. This method is used in many places. However, no method is perfect. For example, who could claim to have the whole list of relevant assets or to have the relevant prices to build the most appropriate weights? We have collected valuable information about assets and the question is how to use these data in the best way to answer the following question: is there an impact on assets? Building an index with some measure of prices, as we did, or as BGMR do with other prices, is just one possible answer to this question.

A second comment is that, at the item level, we actually only have a noisy measure of each item. When we build an index we also have a measurement error, which mixes all the measurement errors on the disaggregated components of the asset index. This error is then transmitted into the residual of the regression. Adding more mismeasured items makes the relation noisier and in the end leads to a lack of power, especially if the error on the measurement of the item is large and the impact predicted is small. This is why we decided not to include tractors and reapers. 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. As BGMR note, "*Including tractors and reapers in the asset appraisal at endline increases average asset value in the sample from 1377 to 5111. It also modifies the impact estimation on total assets at endline. This was 1,448** (658) in CDDP Table 2, which is substantial and significant. It becomes 1,741 (1,255), which is larger but insignificant, when we include tractors and reapers in total assets, while keeping the same control variables as CDDP."* **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.**

C.4.2 Supplementary analysis

That said, just excluding tractors and reapers because they are large and there are two few price points for them is not disciplined. It would have been much better on our part to propose a principled rule of selection.

In this section, given the legitimate uncertainty about the best weighting system to use, we provide a new analysis on assets that rely only on micro-aggregations of similar assets. This seems to be a better way to proceed moving forward for studies of this type. In table 9, we decompose the assets into 18 basic items, simply look at all the basic items one by one, and then apply the Benjamini-Hochberg False discovery rate correction of multiple testing. We also add a test for the assumption of joint nullity of impacts on all the coefficients.

Table 9 shows the results of this analysis. For each family of outcomes that we consider, we first show the coefficient and its standard error and then the p-values, first uncorrected and then adjusted for multiple testing. The last line of the table shows the result of the p-value of the joint test. As can be seen from the table, the adjustment of p-values makes a substantial difference. Second, to side-step the issues of prices, we present results both on the number of assets of each class, and on their value (priced as we did price them). The qualitative conclusion is the same with valued assets or just with counted assets. When considering all the assets jointly, we reject the null of no effect.

Last, (as would be expected) there is no impact on the aggregate of "big assets", aggregating vehicles, tractors and reapers. The coefficient and standard error, however, is quite large. This shows why introducing these items into computation simply adds noise to the measure and reduces the ability to detect an impact if there is in fact one. This is precisely why we chose not to introduce them. Note that with an approach like Kling et al. (2007), this item would have had no influence on the coefficients and standard errors (as the standardization would have neutralized its large variance.)

In Table 10 we apply the same methodology to items entering the definition of sales and expenses. We reject the assumption of no effect for several items, even after the correction of multiple testing. We also reject the null of no effect when considering all items jointly for sales and expenses respectively.

To summarize:

(1) Using a different set of prices to value assets actually delivers the same result as what we proposed.

(2) Introducing tractors and reapers in the list of assets leads to large coefficients and larger standard errors.

(3) Looking at basic items of assets one by one shows a clear and logical pattern, and suggests that jointly there is in fact an increase in asset ownership.

C.5. Sampling strategy

In this section BGMR criticize the sampling strategy followed by Crépon et al. (2015) to measure the impact of microcredit access at the village level. The strategy consists of selecting the top quartile of households with high predicted probability to borrow, plus five households randomly selected from the rest of the village. They also criticize the way this strategy has been implemented. They finally propose alternative estimates of treatment effects based on the selection of a specific sample within the main sample of Crépon et al. (2015) and the use of an alternative weighting scheme.

C.5.1 Preparatory and baseline survey, and changes in household composition.

A first claim of BGMR is that the data collected in the preparatory survey used by Crépon et al. (2015) to predict who will borrow are not consistent with the data collected at the baseline survey.

A second comment is about the changes in household composition between the baseline and endline survey.

Comment:

It is known that data collected using different survey instruments may vary, even if they are administered within a small interval of time. 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. Households tend to spontaneously include individuals that do not respond to the exact definition of a member (21% of households have declared a larger number of household members in the preparatory survey compared to the baseline household survey. The opposite is less frequent (10% of baseline surveyed households)). It is important thus to note that for 69% of households' data the two measures are identical, and for 88% of the households the two measures are within 2 members of one another.

There might also be differences in the composition of households between baseline and endline surveys. At the time of the administration of the endline survey, printed sheets containing the full list of household members, their age, gender and relationship to the household head were distributed to enumerators together with a survey plan. Surveyors had to copy the full list of baseline members in the endline questionnaire (question A2), respecting the member ID assigned at baseline for each of them, and ask if additional members had joined the household. They were also asked to verify, update and correct data from baseline included in the household identification sheet (questions A3 - relationship to the household head, A4 - gender, and A7 - age) as well as record each household member's condition of residency at that time (question A5 - condition of residency). There could nevertheless be true changes in household composition, or different reporting. That is certainly standard in every panel data collection (RCT or not).

But again, as discussed before, we do not need to control for baseline covariates to get consistent estimates, and even controlling for imperfect measures would not introduce bias (due to the randomization). Thus, the introduction of accurate information on household composition at baseline is not a key step in our identification strategy.

C.5.2 Contradictions in sampling scores used as sampling criteria

In the introduction of section 6.3 of BGMR, two sentences (p28) reproduced below suggest that the authors miss the main purpose of the Crépon et al. (2015) paper, and give the impression that they did not fully understand the various steps of the analysis.

"The cornerstone of this RCT protocol and the corresponding article's identification strategy is the household propensity to borrow, which was evaluated by scores."

"All average treatment effects estimated by CDDP (Tables 2 to 7) were calculated for the "high" and "very high" propensity to borrow subsamples and presented as the treatment-on-the-treated (TOT) impact. The analysis of the entire sample ("low", "high" and "very high" propensity groups) is presented as the intention-to-treat (ITT) impact."

Second, BGMR claim, showing boxplots of the distribution of scores, that there is little association between estimated scores and actual borrowing.

Comment:

It seems important to clarify what Crépon et al. (2015) do and what the different estimates produced are intended to measure. The main aim of the study is to answer the following question: does offering microcredit have an impact on households' economic activities and living conditions? Note that the experiment was launched at a moment (in 2006-2007) where some of the first preliminary results of other microcredit impact evaluations (Banerjee et al., 2015) showed fairly limited impacts (as published in the AEJ: Applied special issue). Given those findings, our objective was to conduct the most informative study possible with the highest possible power to detect an effect. With a place-based RCT, it is not possible to compare microcredit borrowers to non-borrowers, since we do not know who would have borrowed.

In this paper, we developed an innovative strategy to maximize the chances of detecting an effect of microcredit *if such an effect did in fact exist* and to provide an estimate of microcredit availability in the village for the average person in the village. What BGMR misunderstood as an error in sampling was actually one of the methodological innovations of the paper.

First, to test whether a microcredit offer has an impact on anyone with the maximum power possible, the idea of the strategy was to select a sample for which the expected take-up to Al Amana was the highest, based only on baseline covariates. To this end, we conducted a very short survey on a random set of households at the village level and then computed for each of these household a score of whether or not they are a "likely borrower". The prediction model had initially been computed on the first few villages to be included, and was adapted and improved with the progressive expansion of the experiment and the inclusion of more villages. This is why three scores were successively estimated and used to include households in the sample. Note that these scores are used similarly in the treatment and the control group. As all the other studies in the AEJ: Applied special issue, we present reduced-forms Intention-To-Treat estimates (ITT) on a sample that was "high probability to borrow" (as clearly explained in section 3 and the footnotes to each table). Those estimates correspond to the result of the regression of the outcome variable we consider on the assignment to treatment (in a village where Al Amana is available) for this population of "possible borrowers". We are very clear that this is not the effect on the average person in the village (nor is it the effect for the average borrower since take up of microcredit is low in this sample; this is an ITT, not a TOT). This is our best effort to maximize power for testing our null hypothesis H_0: microcredit has no effect.

Second, to be able to document *village level impacts* of introducing microcredit on the average households, we also collected data on a random set of households in the village. Using an appropriate

weighting scheme (giving more weights to these households) we are able to identify the impact of Al Amana at the village level (results are displayed in panel B, Table 8 of Crépon et al., 2015). These estimates are probably the ones that are most comparable to the other RCTs that randomized at the place-base level. As we discuss in section 4.1, the impact on formal credit goes from 17 to 13, showing that our prediction model was not very good. Nevertheless, the point estimates are generally smaller: the impact on assets and profits disappears. The effect of sales is still significant but half the size.

Finally, we are also interested in measuring the impact of microcredit on households who took-up a loan from Al Amana. This is the instrumental variable estimate where taking-up a loan (our endogenous variable) is instrumented with the random assignment variable (the instrument). This estimate only makes sense if there are no externalities. To check for externalities, we predict the probability to borrow in the treatment villages based on the baseline variables, and we impute for each household a "predicted probability to borrow". Panel C in Table 8 shows the coefficient of the treatment dummy (effect on those unlikely to borrow) and the interaction between probability to borrow and treatment. Since the coefficients on the treatment dummy are small and insignificant, we go ahead and, in Table 9, use treated village as an instrument for client to compute the LATE (or TOT). We emphasize that these results are suggestive, since there may be externalities that we were not able to pick up due to noise, for example, and would invalidate the IV (moreover, we ideally could have used some method robust to overfitting to predict the probability to borrow, since the model is estimated at endline).

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.

The second claim we reproduce above is factually incorrect. 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. The effect of the "Al Amana village dummy" on those households is never presented as a "Treatment on the Treated" estimator in the paper, but as the ITT on this group. The ITT on the average village person is estimated in Table 8, Panel B. Table 9 provides the TOT of "taking any credit" under the assumption of no externality.

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. This is why we kept refining the score and did not stick to the initial one[14]. Even with this

[14] The T and rank sum test they show in their Table 16 are about the link between the score and being a borrower. It is as if you regress the prediction of y on y. This is quite surprising as a procedure. Even if surprising, it shows a negative

effort, the prediction was not very good, as we also emphasize. All this does, however, is give us lower power to detect effects in our base sample (used in Table 2-7).

C.5.2.1 Supplementary analysis

In Chernozhukov at al. (2019) we apply a machine learning method to discover heterogeneity in this data set, using a double machine learning method that is robust to overfitting. We find significant heterogeneity in loan take up, although most of it is accounted for by the pair dummies. There are very few household characteristics that emerge as robust predictors of take up (age of the household head is one of them). This confirms that using observable characteristics to predict take-up of loans is a difficult exercise: take-up of microcredit at the individual level seems hard to predict. The fact that we struggled with it during the experiment (with much less rich data since we had to rely on the short form survey) was therefore not accidental. What this means is that our effort to get a sample with high probability to borrow was less effective than we hoped for, and may not have improved our power very much. This does not affect the validity of our conclusion for Tables 2-7, as well as Panels A and B of Table 8. Unfortunately, this ML analysis also casts some doubt on our ability to use an observable determinant of credit take-up to separate direct effects from externalities (Panel C of Table 8, Crépon et al., 2015).

To summarize:

(1) BGMR rediscover an issue that we already highlight in the paper: predicting who borrows from microcredit institutions on the basis of observables is very difficult.

(2) They misunderstand the implication it has for our analysis. The ex-ante prediction is never used in the analysis, and the endline sample is what is. Since the weights are known, they can be used to recover average effects.

(3) The sampling represents our best effort to reject H_0: microcredit has no effect even on those who are most likely to take it up. A more precise prediction would have given us more power.

(4) The external validity to a less specific sample is ensured by our care to select 5 households per village in the group that was not likely to borrow, which allows us to reconstruct average village level estimates (presented in Table 8).

C.5.3 Inadequate weighting procedure for ITT and IV calculation

BGMR criticize our weighting procedure to measure impacts at the village level and find it to be over-complicated. For example, they criticize our procedure to limit weights: "*The probability predicted was inversed and the inversed probability censored to 10. This means that a null weight was attributed to any household with an estimated sampling probability below 0.1, such that this household was not taken into account into the weighted ITT and IV estimation".*

coefficient for score 1 (they just comment on the p-value, not on the sign!), thus it again simply tells us that we reacted well when in the field by deciding to adapt our score procedure.

Comment:

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 also criticize us for not accounting for differences in village sizes in the estimation. It seems a matter of judgement, not a mistake. Our procedure leads to weighted results which are representative of village-level impacts at the level of a village rather than population-level impacts in the sample. The results of using a different set of weights are not shown in BGMR so it is not clear whether this has an impact.

C.5.4 Results with a "consistent panel" sample and correcting some coding and measurement errors

Last, BGMR produce alternative results where they restrict the sample to keep only households for whom there is data at baseline and endline and for whom, in addition, the household composition between baseline and endline are broadly consistent. They also add a broader set of covariates as controls in the regression. They claim that estimates are not the same. BGMR also criticize the fact that we do not account for the differences associated with the different scores used to select households.

Comment:

BGMR make a number of incorrect decisions in their "consistent panel":

(1) They discard households which don't have a "compatible" age or gender composition at endline compared to baseline. This introduces sample selection.

(2) They "correct" the specifications in ways we commented above (adding covariates, mixing high and low probability clients without proper weighting procedures)

(3) They exclude 1,400 households that were added at endline. But the households were added at endline because both the baseline data and the take-up data made it clear that the initial power calculations were insufficient. **There is absolutely no reason to throw this data away!**

The most remarkable fact, though, is that after all this effort, rows 1 and 2 of Table 17 have point estimates that are virtually identical. The one thing that has changed is that the standard errors are larger. This makes sense since the sample has 1666 fewer observations (34%).

C.6 External validity

BGMR question the external validity of the findings of the paper. First, they mention that the *"inconsistent scoring system [...] skewed the representativeness of the baseline sample towards a population subset".* Second, they compare the level of consumption in our survey and the ones from the National Living Standard surveys in rural Morocco and find differences.

Comment:

As we mentioned, we also selected a random sample of five households per village to obtain results that are representative for our sample.

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.

C.6.1 Supplementary analysis (Meager, 2018, 2019)

The question of external validity in an RCT is whether or not our estimates are indicative of what we would find in a different context. Meager (2018, 2019) reanalyzes our data and that of six other experiments to assess whether the impacts appear to be heterogenous from site to site, by estimating a Bayesian Hierarchical model with all the data. She finds that there is only a moderate degree of heterogeneity across all studies. 60% of the observed heterogeneity across studies stems from sampling variation, and the rest from truly different effects.

That said, the Morocco results tend to be more positive than those of the other studies. Under the assumption of her analysis (basically the idea that the true treatment effect is drawn from a standard normal), it is possible to obtain a BHM posterior and a 95% interval of the "true" treatment effect on each site, that incorporate what we know from other sites. The conclusion is similar, although the estimates "shrink" towards zero: the posterior mean is positive for profits, business revenues, and business expenditures. The 95% interval includes zero for profit and expenditures, and excludes zero for revenues. As we had found, it is very close to zero for consumption, durable consumption, and temptation goods.

The Morocco site is the only one that seems to find some moderate impacts on business revenues (the impact on profits were always marginally significant) even after shrinking the estimates. It was published as part of a group of papers that came to the collective conclusion that microcredit was not transformational, something that Meager confirms in her analysis.

D. Conclusion

To conclude, we mainly reproduce Table 17 in BGMR, which presents some of the "headline results". Table 18 has more modifications.

Table 17: Replicated impact estimates correcting some measurement, coding and sampling errors

	Assets	Sales and home consumption	Expenses	Of which: Investment	Profit
For memory: initial CDDP results	1,448** (658)	6,061*** (2,167)	4,057** (1,721)	-224 (223)	2,005* (1,210)
Consistent panel and some error corrections	1,277* (767)	5,990** (2,680)	3,815** (1,893)	-5.46 (148)	2,175 (1,722)

Source: Our reproduction of CDDP Table 3 with R using the same raw data, resampling for a consistent panel and correcting the coding and measurement errors listed in Section 5: omission of credits from other MFIs in total access to credit; omission of credits that matured before the survey in the variable; omission of agricultural assets in the total of assets owned by households; erratic prices used to appraise agricultural assets; livestock assets excluding non-existent units; business earnings omitted some business sales; confusions between prices before, during and after harvest to appraise agricultural sales and consumption; and inconsistent amortisation rules for agricultural investments. The sample includes 3,268 households interviewed both at baseline and at endline and which member gender and age composition is compatible between baseline and endline. 0.5 percent of observations are trimmed using the method applied by CDDP at baseline for Table 3. Coefficients and standard errors (in parentheses) from an OLS regression of the variable on a treated village dummy, controlling for strata dummies (paired villages), number of household members, number of adults, head age, does animal husbandry, does other non-agricultural activity, had an outstanding loan over the past 12 months, HH spouse responded to the survey, and other HH member (excluding the HH head) responded to the survey and variables specified below. Standard errors are clustered at the village level.

The footnote in the table reveals the extent of specification searching that went into this exercise. BGMR left no stone unturned to see if our results could somehow change. And yet, after they remove 1666 observations, they find point estimates that are virtually identical, and standard errors, which, unsurprisingly, are larger. The effect on profit, which was marginally significant, is now completely insignificant, but this is due to larger standard errors.

Remarkably, they go on to conclude: *"We see that focusing the analysis on this consistent panel yields different results"*. In our opinion, this discrepancy between this sentence and the table that it is supposed to describe fully captures the spirit of their exercise. The objective was never to find out the truth, but mainly to show that our paper was wrong. When the data did not cooperate, that did not stop them from making the point anyway.

We are very happy for people to re-analyze our data and probe robustness. Both the AEA and J-PAL, where we play a role, have been leaders in transparency in economics. Anyone can re-analyze our data starting from the raw data and the full set of well documented codes.

In fact, Meager's work and our own re-analysis (in this document) convinced us that the results on profits are fragile, driven by very large impacts on the higher quantiles that may or may not be externally valid. We also never advocated for expansion of microcredit based on the Morocco results. In fact, we show no effect on consumption. These results were published as part of a group of papers that shows that the effect of microcredit is very moderate.

But this paper is the example of what a re-analysis should not be. The combination of misunderstanding, false assertions, and specification searching, culminating in the blatant mischaracterization of the paper's own results reflects badly on the authors.

We hope that it will not discourage others from continuing to analyze our data or others'.

References

Athey, Susan, and Guido W. Imbens. 2017. "The Econometrics of Randomized Experiments." In *The Handbook of Field Experiments*, Abhijit Banerjee and Esther Duflo (eds.), Volume 1: 73-140.

Athey, Susan, and Guido W. Imbens. 2018. "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption." Working Paper.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics,* 7(1): 1-21

Bédécarrats, Florent, Isabelle Guérin, Solène Morvant-Roux, and Françious Roubaud. 2018. "Verifying the internal validity of a flagship RCT: A review of Crépon, Devoto, Duflo and Pariente (American Economic Journal: Applied Economics, 2015)." Document de Travail UMR DIAL.

Bédécarrats, Florent, Isabelle Guérin, Solène Morvant-Roux, and Françious Roubaud. 2019. "Estimating microcredit impact with low take-up, contamination and inconsistent data. A review of Crépon, Devoto, Duflo and Pariente (American Economic Journal: Applied Economics, 2015)." *International Journal for Reviews in Empirical Economics*, 3: 1-53.

Bédécarrats, Florent, Isabelle Guérin, Solène Morvant-Roux, and Françious Roubaud. 2019. "Lies, damned lies, and RCT : une expérience de J-PAL sur le microcrédit rural au Maroc." Working Paper DT/2019/04, DIAL (Développement, Institutions et Mondialisation).

Belloni, A., V. Chernozhukov, and C. Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies*, 81(2): 608-650.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments." Working Paper.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté. 2015. "Estimating the Impact of Microcredit on Those who Take it Up." *American Economic Journal: Applied Economics,* 7(1): 123-150.

Deaton, Angus, and Nancy Cartwright. 2016. "The Limitations of Randomised Controlled Trials." VOX: CEPR's Policy Portal. https://voxeu.org/article/limitations-randomised-controlled-trials.

de Mel, Suresh, David J. McKenzie, and Christopher Woodruff. 2009. "Measuring microenterprise profits: Must we ask how the sausage is made?" *Journal of Development Economics*, 88(1): 19-31.

Devoto, Florencia, Esther Duflo, Pascaline Dupas, William Parienté, and Vincent Pons. 2012. "Happiness on Tap: Piped Water Adoption in Urban Morocco." *American Economic Journal: Economic Policy,* 4(4): 68-99.

Giordano, Ryan, Tamara Broderick, Rachael Meager, Jonathan Huggins, and Michael Jordon. 2016. "Fast robustness quantification with variational Bayes." Work in progress.

Imbens, G. W. 2015. "Matching methods in practice: Three examples." *Journal of Human Resources,* 50(2): 373-419.

Imbens, G. W., and D.B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kling, Alexander, Andreas Richter, Jochen Russ. 2007. "The interaction of guarantees, surplus distribution, and asset allocation in with-profit life insurance policies." *Insurance: Mathematics and Economics*, 40(1): 164-178.

Koenker, Roger, and Gilbert Bassett Jr. 1978. "Regression Quantiles." *Econometrica,* 46(1): 33-50.

Koenker, Roger, and Kevin K. Hallock. 2001. "Quantile Regression." *Journal of Economic Perspectives*, 15(4): 143-156.

Ludwig, J., S. Mullainathan, and J. Spiess. In preparation. *Machine learning tests for effects on multiple outcomes*. Princeton University Press.

Meager, Rachael. 2018. "Aggregating Distributional Treatment Affects: A Bayesian Hierarchical Analysis of the Microcredit Literature." Working Paper.

Meager, Rachael. 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.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics,* 134(2): 557-598.

Figure 1: Sensitivity of Original Estimates to BGMR robustness checks

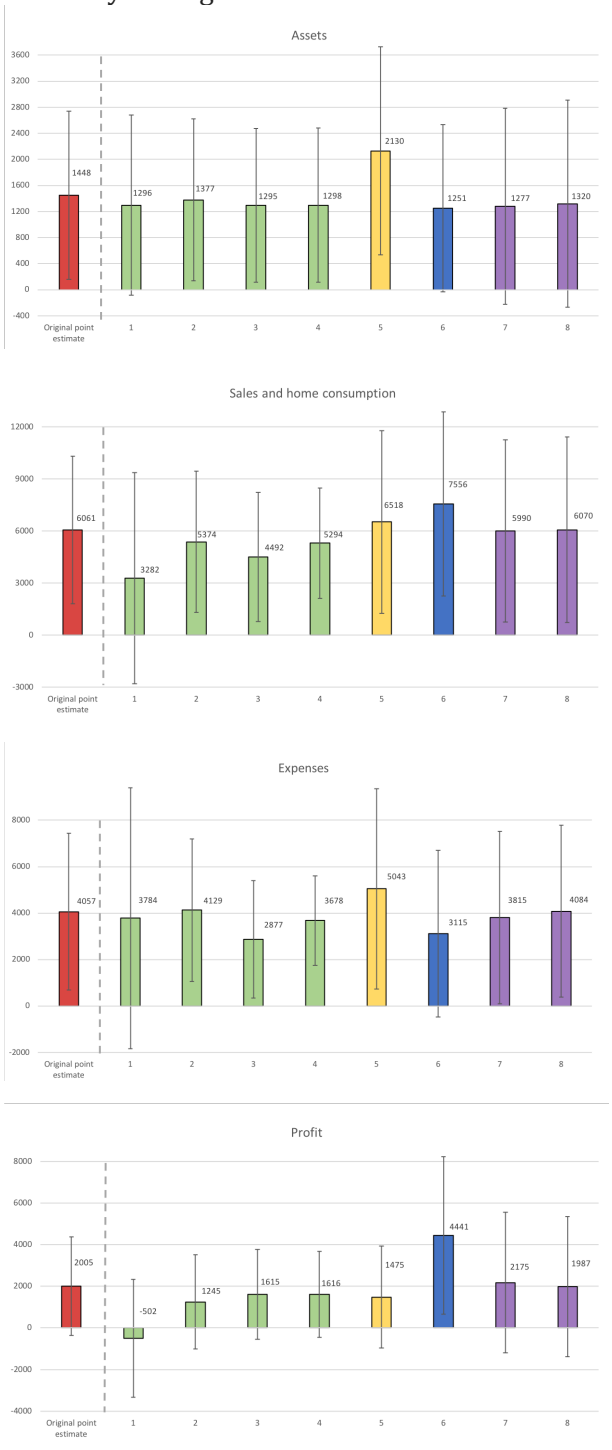

Figure 2A: BGMR point estimates of profits and assets, trimmed at 2%, 3%, and 5%

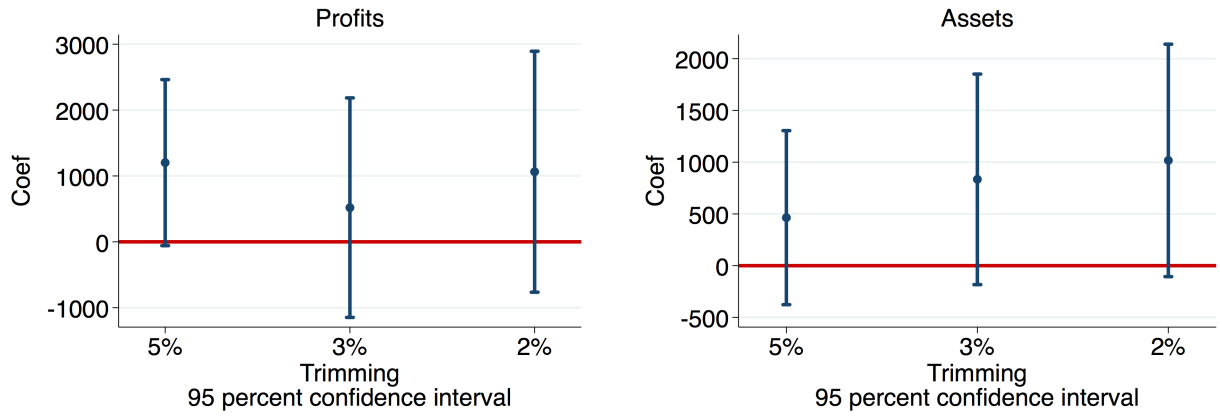

Figure 2B: CDDP quantile regression estimates of profits and assets (ITT)

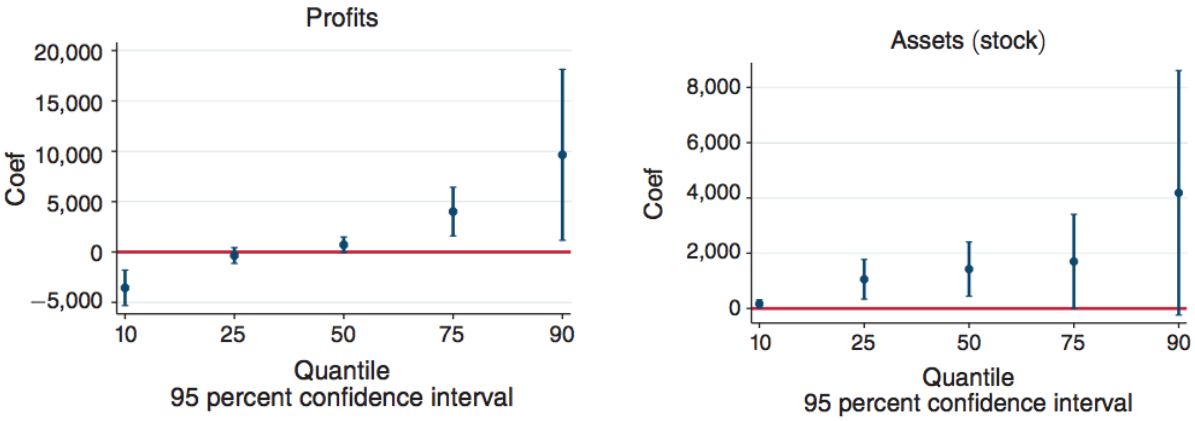

Figure 2C: Meager quantile regression estimates of profits, no pooling and partial pooling (USD PPP per two weeks)

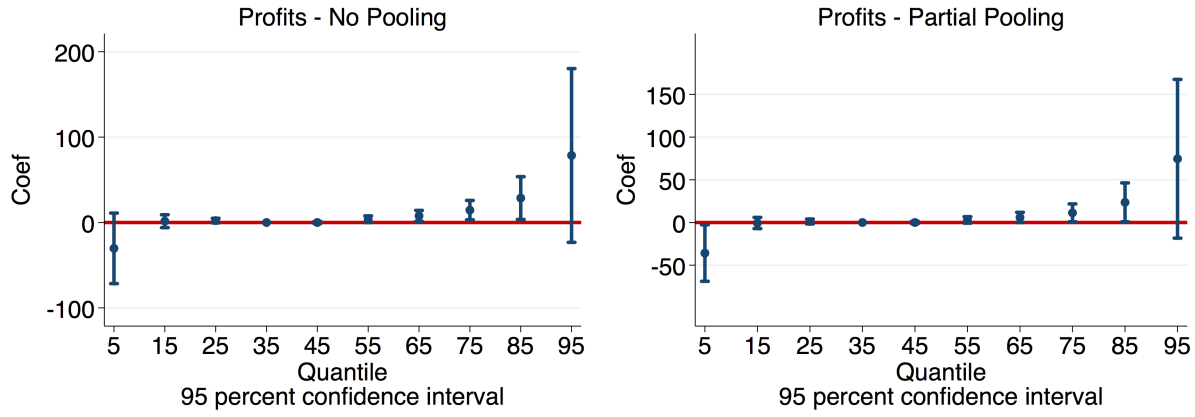

Table 1. Permutation and Rank tests for the main outcome variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Assets (stock)	Sales and home consumption	Expenses	Profit	Income from day labor/ salaried	Weekly hours worked by HH members ages 16-65		Monthly HH consumption (in MAD)	Joint Test
						self-employment	outside		
Panel A. Results from Crépon et al. (2015)									
Treated village	2,085	6,232	4,014	2,218	-1,079	1.48	-2.95	-19	
	(693)	(2362)	(1855)	(1234)	(507)	(1.51)	(0.99)	(49)	
p-value	0.003	0.009	0.032	0.074	0.035	0.327	0.003	0.690	
Panel B. P-values from permutation tests									
Mean	0.035	0.069	0.144	0.220	0.130	0.506	0.036	0.784	0.042
Deciles	0.023	0.054	0.151	0.028	0.321	0.422	0.094	0.615	0.027
Panel C. P-values from rank test									
Van Elteren	0.003	0.011	0.015	0.290	0.163	0.174	0.006	0.662	0.000

Notes: Data source: Endline household survey. Observation unit: household. Sample includes households with high probability-to-borrow scores. All panels include sample after 0.5% trimming of observations. Coefficients and standard errors (in parentheses) in Panel A are from Table B7 of Crépon et al. 2015. Panel B presents results of permutation tests. The line "mean" considers the absolute value of the weighted mean of the difference between the mean value of the variable in treatment and control villages. The p-value shown is the share of assignments of villages to pseudo treatment and control within each pair (out of 5000 pseudo assignments) that produce a statisitic larger than the one obtained with the true assignment. The line "deciles" does the same thing but considers each decile of the distribution of the variable in each village, instead of the mean, and then aggregates the statistics obtained for each of the nine deciles. Panel C presents the results of the so-called "Van Elteren test", which is a version of the Mann-Whitney ranksum test adapted to clustered designs such as what we have.

Table 2. Comparison of results from Crépon et al. (2015) and from selecting baseline covariates using a double post lasso procedure

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Client Al Amana - Admin data	Assets (stock)	Sales and home consumption	Expenses	Profit	Income from day labor/ salaried	Weekly hours worked by HH members aged 16-65		Monthly HH consumption (in MAD)
							self-employment	outside	
Panel A. Results from Crépon et al. (2015)									
Treated village	0.167***	1,448**	6,061***	4,057**	2,005*	-1,050**	0.6	-3.0***	-46
	(0.012)	(658)	(2,167)	(1,721)	(1,210)	(478)	(1.3)	(1.0)	(47)
Panel B. Results using a double lasso procedure to select baseline covariates									
Treated village	0.166***	1,616**	6,663***	4,585***	2,097*	-904**	0.6	-2.7***	-40
	(0.012)	(641)	(2,157)	(1,737)	(1,223)	(458)	(1.3)	(0.9)	(44)
Observations	4,934	4,934	4,934	4,934	4,934	4,934	4,918	4,918	4,924
Control mean	0	15984	30450	21394	9056	15748	40.61	30.40	3057

Notes: Data source: Endline household survey. Observation unit: household. Sample includes households with high probability-to-borrow scores. All panels include sample after 0.5% trimming of observations. Panel A: coefficients and standard errors (in parentheses) from an OLS regression of the variable on a treated village dummy, controlling for strata dummies (paired villages) and variables specified below. Panel B: results of the estimates of the treatment effects when adding to the regression a set of control variables selected following the the double post lasso procedure of Belloni et al. 2014. All panels: standard errors are clustered at the village level. ***, **, * indicate significance at 1, 5 and 10%.

Table 3. Take-up: alternative measures

	(1) Client Al Amana - Admin data	(2) Client Al Amana - Admin data	(3) Client Al Amana - Survey data	(4) Client Al Amana
	Crépon et al. (2015)	*Updated*	*Crépon et al. (2015)*	*Upperbound*
Credit access†				
Treated village	0.167	0.176	0.089	0.182
	(0.012)***	(0.013)***	(0.010)***	(0.014)***
Observations	4934	4934	4934	4934
Control mean	0.000	0.008	0.022	0.022

Notes: Data source: Columns 1-2: Al Amana administrative data. Column 3: Endline household survey. Column 4: Al Amana administrative data & Endline household survey. Observation unit: household. Sample includes households with high probability-to-borrow scores surveyed at endline, after trimming 0.5% of observations (3,525 that were administered both the full baseline and endline household survey, plus an additional 1,409 households that were only administered the full endline survey). Coefficients and standard errors (in parentheses) are from an OLS regression of the variable on a treated village dummy, controlling for strata dummies (paired villages) and variables specified below. Standard errors are clustered at the village level. ***, **, * indicate significance at 1, 5 and 10%. Control variables include: number of household members, number of adults, household head age, does animal husbandry, does other non-agricultural activity, had an outstanding loan (revised to include loans from other MFIs as well as loans with missing source but with a declared loan amount), HH spouse responded to the survey, and other HH member (excluding the HH head) responded the survey.

† Column 1-2: dummy variable equal to 1 if the households had borrowed from Al Amana over the two years prior to the survey. Column 3: dummy variable equal to 1 if the household had an outstanding loan from Al Amana over the 12 months prior to the survey. Column 4: dummy variable equal to 1 if the household borrowed from Al Amana over the two years prior to the survey, based on both Al Amana admnistrative data & endline household survey.

Table 4. Externalities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Client Al Amana - Admin data	Assets (stock)	Sales and home consumption	Expenses	Profit	Income from day labor/ salaried	Weekly hours worked by HH members ages 16-65		Monthly HH consumption (in MAD)
							self-employment	outside	
Panel A. Top and bottom 30% predicted using same client definition as Crépon et al. (2015)									
Treated village X Low Predicted Propensity to Borrow	0.015***	1,612	647	1,013	-366	-2,453***	-1.4	-6.2***	82
	(0.003)	(1,132)	(2,701)	(1,737)	(1,734)	(795)	(1.3)	(1.6)	(62)
Treated village X High Predicted Propensity to Borrow	0.363***	1,033	15,774***	10,171***	5,603**	-2,113***	2.9	-7.0***	-93
	(0.011)	(1,296)	(4,154)	(3,555)	(2,452)	(692)	(2.3)	(1.8)	(94)
Observations	3,315	3,315	3,315	3,315	3,315	3,315	3,303	3,303	3,307
p-value: T X Low PTB = T X High PTB	0.000	0.734	0.002	0.022	0.049	0.746	0.106	0.727	0.113
Panel B. Top and bottom 30% predicted using an improved administrative measure of clients									
Treated village X Low Predicted Propensity to Borrow	0.016***	1,124	1,931	2,262	-331	-1,988**	-1.0	-4.5***	88
	(0.003)	(1,060)	(2,851)	(2,000)	(1,686)	(805)	(1.3)	(1.7)	(69)
Treated village X High Predicted Propensity to Borrow	0.353***	1,520	16,585***	9,763***	6,822***	-2,101***	2.2	-6.5***	-87
	(0.013)	(1,167)	(3,904)	(3,099)	(2,308)	(738)	(2.7)	(1.7)	(99)
Observations	3,315	3,315	3,315	3,315	3,315	3,315	3,304	3,304	3,307
p-value: T X Low PTB = T X High PTB	0.000	0.802	0.002	0.037	0.014	0.917	0.288	0.395	0.136
Panel C. Top and bottom 30% predicted using an upper bound measure of clients (administrative records + survey self-declaration)									
Treated village X Low Predicted Propensity to Borrow	0.019***	222	4,102	4,458**	-356	-2,063***	-1.3	-4.8***	107
	(0.003)	(1,085)	(2,629)	(1,852)	(1,683)	(786)	(1.4)	(1.7)	(69)
Treated village X High Predicted Propensity to Borrow	0.350***	1,644	13,994***	6,946**	7,047***	-2,031***	2.9	-5.5***	-57
	(0.016)	(1,274)	(4,220)	(3,446)	(2,199)	(730)	(2.5)	(1.8)	(108)
Observations	3,315	3,315	3,315	3,315	3,315	3,315	3,304	3,304	3,307
p-value: T X Low PTB = T X High PTB	0.000	0.398	0.047	0.526	0.009	0.976	0.146	0.790	0.180
Panel D. Top and bottom 30% predicted using a client measure from survey self-declaration									
Treated village X Low Predicted Propensity to Borrow	0.035***	-1,440	1,211	2,308	-1,096	-1,209*	-1.6	-4.4***	-62
	(0.006)	(999)	(2,630)	(1,751)	(1,680)	(679)	(1.3)	(1.5)	(73)
Treated village X High Predicted Propensity to Borrow	0.296***	2,598**	5,129	-809	5,938**	-1,914**	1.1	-5.4***	-233**
	(0.018)	(1,223)	(4,835)	(3,814)	(2,497)	(832)	(2.7)	(1.8)	(91)
Observations	3,315	3,315	3,315	3,315	3,315	3,315	3,301	3,301	3,307
p-value: T X Low PTB = T X High PTB	0.000	0.011	0.483	0.468	0.023	0.523	0.368	0.665	0.143

Notes: Data source: Endline household survey. Observation unit: household. All panels: sample includes both households with high probability-to-borrow scores and households picked at random, but only those in the top 30% and in the bottom 30% of the predicted propensity to borrow (PTB) distribution. All panels include sample after 0.5% trimming of observations. All panels: coefficients and standard errors (in parentheses) are from an OLS regression of the variable on a treated village dummy interacted with a dummy equal to 1 if HH predicted propensity to borrow is in the 0-30th percentile of the PTB distribution (Low Predicted PTB), on a treated village dummy interacted with a dummy equal to 1 if HH predicted PTB is in the 70-100th percentile of the PTB distribution (High Predicted PTB) and on a dummy equal to 1 if HH predicted PTB is in the 0-30th percentile of the PTB distribution (not shown), controlling for strata dummies (paired villages) and variables specified below. All panels: standard errors are clustered at the village level. ***, **, * indicate significance at 1, 5 and 10%. Same controls as in Table 8 of Crépon et al. (2015).

Table 5. Summary Statistics

	Obs	Control Group			Treatment - Control	
		Obs	Mean	St. Dev.	Coeff.	*p-value*
Access to credit:						
Had an outstanding loan (original)	4465	2266	0.257	0.437	0.053 ***	*0.007*
Had an outstanding loan revised†	4465	2266	0.268	0.443	0.049 **	*0.011*

Notes: Data source: Baseline household survey. Unit of observation: household. Sample includes all households surveyed at baseline. ***, **, * indicate significance at 1, 5 and 10%.

† The definition of the variable was revised in order to include loans from other MFIs and 17 loans from unknown sources but with a positive loan amount.

Table 6. Robustness: main effects controlling by different baseline variables on access to credit

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Client Al Amana - Admin data	Client Al Amana - Survey data	Assets (stock)	Sales and home consumption	Expenses	Profit	Income from day labor/ salaried	Weekly hours worked by HH members ages 16-65		Monthly HH consumption (in MAD)
								self-employment	outside	
Panel A. Same Control Variables as in Crépon et al. (2015)										
Treated village	0.167	0.090	1,448	6,061	4,057	2,005	-1,050	0.6	-3.0	-46
	(0.012)***	(0.010)***	(658)**	(2,167)***	(1,721)**	(1,210)*	(478)**	(1.3)	(1.0)***	(47)
Observations	4,934	4,934	4,934	4,934	4,934	4,934	4,934	4,918	4,918	4,924
Control mean	0.000	0.022	15,984	30,450	21,394	9,056	15,748	40.6	30.4	3,057
Panel B. Adding outstanding loans from other MFIs										
Treated village	0.167	0.089	1,455	6,098	4,090	2,008	-1,066	0.6	-3.0	-46
	(0.012)***	(0.010)***	(655)**	(2,168)***	(1,722)**	(1,210)*	(478)**	(1.3)	(1.0)***	(47)
Observations	4,934	4,934	4,934	4,934	4,934	4,934	4,934	4,918	4,918	4,924
Control mean	0.000	0.022	15,984	30,450	21,394	9,056	15,748	40.6	30.4	3,057
Panel C. Adding outstanding loans from other MFIs and matured loans from any source in the past 12 months										
Treated village	0.167	0.090	1,484	6,200	4,193	2,006	-1,053	0.6	-3.0	-44
	(0.012)***	(0.010)***	(654)**	(2,170)***	(1,733)**	(1,208)*	(478)**	(1.3)	(1.0)***	(47)
Observations	4,934	4,934	4,934	4,934	4,934	4,934	4,934	4,918	4,918	4,924
Control mean	0.000	0.022	15,984	30,450	21,394	9,056	15,748	40.6	30.4	3,057

Notes: Data source: Endline household survey. Observation unit: household. Sample includes households with high probability-to-borrow scores after 0.5% trimming of observations. All panels: coefficients and standard errors (in parentheses) from an OLS regression of the variable on a treated village dummy, controlling for strata dummies (paired villages) and variables specified below. All panels: standard errors are clustered at the village level. ***, **, * indicate significance at 1, 5 and 10%. Panel A controls include: number of household members, number of adults, household head age, does animal husbandry, does other non-agricultural activity, had an outstanding loan, HH spouse responded the survey, and other HH member (excluding the HH head) responded to the survey. Panel B controls include the same controls as Panel A except for the variable "had an outstanding loan", which has been revised to include loans from other MFIs as well as loans with missing sources but a declared loan amount. Panel C controls include: same controls as Panel B except for the revised variable "had an outstanding loan", which has been redefined as "had an outstanding or matured loan over the past 12 months".

Table 7. Other loans: utility company & others

	(1)	(2)	(3)	(4)
			Of which:	
	Utility company and Other loans *Crépon et al. (2015)*	Utility company and Other loans *Revised††*	Utility company	Other loans
Credit access†				
Panel A. Full sample of households with high probability to borrow surveyed at endline (same sample as in Crépon et al. (2015))				
Treated village	0.017	0.017	0.016	0.001
	(0.017)	(0.018)	(0.017)	(0.003)
Observations	4934	4934	4934	4934
Control mean	0.157	0.165	0.150	0.016
Panel B. Restricted sample of households with high probability to borrow surveyed at both endline and baseline (same sample as in Bédécarrats et al. (2019))				
Treated village	0.037	0.036	0.037	0.000
	(0.017)**	(0.017)**	(0.016)**	(0.004)
Observations	3525	3525	3525	3525
Control mean	0.161	0.170	0.154	0.018

Notes: Data source: Columns 1-3: Endline household survey. Observation unit: household. Panel A: sample includes households with high probability-to-borrow scores surveyed at endline, after trimming 0.5% of observations (3,525 who were administered both the full baseline and endline household survey, plus an additional 1,409 households who were administered only the full endline survey). Panel B: sample includes 3,525 households who were administered both the full baseline and endline household survey. Coefficients and standard errors (in parentheses) are from an OLS regression of the variable on a treated village dummy, controlling for strata dummies (paired villages) and variables specified below. Standard errors are clustered at the village level. ***, **, * indicate significance at 1, 5 and 10%. Panel A controls include: number of household members, number of adults, household head age, does animal husbandry, does other non-agricultural activity, had an outstanding loan (revised to include loans from other MFIs as well as loans with missing source but a declared loan amount), HH spouse responded to the survey, and other HH member (excluding the HH head) responded to the survey. Panel B controls include: had an outstanding or matured loan over the past 12 months and same variables for the rest of controls. †

Column 1-3: dummy variable equal to 1 if the household had an outstanding loan over the 12 months prior to the survey.

†† Includes loans with missing source but a declared loan amount and excludes 17 loans reclassified as "other formal".

Table 8. Comparison of results from Crépon et al. (2015) and from selecting baseline covariates using a double post lasso procedure

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Al Amana - Admin data	Al Amana - Survey data	Other MFI	Other Formal†	Informal	Utility Company†	Other†	Total
Panel A. Results from Crépon et al. (2015)								
Treated village	0.167***	0.090***	-0.006	0.009***	-0.003	0.016	0.001	0.078***
	(0.012)	(0.010)	(0.004)	(0.003)	(0.007)	(0.017)	(0.003)	(0.018)
Panel B. Results using a double lasso procedure to select baseline covariates								
Treated village	0.166***	0.089***	-0.006	0.010***	-0.002	0.019	0.000	0.076***
	(0.012)	(0.010)	(0.004)	(0.003)	(0.006)	(0.017)	(0.003)	(0.018)
Observations	4,934	4,934	4,934	4,934	4,934	4,934	4,934	4,934
Control mean	0.000	0.022	0.023	0.017	0.059	0.150	0.016	0.254

Notes: Data source: Endline household survey. Observation unit: household. Sample includes households with high probability-to-borrow scores. All panels include sample after 0.5% trimming of observations. Panel A: coefficients and standard errors (in parentheses) are from an OLS regression of the variable on a treated village dummy, controlling for strata dummies (paired villages) and same variables as in Table 2 of Crépon et al. (2015). Panel B: results of the estimate of the treatment effect when adding in the regression a set of control variables selected following the the double post lasso procedure of Belloni et al. 2014. All panels: standard errors are clustered at the village level. ***, **, * indicate significance at 1, 5 and 10%.

† Revised loan variables (see section C.3.3).

Table 9. Multiple Testing: Assets

	Number of assets owned					*Value of assets owned (in MAD)*				
	Treatment - Control					Treatment - Control				
	Coeff.	se	*p-val*		*adjusted p-val*	Coeff.	se	*p-val*		*adjusted p-val*	
Milk jars	0.077	0.018	*0.000*	***	*0.001*	***	3	1	*0.000*	***	*0.001*	***
Rabbits	0.250	0.078	*0.002*	***	*0.014*	**	10	3	*0.002*	***	*0.013*	**
Draft animals	0.078	0.027	*0.004*	***	*0.023*	**	65	22	*0.004*	***	*0.022*	**
Improved cattle breed	0.118	0.052	*0.023*	**	*0.104*		828	361	*0.023*	**	*0.098*	*
Traditional plowing	0.032	0.017	*0.058*	*	*0.210*		6	3	*0.058*	*	*0.195*	
Local cattle breed	0.092	0.054	*0.094*	*	*0.227*		503	299	*0.094*	*	*0.195*	
Poultry	0.309	0.187	*0.101*		*0.227*		6	4	*0.096*	*	*0.195*	
Livestock pens	0.020	0.012	*0.101*		*0.227*		6	4	*0.101*		*0.195*	
Carts and wheelbarrows	0.033	0.021	*0.123*		*0.245*		10	6	*0.103*		*0.195*	
Goats	0.490	0.382	*0.201*		*0.356*		6	4	*0.126*		*0.214*	
Sewing and weaving machines	-0.007	0.005	*0.217*		*0.356*		196	153	*0.201*		*0.297*	
Tractors, reapers, cars and trucks	0.010	0.009	*0.280*		*0.395*		-17	13	*0.209*		*0.297*	
Other assets	0.019	0.018	*0.285*		*0.395*		-20	19	*0.294*		*0.383*	
Honey wooden hives	-0.038	0.038	*0.320*		*0.410*		850	868	*0.329*		*0.383*	
Small tools	0.089	0.094	*0.342*		*0.410*		4	4	*0.338*		*0.383*	
Oil mills	-0.002	0.003	*0.382*		*0.430*		88	125	*0.480*		*0.510*	
Other livestock	0.010	0.014	*0.480*		*0.509*		22	341	*0.948*		*0.948*	
Sheep	0.028	0.429	*0.948*		*0.948*		n.a.	n.a.	n.a.		n.a.	
Joint Test†			*0.009*			*Joint Test†*			*0.008*			

Notes: Data source: Endline household survey. Unit of observation: household. Sample includes households with high probability-to-borrow scores surveyed at endline, after trimming 0.5% of observations. ***, **, * indicate significance at 1, 5 and 10%. The table presents the estimated treatment effect on different items included in the definition of the total value of assets. The first column gives the estimated coefficient, the second the standard error. The third column provides the usual pvalue and the last one the p-value adjusted for multiple testing and controlling for the false discovery rate following the method proposed by Benjamini and Hochberg (2001). P-values are ranked and the adjusted p-value at rank r is defined as the maximum of p(i)*M/i for i larger than r, and M is the number of items in the family.

† The hypothesis tested is that all the treatment effects of the disaggregated variables are jointly zero. The test is a Wald test using an estimation of the joint variance matrix of the parameters estimated equation by equation. The asymptotic distribution is chi squared, whose number of degrees of freedom is the number of dissagregated items.

Table 10. Multiple Testing: Sales & Expenses

	Treatment - Control			
	Coeff.	se	*p-val*	*adjusted p-val*
Panel A. Sales and self-consumption over the past 12 months prior to the survey (in MAD)				
Crops: other	1325	307	*0.000* ***	*0.000* ***
Animal husbandry: other livestock	129	49	*0.010* ***	*0.043* **
Animal husbandry: cows	744	286	*0.010* **	*0.043* **
Vegetables	358	143	*0.013* **	*0.043* **
Livestock production: other	166	67	*0.014* **	*0.043* **
Animal husbandry: sheeps	618	324	*0.058* *	*0.145*
Crops: barley	442	361	*0.222*	*0.397*
Crops: wheat	460	384	*0.232*	*0.397*
Tree fruits: other	-357	312	*0.255*	*0.397*
Livestock production: milk	304	272	*0.264*	*0.397*
Livestock production: eggs	22	23	*0.334*	*0.434*
Animal husbandry: poultry	-54	58	*0.347*	*0.434*
Non-agricultural business: total sales	1693	1938	*0.384*	*0.443*
Tree fruits: olives	214	399	*0.592*	*0.612*
Crops: durum wheat	167	329	*0.612*	*0.612*
Joint Test			*0.000*	
Panel B. Expenses over the past 12 months prior to the survey (in MAD)				
Animal husbandry: feed	672	194	*0.001* ***	*0.008* ***
Agriculture: plowing	135	46	*0.004* ***	*0.024* **
Animal husbandry: transport	62	23	*0.009* ***	*0.037* **
Agriculture: seeds	151	63	*0.018* **	*0.053* *
Agriculture: rent	269	126	*0.034* **	*0.083* *
Agriculture: labor	476	241	*0.050* *	*0.100*
Animal husbandry: other expenses	377	217	*0.085* *	*0.145*
Agriculture: other expenses	95	80	*0.238*	*0.336*
Animal husbandry: fattening	213	193	*0.271*	*0.336*
Agriculture: fertilizer	35	32	*0.280*	*0.336*
Non-agricultural business: total expenses	1476	1816	*0.418*	*0.455*
Agriculture: harvest fees	53	78	*0.498*	*0.498*
Joint Test			*0.009*	

Notes: Data source: Endline household survey. Unit of observation: household. Sample includes households with high probability-to-borrow scores surveyed at endline, after trimming 0.5% of observations. ***, **, * indicate significance at 1, 5 and 10%. See Table 9.