Estimating microcredit impact with low take-up, contamination and inconsistent data.

A replication study of Crépon, Devoto, Duflo, and Pariente (American Economic Journal: Applied Economics, 2015)

Florent Bédécarrats [*] Isabelle Guérin [†] Solène Morvant-Roux [‡]
François Roubaud [§]

International Journal for Re-Views in Empirical Economics, Volume 3, 2019-3, DOI: 10.18718/81781.12

JEL: C18, C83, C93, G21, 016, 055
Keywords: RCT, microcredit, J-PAL, Morocco, internal validity, data quality, replication study

Data Availability: In this paper we use the original data from Crépon, Devoto, Duflo, and Pariente (CDDP), which are available on the AEJ:AE website. The zip-file can be downloaded at www.aeaweb.org/articles?id=10.1257/app.20130535. The R-code of this replication is available from the website of the journal www.iree.eu.

Please Cite As: Bédécarrats, Florent, Isabelle Guérin, Solène Morvant-Roux, and François 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 Re-Views in Empirical Economics*, Vol 3(2019-3). DOI: 10.18718/81781.12

[*]AFD-EVA (Evaluation unit of the French Development Agency), Paris, France, e-mail: bedecarratsf@afd.fr
[†]IRD-CESSMA (Centre for social science Studies on the African, American and Asian worlds at the French national Research Institute for Sustainable Development), Paris, France, e-mail: isabelle.guerin@ird.fr
[‡]Department of History, Economics and Society and Institute of Demography and Socioeconomics at the University of Geneva, Geneva, Switzerland, e-mail: Solene.Morvant@unige.ch
[§]IRD-DIAL (Joint Research Unit on Development, Institutions and Globalization at the French national Research Institute for Sustainable Development), Paris, France, e-mail: roubaud@dial.prd.fr
Declaration of interests: This research is not the result of a for-pay consulting relationship. One of the four authors joined the Agence Française de Développement in 2014 (that is after this RCT was conducted) and works in its evaluation unit, which is part of AFD research department. AFD research department was the main funder of the initial RCT replicated in this paper. AFD evaluation and research activities are meant to be independent from operational and financial interests of the institutions. Microfinance has very limited importance in AFD's portfolio and does not constitute a substantial financial interest for the institution. AFD is however deeply involved in the scientific and methodological debate about the appropriate methods to evaluate development interventions. The other co-authors belong to academic institutions, maintain a long standing scientific collaboration on this subject with the aforementioned AFD co-author and did not receive any funding for this replication analysis.

Received June 25, 2018; Revised February 15, 2019; Accepted February 18, 2019; Published March 13, 2019.

Abstract

We replicate a flagship randomised control 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 observation with the highest values on some variables. However, the applied trimming procedures are inconsistent between the baseline and the endline. Using identical specifications as the original paper reveals large and significant imbalances at the baseline and, at the endline, impacts on 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. We find a series of coding, measurement and sampling errors. Correcting the identified errors lead to different results. After rectifying identified errors, we still find substantial imbalances at baseline and implausible impacts at the endline. 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.

1 Introduction

Randomised control trials (RCTs) are increasingly considered as the gold standard for producing evidence on what works and what does not, and this trend is particularly strong in development economics (Bédécarrats, Guérin, and Roubaud 2017). In this field, microfinance is the sector most frequently evaluated by RCTs. J-PAL (a global research centre promoting this method for poverty reduction) posts 262 "finance" RCTs out of its 902 completed and ongoing RCTs.[1] A highlight of this undertaking was the 2015 publication of a special issue in the *American Economic Journal: Applied Economics* (AEJ:AE) featuring six RCTs on microcredit (Banerjee, Karlan, and Zinman 2015). This special issue is seen by leading RCT movement figures as the decisive contribution to settle a long-standing debate on the subject (Ogden 2017). It quickly attracted massive coverage: 2,557 citations in other scientific studies[2] and J-PAL's publication of a policy briefcase based on the six papers and drawing general conclusions for finance access strategies worldwide (Loiseau and Walsh 2015).

To strengthen the robustness of empirical research, the scientific community increasingly recommends systematic replication. A replication is a "*study whose main purpose is to determine the validity of one or more empirical results from a previously published study*" (Duvendack, Palmer-Jones, and Reed 2017: 47). Clemens (2017) defines two categories and four subcategories of tests that can be used to this effect. The first *replication test* category uses the same specifications as the original paper, focuses on the same population of interest and is expected to produce the same results. Replication tests can be divided into two subcategories. The *replication-verification* subcategory retains the same sample as the original, to ensure that the reported statistical analysis does indeed produce the same results. Its purpose is mainly to identify flawed measurements, codes, datasets, etc. The *replication-reproduction* subcategory resamples, but from the same population and with the same distribution as the original paper. This is designed to turn up sampling errors, statistical power issues and other errors found by verification. The second *robustness test* category uses different specifications to the original paper. They are not expected to produce the same results, but the results should remain consistent with the conclusions of the original paper to hold. Robustness tests can also be divided into two subcategories. The *robustness-reanalysis* subcategory alters the statistical procedures to include new recoded variables or run different types of regressions for instance. It may or may not entail resampling, but it refers to the same population of interest. The *robustness-extension* subcategory uses different data from a different population or from the same population at a different point in time, but applies the same data analysis procedure.

Replications are still seldom performed, and most of them belong to the *robustness-reanalysis* category. Sukhantar (2017) systematically reviews development economics articles published in ten top-ranking journals[3] since 2000. He finds that 71 (6.2%) of the 1,138 empirical articles studied have been the subject of replication or robustness tests in a published or working paper. This

[1] Source: The Abdul Lateef Jameel Poverty Action Lab (J-PAL) website: www.povertyactionlab.org/evaluations, visited on April 23, 2018.

[2] Source: Google Scholar citation indexes for the articles featured in this special issue, see scholar.google.fr/scholar?hl=fr&as_sdt=0%2C5&as_ylo=2015&as_yhi=2015&q=microcredit+source%3A%22American+economic+journal+applied+economics%22&btnG=, visited on April 23, 2018.

[3] *American Economic Review*, *Quarterly Journal of Economics*, *Journal of Political Economy*, *Econometrica*, *Review of Economic Studies*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *Economic Journal*, *Journal of the European Economic Association*, and *Review of Economics and Statistics*.

rate rises to 12.5% when considering solely the 120 RCTs covered in this systematic review. Yet when the scope is narrowed to reviews conducting *replication tests* (verification or reproduction), the ratio falls to just 0.20% for all empirical papers and 0.16% (only two cases) for RCTs. These rates suggest that economists generally take for granted the reliability of the data, sampling and codes of the work produced by their peers and that, when they do take an interest in challenging a publication, they focus the discussion on modelling techniques.

Replication tests can only be performed if the raw microdata is available. So in order to encourage these tests, a growing number of journals now systematically publish articles jointly with the data and analysis procedure on which they are based. The AEJ:AE data availability policy[4] states that the raw data should be made available, in particular in the case of experiments. However, in the above-mentioned special issue on microcredit, the raw data is available for just three of the six RCTs: (Crépon et al. 2015; Attanasio et al. 2015; Augsburg et al. 2015). A subset of pre-processed aggregated variables is provided in two cases (Banerjee et al. 2015; Angelucci, Karlan, and Zinman 2015), and no data is made available at all in one case (Tarozzi, Desai, and Johnson 2015). We chose to replicate the Moroccan study by Crépon, Devoto, Duflo and Pariente (hereafter referred to as CDDP). This is the most cited paper of this reproducible half of the AEJ:AE special issue on microcredit. It is also co-authored by two researchers who play a central role as standard setters at J-PAL: Crépon and Duflo (Jatteau 2016: 313). It could therefore be indicative of common RCT practices in the development field.

CDDP conducted this RCT impact evaluation with Morocco's largest microcredit institution (Association Al Amana, hereafter AAA), which was launching microcredit in rural areas not yet covered. The team took advantage of this expansion to new places to perform a RCT at area level. 162 villages were chosen around a central zone where the MFI had decided to start up new operations. The villages were then divided into 81 pairs of similar villages based on observable characteristics such as the number of households, accessibility to the centre of the community, existing infrastructure, type of activities carried out by the households and type of agricultural activities.

AAA started up operations in randomly assigned villages offering joint-liability loans to local men and women living there. The loans granted were similar to urban area loans: group loans with amounts ranging from MAD (Moroccan dirhams) 1,000 to 15,000 (USD 124 to 1,855) per group member. In March 2008, AAA launched individual loans in rural areas: housing and non-agricultural businesses were eligible for larger amounts, but with additional conditions. Most of the loans taken in these areas, however, were group loans.

Loan periods ranged from 3 to 18 months and repayments were made weekly, fortnightly or monthly excepting stockbreeding loans, which benefited from a two-month grace period. Annual interest rates ranged between 12.5% and 14.5% at the time of the study. The authors argue that there was enough distance between pairs of villages to prevent any contamination between treatment and control villages. The RCT was performed from 2006 to 2010 over four expansion periods. The baseline was conducted in four phases between 2006 and 2007.

[4] All journals from the American Economic Association, including AEJ:AE, are subject to the same data availability policy, available online: www.aeaweb.org/journals/policies/data-availability-policy. This data availability policy has remained the same since at least 2012. This is the same clause as found in this review of journal data policy: www.edawax.de/wp-content/uploads/2012/07/Data_Policies_WP2.pd.

The sample as a whole was broken down into three household categories: 1) households in the top quartile identified along the line of the propensity score (25% of households with the highest probability of taking out a microloan); 2) five randomly selected households in the three other quartiles added to this sample in each village (treatment and control); and 3) a last (third) group of 1,433 households added only at the endline by re-estimating take-up scores across the entire sample and matching with administrative data provided by the MFI. The total sample contained 4,465 households at the baseline, 92% of which (4,118) could be re-interviewed at the endline, plus the 1,433 new households added at the endline. The total sample came to 5,551 households.

The authors state that these three categories of potential borrowers capture the heterogeneity across households (borrowers versus non-borrowers) and thus enable them to assess the spillover effect on non-borrowers and "*measure the impact of microcredit expansion on the community as a whole*" (Loiseau and Walsh 2015: 3).

The main findings of the RCT on the entire population of a village are reported in Crépon et al. (2015), and Loiseau and Walsh (2015). The first finding is that demand (take-up) for microcredit was low and lower than the researchers and the partner MFI expected. While this pattern is similar to other countries such as Ethiopia, India and Mexico, the uptake rate was particularly low in Morocco (16% of eligible borrowers), despite active promotion of microcredit by AAA loan officers during the RCT.

The authors find that the programme had no impact on business start-up, but positive effects were found on a number of business-related outcome variables such as income, assets, investment and profits. Overall positive results were highly heterogeneous, meaning that some households benefited (larger business owners) while others did not (negative impact).

Heterogeneity aside, positive impacts on business earnings were offset by significant decrease in labour supplied outside the home and in salary income. Consumption across an entire village population also decreased, albeit not significantly. Lastly, in terms of empowerment, microcredit impact on two major outcome variables (education and women's empowerment) is unlikely to change women's bargaining power in rural Morocco.

The main conclusion the authors derive from their study is that the aggregate impact of microcredit should not be overestimated, as their study finds an overall fairly limited effect on the population at large, at least over a short period of time (two years).

This replication paper is structured as follows. In Section 2, we describe the data and our replication method. Section 3 discusses the trimming procedures used by CDDP and assesses their results' sensitivity to the trimming threshold. Section 4 highlights several significant imbalances at baseline and disconcerting impacts on other outcomes produced with the same specifications as CDDP. Section 5 focuses on coding and measurement errors, while Section 6 addresses sampling errors. Section 7 discusses shortcomings related to external validity and our concluding comments are found in Section 8.

2 Data and method

The data and code used by CDDP can be found on the American Economic Association website's subsection on AEJ:AE, as links on the page on this article.

The download contains three datasets, in Stata (.dta) format: the short preparatory survey (15,145 observations and 25 variables), the baseline survey (4,465 observations and 3,733 variables) and the endline survey (5,551 observations and 4,790 variables). It also includes the endline survey questionnaire, in French and English. Neither the simple preparatory survey questionnaire nor the baseline survey questionnaire is provided. Lastly, the download includes five data processing scripts, also in Stata format (.do): "Outcome construction at baseline", "Outcome construction at endline", "Analysis", "Graphs" and "Master". In the following replication, we refer to specific code sections, giving the Stata files these code sections come from (abbreviated respectively as BL, EL, AN, GR and MA), followed by the line number. For example, "BL:43" refers to line 43 of the file "Outcome construction at baseline". We also refer to specific survey questions and microdata variables, giving their code in single quotation marks. Modalities are placed in italics. For example, '*Al Amana*' and '*Zakoura*' are two possible answers to survey questionnaire question 'i3' on whom the household has borrowed from during the past year.

To ensure that our procedures are fully transparent and reproducible, we computed them using `R` (R Core Team 2018) in `RMarkdown` (RStudio Team 2018) format. We published, jointly with this paper, its source file with a .rmd extension, which contains all the scripts to access, download, import, prepare and compute the data (Bédécarrats et al. 2019)[5]. No data or figure was added outside of the script and the results, tables and figures displayed in the document are produced solely by this code.

Taking Clemens' typology (2017), our analysis includes *replication-verification*, *replication-reproduction*, and *robustness-reanalysis* tests. These tests are interdependent. Our verification turns up not only measurement errors, but also sampling errors, calling for resampling analysis. Our verification also raises concerns as to the robustness of the paper's conclusions. This was assessed by using the same specifications as CDDP, but by completing the independent variables they included in their regression to control for imbalanced variables at baseline, with other variables on which we also found major imbalances at baseline. The primary focus of this re-analysis is assessing the internal validity of CDDP published results and, if not stated otherwise, the shortcomings discussed below all refer to internal validity. Some of the issues we identified to assess internal validity also have implications for external validity, so we also discuss this in the last section of this replication paper.

Verification tests are often restricted to "push button replications", as the International Initiative for Impact Evaluation (3IE) describes them[6]: rerunning the script code provided by the authors with the same data and checking that it produces the same outputs. Here, we conducted a more exacting process, consisting of translating the analysis procedure into a different statistical language (R) to the one used by the authors (Stata). Translating the code into another programming language requires the replicators to understand the original authors' intention, design a script that

[5] Downloadable from the data archive of IREE. DOI: 10.15456/iree.2019071.090421
[6] See the "Push Button Replication Project" page from the 3IE website: www.3ieimpact.org

executes this intention (instead of simply copy-pasting), and analyse any discrepancies between replicated and original results at all stages of the data analysis process until the cause of each and every difference can be understood. We ended up refining a code where each step of data analysis is a function. Every time a coding error was identified in the original paper, this coding error was included as an optional parameter in the corresponding function. If the option is activated, the function reproduces the error made by CDDP. If it is deactivated, it produces a corrected output.

We verified data quality and sampling integrity using basic good practices for survey analysis (United Nations Statistical Division 2005), in particular to check the consistency of household composition with respect to simple criteria such as gender and age. We also verified the variation in respondents' answers to identical survey questions repeated across the questionnaires.

The original code and paper run regressions on 110 constructed dependent variables, each one built upon information contained in a number (sometime dozens) of variables from the raw dataset. These variables can be clustered into five groups: credit, self-employment activities, work, consumption and socio-economic variables. We focused here on a subset of three of these groups, namely credit, which corresponds to the treatment being evaluated, self-employment activities, which is where the main impacts have been found, and consumption, as it is used for trimming (see Section 3.1).

We first reproduced with R the analysis of the original paper to show that we did have the same data and that we had understood every detail of the analysis procedures applied by CDDP. Table 1 below reproduces some of the balance test presented in CDDP Table 1. Table 2 below reproduces the average impact estimates of the experiment on access to credit, as in CDDP Table 2. Table 3 below presents the average treatment effect on variables related to self-employment activities, which include the most significant results of this RCT, as in CDDP Table 3.

Table 1 shows that CDDP identified some small but significant imbalances at baseline: households in treatment villages have older heads, carry out more frequently animal husbandry and non-farm businesses and borrow more frequently from formal and informal credit sources. The baseline values of these imbalanced variables have been used by CDDP as controls for the regressions estimating the average treatment effects at endline, for instance Table 2 and Table 3. Table 2 suggests that the experiment worked, that is that the households in the village assigned to the treatment group received significantly more loans from Al Amana and not from other sources. Table 3 shows substantial and significant impacts of the treatment on assets, outputs, expenses and profits of self-employment activities.

While reproducing CDDP results, however, we identified issues with the trimming procedure, other imbalances at baseline and significant impacts on unlikely outcomes. In-depth verification revealed sampling errors, measurement errors and coding errors. These errors are not acknowledged by CDDP. After correcting the errors that could be corrected, we found different results, whose validity nevertheless remains uncertain.

Table 1: Summary statistics: reproduction of CDDP balance tests at baseline

		Control group			Treatment - Control	
	Obs.	Obs.	Mean	SD	Coeff.	p-value
Number of household members	4,465	2,266	5.14	2.69	0.043	0.582
Number of adults	4,465	2,266	3.45	1.99	0.031	0.563
Head age	4,465	2,266	47.8	16	1.08**	0.011
Does animal husbandry	4,465	2,266	0.533	0.499	0.042**	0.026
Runs a non-farm business	4,465	2,266	0.217	0.412	-0.034**	0.01
Loan from Al Amana	4,465	2,266	0.007	0.084	-0.003	0.424
Loan from other formal institution	4,465	2,266	0.06	0.238	0.03**	0.022
Informal loan	4,465	2,266	0.068	0.251	0.023***	0.005
Electricity or water connection loan	4,465	2,266	0.156	0.363	0.013	0.522

Source: Our reproduction of CDDP Table 1 with R, using the same raw data and specifications and producing the same results. Coefficients and p-values from an OLS regression of the variable on a treated village dummy, controlling for strata dummies (paired villages). Standard errors are clustered at the village level.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

Table 2: Credit: reproduction of CDDP regression results

	AAA admin data	AAA survey data	Other MFI	Other formal	Utility company	Informal	Total
Treated villages	0.167***	0.09***	-0.006	0.007**	-0.003	0.017	0.076***
	(0.012)	(0.01)	(0.004)	(0.003)	(0.007)	(0.017)	(0.017)

Source: Our reproduction of CDDP Table 2 with R, using the same raw data and specifications and producing the same results. Sample includes 4,934 households classified as high probability-to-borrow and surveyed at endline, after trimming 0.5 percent of observations. 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.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

Table 3: Self-Employment Activities: reproduction of CDDP results

	Assets	Sales and home consumption	Expenses	Of which: Investment	Profit
Treated villages	1,448** (658)	6,061*** (2,167)	4,057** (1,721)	-224 (223)	2,005* (1,210)

Source: Our reproduction of CDDP Table 3 with R using the same raw data and producing the same results. Sample includes 4,934 households classified as high probability-to-borrow and surveyed at endline, after trimming 0.5 percent of observations. 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.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

3 Results rely primarily on the trimming procedure and threshold

Deaton and Cartwright (2016) issue the following warning regarding trimming in RCTs: "*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.*"

Examining the trimming procedure applied by CDDP reveals that different procedures were applied at baseline and endline and that the final results are heavily dependent on the trimming threshold.

3.1 Different trimming procedures were applied at baseline and at endline

CDDP present the procedure they used for trimming as follows: "*Out of the 5,551, to remove obvious outliers without risking cherry-picking, we trimmed 0.5 percent of observations using the following mechanical rule: for each of the main continuous variables of our analysis (total loan amount, Al Amana loan amount, other MFI loan amount, other formal loan amount, utility company loan amount, informal loan amounts, total assets, productive assets of each of the three self-employment activities, total production, production of each of the three self-employment activities, total expenses, expenses of each of the three self-employment activities, income from employment activities, and monthly household consumption), we computed the ratio of the value of the variable and the ninetieth percentile of the variable distribution. We then computed the maximum ratio over all the variables for each household and we trimmed 0.5 percent of households with the highest ratios. Analysis is thus conducted*

over 5,424 observations instead of the original 5,551, and no further trimming is done in the data" (Crépon et al. 2015: 130).

However, this account is inaccurate: it should have read 5,524 instead of 5,424, which corresponds to the number of remaining observations once 0.5% of 5,551 has been removed. Secondly, most of the analysis's continuous variables were included in the trimming exercise, but not all of them: the number of worked hours was not included, for instance. In addition, this systematic trimming was applied only to endline data. The baseline data was the subject of far more erratic and extended trimming. Table 4 compares the variables and thresholds applied at baseline and endline.

As can be seen from Table 4, a number of trimmings were performed on different variables using different thresholds and at least two different procedures. The above-quoted complex procedure described by CDDP was used at endline. A simpler procedure was used for 24 variables at baseline, consisting of removing a variable value where this value was above a given variable distribution threshold. The thresholds determined for this "simple" trimming varied from one variable to another, from 0.1% to 0.4%. A total of 459 observations have been trimmed this way, out of a total of 4,465 observations in the baseline sample, that is a percentage of 10.3% of observations on which some variables have been trimmed at baseline. This raises three concerns. First, it is not true that "*no further trimming is done in the data*" (Crépon et al. 2015: 130). Second, setting fixed cut-offs for trimming lacks objectivity and is a source of bias, as it does not take into account the structure of the data distribution. Good practice for trimming experimental data consists of using a factor of standard deviation and, ideally, defining this factor based on sample size (Selst and Jolicoeur 1994). Third, the impact estimations are highly sensitive to the selected trimming threshold, as illustrated in the next section.

3.2 Variation in impact estimates depending on trimming threshold

In Table 5, we use the exact same data preparation and regression specifications as CDDP, and test other thresholds.

Table 5 shows that the results published by CDDP are highly sensitive to the threshold results and other thresholds than 0.5% point towards different interpretations. Thresholds below 0.5% produce results with no statistically significant impacts on self-employment activity outputs (sales and home consumption) or profits. The logical interpretation would then be that microcredit has no clear impact on self-employment activities. Thresholds above 0.5% generate a statistically significant impact in terms of an increase in expenses and decrease in investment, but no statistically significant impact on profits. It would be harder to produce a coherent interpretation of such results as, in particular, a decrease in investment is contradictory with an increase in assets. Initial conclusions on microcredit effects are also minimised if the provision of liquidity only results in an increase in turnover (sales and expenses), with no effect on investment or profits.

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.

Table 4: Inconsistent trimming procedures and threshold between baseline and endline by CDDP

Variable	Trimming threshold at baseline	Trimming threshold at endline
Amounts of active loans from AAA, informal & utilities	0.1% (BL:89-92)	
Amounts of active loans from other formal sources	0.3% (BL:94-5)	
Amounts of matured loans		0.5% (AN:247-72)*
Agriculture, livestocks and business assets	0.3% (BL:366-9)	0.5% (AN:247-72)*
Livestock & business investments	0.3% (BL:366-9)	
Agricultural investments	0.4% (BL:371-2)	
Agricultural sales	0.4% (BL:514-5)	0.5% (AN:247-72)*
Livestock sales	0.3% (BL:564-5)	0.5% (AN:247-72)*
Business sales	0.4% (BL:593-4)	0.5% (AN:247-72)*
Agriculture, livestock and business expenses	0.3% (BL:631-2, 675-6, 701-2)	0.5% (AN:247-72)*
Agricultural savings	0.3% (BL:756-7)	0.5% (AN:247-72)*
Livestock & business savings	0.3% (BL:785-6, 823-4)	
Consumption	0.1% (BL:923-4)	0.5% (AN:247-72)*
Income from dependent activities		0.5% (AN:247-72)*
Loan repayments	0.1% (BL:930-1)	
Income from self-employment activities	0.3% (BL:1016-7)	
Employment in agriculture and livestock	0.3% (BL:1073-6)	0.3% (EL:1299-302)
Work from family members in agriculture and livestock	0.3% (BL:1101-4)	
Distance to markets	0.1% (1256-71)	

Source: Examination of CDDP scripts for data preparation at baseline (BL) and at endline (EL).

* Those cases are trimmed using the procedure described in Crépon et al. (2015) and presented above (under 4.5): the whole observation is removed for each trimmed observation. For the other cases, only the outlying values of the trimmed variables were truncated as missing.

Table 5: Identical analysis to CDDP, but with varying trimming thresholds

Treshold	Obs.	Assets	Sales and home consumption	Expenses	Of which: Investment	Profit
Trimming at 0%	4,961	1,296*	3,282	3,784	22.1	-502
		(706)	(3,107)	(2,865)	(354)	(1,442)
Trimming at 0.3%	4,945	1,223*	4,231*	3,484*	-192	747
		(656)	(2,422)	(1,846)	(221)	(1,382)
Trimming at 0.5%	4,934	1,448**	6,061***	4,057**	-224	2,005*
		(658)	(2,167)	(1,721)	(223)	(1,210)
Trimming at 0.7%	4,923	1,377**	5,374***	4,129***	-358*	1,245
		(634)	(2,073)	(1,569)	(202)	(1,154)
Trimming at 1%	4,907	1,295**	4,492**	2,877**	-378*	1,615
		(602)	(1,898)	(1,290)	(202)	(1,098)
Trimming at 1.5%	4,880	1,298**	5,294***	3,678***	-377*	1,616
		(603)	(1,619)	(983)	(207)	(1,057)
Trimming at 2%	4,853	1,017*	3,107**	2,043**	-479**	1,064
		(573)	(1,388)	(880)	(202)	(933)
Trimming at 3%	4,802	834	2,216	1,698*	-435**	519
		(519)	(1,369)	(890)	(204)	(850)
Trimming at 5%	4,702	464	1,989**	788	-331**	1,202*
		(429)	(1,010)	(654)	(140)	(643)

Source: Our reproduction of CCDDP Table 3 with R, using the same data and same trimming procedure at endline, but with varying trimming thresholds. The sample includes the households surveyed at endline, minus the households considered as low probability-to-borrow and minus the trimmed observations. The other specifications are the same as CDDP 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.

*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

4 Imbalances at baseline and impacts on implausible outcomes

CDDP started their analysis by testing the balance between treatment and control groups on a limited number of variables. They found some small, but significant differences for some of them: households in treatment villages have more access to credit, more livestock activities and livestock assets, less non-farm business, and household heads are slightly older (see Table 1). The baseline values for these variables were therefore included as controls in the regressions to estimate impacts (Table 2 and Table 3 among others).

However, CDDP did not report the balance for the most important variables in their analysis, namely the outcomes they used to estimate the experiment's impact. They also did not report the balance on the characteristics that have been highlighted as essential in a qualitative research aiming at providing contextual insights for this RCT (Morvant-Roux et al. 2014): socio-economic status, belonging to a particular language or ethnic group, attitude towards female empowerment. It seems also important to check the balance on access to water and electricity services, as we will see in Section 5.1.4 that loans to finance connexions to these utilities are the main source of credit in the area, with a significant variation between baseline and endline. In Table 6, we use the same specification as in Table 1, to assess the balance between control and treatment groups at baseline, but with regression on these additional variables. We also estimate in Table 6 the average treatment effect on those additional variables, first with the exact same specifications as CDDP Table 3, second adding as controls the additional variables that appeared as imbalanced at baseline.

Table 6 reveals that, at baseline, households in the treatment group had significantly less sales and profits from self-employment activities than households in the control group. They were also making higher investments. There are also imbalances at baseline on several important variables, such as the area of owned land, access to basic services or women empowerment. When using the same specifications as CDDP, we also find significant treatment effects on outcomes for which microcredit impact is hardly plausible: household head gender, absence of education and spoken language.

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. However, the impact on profits no longer appears significant. Some impacts on unlikely outcomes are no longer significant, but others remain or appear, like household head gender, education and household members leaving the household.

The variables regarding access to electricity, water and sanitation deserve a specific attention. They show significant imbalances at baseline, but also a strong average treatment effects at endline. This is notable as we will see that branching credit and expansion campaigns from those utilities appear as a possible co-intervention that might have contaminated the experiment (see 5.1.5).

These imbalances at baseline and unlikely average treatment effects call for a closer examination of data quality and experiment integrity. We start with reviewing measurement and coding errors.

Table 6: Balance tests at baseline and impact estimates at endline, without correcting coding, measurement and sampling errors

	Balance at baseline						Impact at endline		
	N	Control group			Treatment - Control		ATE estimates		
Variable	Obs.	Obs.	Mean	SD	Coeff.[1]	p-value	Obs.	As in CDDP[2]	Adding controls[3]
Outcomes on self-employment activities									
Assets	4,440	2,251	13233	26469	923	0.254	4,934	1,448** (658)	2,130*** (814)
Sales and home consumption	4,440	2,251	34346	143720	-7510**	0.037	4,934	6,061*** (2,167)	6,518** (2,690)
Expenses	4,440	2,251	18671	67682	3202	0.225	4,934	4,057** (1,721)	5,043** (2,203)
Of which: Investment	4,440	2,251	732	5356	449**	0.042	4,934	-224 (223)	-31.4 (194)
Profit	4,440	2,251	15675	145624	-10712***	0.01	4,934	2,005* (1,210)	1,475 (1,250)
Household characteristics									
Male head	4,440	2,251	0.935	0.247	0.001	0.805	4,934	0.01* (0.006)	0.013** (0.006)
Head is a public servant	4,440	2,251	1.11	0.415	-0.024**	0.012	4,934	-0.014 (0.01)	-0.012 (0.011)
Head born in the same village	4,440	2,251	0.87	0.428	-0.025**	0.014	4,934	-0.008 (0.007)	-0.002 (0.008)
Head without education	4,440	2,251	0.623	0.5	-0.017	0.237	4,934	-0.029** (0.013)	-0.027* (0.014)
Members left in the last 5 years	4,440	2,251	0.093	0.361	0.011	0.319	4,934	0.009 (0.019)	0.044* (0.023)
Household head spoken language									
Darija	4,440	2,251	0.88	0.393	-0.019***	0.005	4,934	0.002 (0.008)	0.009 (0.008)
Berber	4,440	2,251	0.405	0.523	-0.013	0.301	4,934	-0.025* (0.014)	-0.017 (0.017)
Classical Arabic	4,440	2,251	0.188	0.402	0.001	0.943	4,934	0.021** (0.01)	0.017 (0.011)
French	4,440	2,251	0.067	0.249	0.002	0.721	4,934	0.005 (0.006)	-0.003 (0.007)
Household assets									
Number of color TVs	4,440	2,251	0.439	0.511	0.026	0.157	4,934	0.001 (0.02)	0.006 (0.016)
Owns land	4,440	2,251	0.61	0.488	0.011	0.51	4,934	0.004 (0.014)	0.001 (0.014)
Area of owned land	4,440	2,251	2.72	9.01	0.452*	0.094	4,934	0.01 (0.2)	0.018 (0.268)
Access to basic utilities									
Electricity from grid	4,440	2,251	0.616	0.486	0.057**	0.021	4,934	-0.011 (0.019)	-0.018 (0.016)
Sewage network	4,440	2,251	0.021	0.142	-0.014**	0.046	4,934	-0.013* (0.007)	-0.012** (0.006)
Septic tank	4,440	2,251	0.323	0.468	-0.029**	0.016	4,934	0.043*** (0.014)	0.041*** (0.015)
Private connection to piped water	4,440	2,251	0.344	0.475	-0.037	0.233	4,934	-0.045* (0.024)	-0.05** (0.026)
Shared connection to public tap	4,440	2,251	0.143	0.35	0.034**	0.045	4,934	0.029** (0.011)	0.026** (0.01)
Respondent considers that women should not:									
Go to the souk alone	4,440	2,251	0.716	0.451	-0.03**	0.017	4,934	-0.009 (0.014)	0.014 (0.015)
Take the bus alone	4,440	2,251	0.69	0.463	-0.029**	0.024	4,934	-0.014 (0.014)	0.02 (0.015)

*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

[1]Same specifications as in Table 1; [2]Same specifications as in Table 3; [3]Same specifications as in Table 3, adding as controls the baseline values of sales, investments, profits, head is a public servant, head was born in the same village, speaks Darija, area of owned land, household has a connexion to electricity, to the sewage network, to a septic tank, access to a public tap, respondent considers that women should not go to souk alone and that women should not take the bus alone.

5 Measurement and coding errors

Measurement errors can be observed in all sections of the dataset. We focus here on the variables used in the regression, which therefore have a direct incidence on identification and impact estimates. We also present the coding errors that have an incidence on the results. Other coding errors are listed in Appendix 3.

5.1 Inconsistent treatment (credit) measures

Credit measures are essentials to characterise the treatment and confirm that no contamination or co-interventions pose a threat to the experiment integrity. The analysis of coding and measurement errors on access to credit shows that the administrative data appended to the survey data is not reliable and indicates a lower take-up, as well as possible contamination and co-interventions.

5.1.1 Discrepancies between administrative and survey data

Household access to AAA credit was captured by two different questions, present in both the baseline and endline questionnaires:

- Question 'i3': Did you or a member of the household have a loan from '*[NAME OF SOURCE]*'? Is it outstanding or mature? (previous question specifies that recall period for matured loan is 12 months);
- Question 'i62': Do you or any household member have an outstanding loan or a loan that matured during the last 12 months from Al Amana?

Besides variables 'i3' and 'i62' that derive from the survey, CDDP built a third variable named 'client' out of data gathered from the AAA client registry.

The variable 'i3' indicates a low average level of borrowing from AAA at endline: 10.5% (289 households) in the treatment group and 2% (57 households) in the control group. The variable 'client' indicates a higher average level of borrowing from AAA at endline: 15.9% in the treatment villages (435 households) and 0% in the control villages. CDDP argue that more than a third of the households that took a loan from AAA did not report it in the survey and propose two interpretations: the household might not admit to borrowing because it is frowned upon by Islam; or they might confuse credit from AAA with credit from other formal sources. They conclude that administrative data must be regarded as more reliable than survey data to capture take-up (Crépon et al. 2015: 133-134).

Qualitative research in the settlements targeted by this RCT confirms that religious norms strongly influence practices and discourses related to credit (Morvant-Roux et al. 2014). Islam frowns upon two aspects. First, interest rates are explicitly illegal according to the sharia, which mostly applies to formal credit. Second, being in debt is regarded as a disgrace, which applies to all forms of credit. There is no question in the survey questionnaire that assesses religious practices or observance. If it were, we would probably notice some correlation between religious indicator and

credit. It would, however, be difficult to assess what arises from a lower credit taking and from a lower credit reporting, as religious norms might lead believers not to borrow rather than to borrow and refrain from reporting it to interviewers.

Table 7 presents cross-tabulation of the three variables that report household borrowings from AAA. It reveals that inconsistencies are much broader than the differences in averages of reported borrowings. Such inconsistencies contradict the assertion that the administrative data can be regarded as more reliable than the survey data.

Table 7 yields two insights. First, there are limited inconsistencies across different questions of the same survey: 20 households reported credit from AAA in Question 'i3', but not in Question 'i62'. Conversely, 26 households did not report credit from AAA in Question 'i3', but did so in Question 'i62'. Second, there are major inconsistencies between the survey data and the 'client' variable extracted from the AAA administrative data: 152 households declare having contracted a loan from AAA in Question 'i3' but do not appear in the 'client' variable retrieved from AAA administrative registries. 241 households identified in the latter as AAA borrowers declare not having an outstanding or matured loan from this microfinance institution (MFI) in Question 'i3'.

Table 7: Number of households borrowing from Al Amana at endline: contradictions between survey information and administrative data

	Credit from AAA in 'i62'		Credit from AAA in 'client'	
	Yes	No	Yes	No
Credit from AAA in 'i3'	320	26	194	152
No credit from AAA in 'i3'	20	5,185	241	4,964

Source: Our analysis using CDDP microdata retrieved from endline survey ('i3' and 'i62') and AAA administrative data ('client').

Of the 241 households identified as clients at endline based on the AAA administrative data and who declared not having an outstanding or matured loan from this microfinance institution (MFI) in Question 'i3':

- 27 reported at least one other formal credit[7] at endline;
- 25 reported at least one other formal credit at baseline (and 18 of those did not do so at endline);
- 2 reported filing a credit application that was refused (one of these two was not already reported in the above cases).

To sum up, the religion-driven shame argument clearly does not apply to 46 (27+18+1, i.e. 19%) of these 241 households, as they declare borrowing from formal sources elsewhere and, as explained above, the religion-driven shame argument applies equally to AAA microcredit and to other

[7] CDDP classify as formal credit: '*Al Amana*'; '*Zakoura*'; '*Crédit Agricole Foundation*'; '*Other MFI*'; '*Crédit agricole*'; and '*Other bank*'. See more details on credit sources in 3.1.4.

formal sources of credit. On the other hand, an argument of "credit shame" for these 241 households would call for an explanation of "credit pride" for the 152 households who reported having a loan from AAA even though they did not appear in the AAA registries.

Turning to the second argument regarding confusing AAA with other sources of formal credit, we show in Section 5.1.5 that access to formal credit did not increase in the treatment group, but remained stable with other formal sources replaced by AAA. In the control group, access to formal credit fell between the baseline and endline. The fact that the other formal sources of credit fell significantly in both groups between the baseline and endline does not leave much room for substantial confusion between AAA and other formal sources at endline.

Another plausible hypothesis to explain these discrepancies between survey data and administrative data is that the administrative data is inaccurate, or that it was not properly matched with the survey data. As we will see in Section 6.3, the sampling strategy failed to identify the households with a high propensity to borrow. It is therefore likely that a large part of the households that did borrow from AAA in the treated villages were not included in the survey sample. Besides, the microfinance sector in Morocco suffered a serious crisis from 2008 to 2012 (the endline surveys were conducted from May 2008 to January 2010) due to uncontrolled growth, over-indebtedness and widespread fraud by credit officers who used nominees to embezzle loans (Chen, Rasmussen, and Reille 2010; Rozas et al. 2014; D'Espallier, Labie, and Louis 2015). A Master's student who did an internship in AAA's internal audit division in 2009 substantiated the existence of such fraud in the MSc thesis he published on this subject (Hejjaji 2010). AAA had to write off 23%[8] of its portfolio in the following years as many loans were deemed uncollectable. To this should be added the rather frequent practice of borrowers themselves using nominees to bypass restrictive eligibility rules. These observations show that the reliability of the MFI administrative data should be viewed with caution, and that administrative data cannot be automatically considered to be more reliable than survey data. As the dataset is anonymised, we are unable to review the quality of the matching between survey and administrative variables.

In sum, the identification of the households borrowing from Al Amana matches across sources in 194 cases, versus 587 cases (241 + 152 +194) where households appear as borrowing from AAA in either the administrative data or the survey data. That is a concordance rate of 33%, which is small considering that credit from AAA corresponds to the "treatment" which effectiveness is being tested.

A large portion of CDDP's demonstration relies on these credit-taking variables. CDDP use the baseline values of variable 'i3' to produce their Table 1 and as control variables for their Tables 2 to 8. CDDP did not use variable 'i62' in their statistical analysis. The 'client' variable created from administrative data was used by CDDP to recompute a new borrowing propensity score, used to test externalities (Crépon et al. 2015: Table 8), in order to argue that there is no externality of microcredit and to justify the Local Average Treatment Effect (LATE) estimation. 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. We cannot rectify these inaccuracies with the available data,

[8]Data from Mix Market database: Write-off ratios from 2006 to 2016 are for each subsequent year: 0.5%, 1.3%, 3.7%, 6.4%, 3.5%, 8.7%, NA, 4.5%, 3.7%, 3.7%, 5.1%. The figure for 2012 is not known.

nor measure their incidence on the impact estimates. However, this imprecision regarding which households, and how many of them, benefited from the evaluated intervention undermines the internal validity of the RCT results, and in particular of the local average treatment effect estimations.

5.1.2 Credit from other MFIs was omitted at baseline

CDDP did not take into account loans from other MFIs when reporting access to credit and assessing the balance between treatment and control groups at baseline, as explained in more detail in Appendix A.2.1. In their Table 1, CDDP used the number of loans and the dummy (having a loan or not) variables to assess the balance between the treatment and control groups. The 'total access to credit at baseline' variable was also one of the control variables used for all regressions presented by CDDP (Tables 2 to 7).

Correcting this error increases by 3% the level of total access to credit in treatment and control group at baseline. This error combines with the one presented in Section 5.1.3, which has a larger incidence on measured credit access at baseline. This result has an incidence on the impact evaluation results, as illustrated in the following section.

5.1.3 Only outstanding loans were taken into account at baseline

When assessing access to credit at endline, CDDP included the loans outstanding at the time of the survey, plus the loans that were not outstanding any more at the time of the survey, but that had been outstanding in the past 12 months. When assessing access to credit at baseline, CDDP only included the loans outstanding at the time of the survey. They did not include the loans outstanding in the past 12 months that ended before the survey. Appendix A.2.2 details the coding error that led to this difference.

This 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. The identical naming and commenting on the code files suggests that the difference was not made on purpose. Besides, CDDP reiterate on three different occasions in their paper that this variable at baseline indicates whether a household "*had an outstanding formal loan over the past 12 months*" (pages 129, 132 and 133).

Correcting this error increases by 15% the measured level of total access to credit at baseline in treatment and control group. The revised levels of access by source and treatment or control group are detailed in Section 5.1.5, Table 11. 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*** (0.009) when correcting this error, which indicates an impact lower by 30% of the experiment on credit take-up. The average treatment 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.

5.1.4 All "other" credits were incorrectly recoded as "utilities" credit

In the baseline and endline surveys, credit sources were collected by the above-mentioned question 'i3'. Each loan was registered on a specific line of the questionnaire depending on its credit source. Sixteen possible sources were proposed to respondents at baseline (we reproduce here the English translations by CDDP): '*Crédit agricole*'; '*Other bank*'; '*Al Amana*'; '*Zakoura*'; '*Crédit Agricole Foundation*'; '*Other MFI (Microfinance Institution)*'; '*Usurer/Rhnane*'; '*Jeweler*'; '*Family*'; '*Neighbor*'; '*Friend*'; '*Shop*'; '*A client*'; '*A supplier*'; '*Cooperative*'; '*Other, specify:*'. A 17th option was added at endline: '*Utilities credit*'. Table 7 presents the number of respondents reporting one or more loan in the '*Other, specify:*' and '*Utilities credit*' categories:

Table 8: Number of households reporting one or more loan in the 'Other' and 'Utility' categories

	Surveyed households	Other	% other	Utilities	% utilities
Baseline	4,465	791	17.7	0	0.0
Endline	5,551	263	4.7	675	12.2

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

However, when recoding these variables, all sources registered as '*Other, specify:*' were reclassified as '*Utilities credit*' (see code in Appendix A.2.3). In other words, CDDP considered that all credit from sources other than those listed in the questionnaire was credit from water or electricity companies, even at endline where loans from water or electricity companies were specific options listed in the questionnaire.

To check for consistency, we first correlated the '*Other, specify:*' answers to Question 'i3' with the variable indicating whether households had water or electricity supply, both at baseline and endline (Table 9).

Table 9: Number of observations for which 'other' credit was recoded as 'utility' credit, whether they had access to utility services

Has electricity or water	"Other" credit at baseline	%	"Other" credit at endline	%
No	44	5.6	38	14.4
Yes	747	94.4	225	85.6

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

The vast majority of surveyed households were connected to water and electricity, with 69.9% having access to one of these services at baseline and 81% at endline (these two rates are simple averages without weighting). However, it does not seem appropriate to have recoded all declared "other" credit sources as "utility credit". It appears, for instance, implausible that households without water and electricity (first row in Table 9) could have received a "utility credit".

In the questionnaire, the '*Other, specify:*' option was followed by a field where the respondent was supposed to give the name of this unspecified source. We present in Appendix 1 the occurrences

encountered in this complementary variable and their corresponding frequencies. At baseline for instance, a specification corresponding to a utility company was provided in 29% of the cases, but in the others, the specifications corresponded to other types of sources (local stores, consumer lending, real estate purchase, etc.) or were missing. This indicates that, both at baseline and endline, credits registered as '*Other*' should not have been systematically reclassified as '*utility credit*'.

In addition, 17 credits at baseline and 17 credits at endline were registered with the amount, guarantee and other fields, but no source. Due to these missing values in the 'source' variable, these credits were not taken into account in the computations made by CDDP. For our replication, to avoid omitting them from descriptive statistics on access to credit, we replace these empty values with '*Other*' in the '*source*' field.

This approximation regarding credit from utility companies is noteworthy since they appear as the most important credit source in the surveyed villages, and it is also the type of credit source whose penetration varies the most between baseline and endline. After this correction, the results of the balance tests computed in CDDP Table 1 are modified, as shown in Table 10.

Table 10: Summary statistics: rectified balance at baseline on credit variables

		Control group			Treatment - Control	
	Obs.	Obs.	Mean	SD	Coeff.	p-value
Loan from Al Amana	4,465	2,266	0.007	0.084	-0.002	0.59
Loan from other formal institution	4,465	2,266	0.072	0.259	0.026*	0.068
Informal loan	4,465	2,266	0.085	0.279	0.026***	0.003
Electricity or water connection loan	4,465	2,266	0.039	0.194	0.023**	0.027
Other source	4,465	2,266	0.146	0.353	-0.024	0.225

Source: Our reproduction of CDDP Table 2 with R using the same specifications and correcting loan reclassification.
Coefficients and p-values from an OLS regression of the variable on a treated village dummy, controlling for strata dummies (paired villages). Standard errors are clustered at the village level.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

It is unclear whether this significant increase in access to utility credit in treatment villages is an unexpected impact of increased AAA credit or contamination by a co-intervention. In any case, further analysis would be required to disentangle the impact of microcredit and the impact of utility credit in this context. The existence of such imbalance at baseline and effects at endline is a threat to the internal validity of this RCT. This is an indication of an possible alteration of the experiment integrity and, if it is the case, part of the measured results would be attributable to utility credit instead of microcredit. Another RCT conducted during the same period in Morocco found significant impacts of utility credit on household well-being (Devoto et al. 2012). These results also raise questions regarding the external validity of the experiment: would the results apply to a context where there are no important efforts by water and utility companies to expand their services?

5.1.5 Credit access and identification of the treatment

In their published article, CDDP are very straightforward in the way they describe the difference in access to credit between treatment and control villages:

> "*Thirteen percent of the households in treatment villages took a loan, and none in control villages did.*" (Crépon et al. 2015: abstract)

> "*The study has three features that make it a good complement to existing papers. First, it takes place in an area where there is absolutely no other microcredit penetration, before or after the introduction of the product, and for the duration of the study.*" (Crépon et al. 2015: 124)

> "*The experimental design was generally well respected, and we observe essentially no entry of Al Amana (or any other MFI, as it turns out) in the control group. Villagers did not travel to other branches to get loans either.*" (Crépon et al. 2015: 130)

We computed credit prevalence in the treatment and control group at baseline and endline. As a substantial number of households (1,433) were added at endline without having been surveyed at baseline, we present the same analysis on the different subsets:

- One with the 5,551 households surveyed at endline and the 4,465 households surveyed at baseline (cross sections), and
- One with only the 4,118 households surveyed both at baseline and endline (panel).

Figure 1 focuses on panel households and Table 11 presents credit access for both panel and cross section households.

The difference in Table 11 between the repeated cross-sections and the panel households highlights sample errors, which we will analyse in more detail in Section 6 of this replication. At this stage, Table 11 shows that the attrition households and the households added at endline are very different in terms of borrowing levels to the households that were interviewed both at baseline and endline. This tends to rule out cross-section analysis and calls for a panel analysis instead. If we focus on growth in access to credit for panel households, as presented in Figure 1, we observe three striking phenomena that undermine the identification strategy used by CDDP.

First, access to formal credit did not notably increase in the treatment group (from 11.31% at baseline to 11.41% at endline). What we observe instead is a substitution of other formal credit sources by AAA.

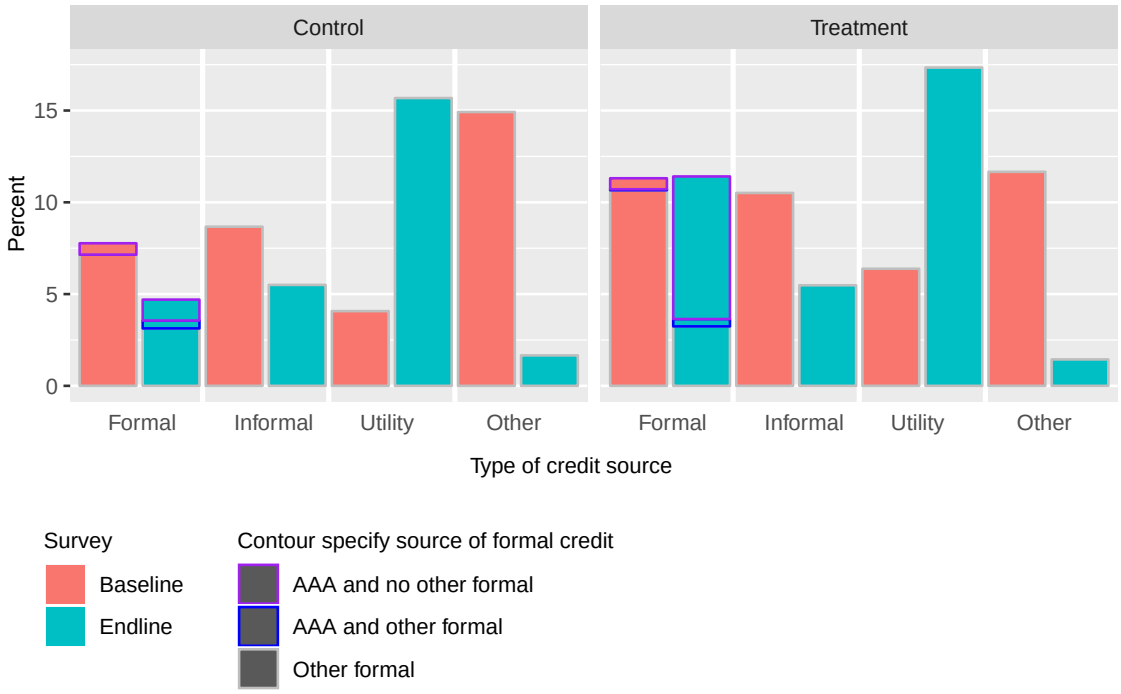
Second, access to formal credit significantly decreased in the control group (from 7.77% at baseline to 4.70% at endline). This might be due to the microcredit crisis that hit Morocco in 2008 (Chen, Rasmussen, and Reille 2010; Rozas et al. 2014; D'Espallier, Labie, and Louis 2015). It could also be explained by an agreement reached at the beginning of the RCT with the leading financial institutions that they would not intervene in the study areas. It might also be caused by AAA (which

Table 11: Changes in access to credit sources

		Control		Treatment	
	Credit source	Baseline	Endline	Baseline	Endline
Cross-section					
Formal					
	AAA and no other formal	0.66	1.53	0.59	10.00
	AAA and other formal	0.04	0.50	0.05	0.55
	Other formal only	7.19	3.35	10.05	3.72
	Total formal sources	7.89	5.38	10.69	14.27
Informal					
	Any informal source	8.47	6.01	10.41	5.62
Utility					
	Water or electicity company	3.93	15.30	6.14	17.04
Other					
	None of the above or not specified	14.61	1.49	11.37	1.53
Panel					
Formal					
	AAA and no other formal	0.62	1.14	0.60	7.77
	AAA and other formal	0.00	0.43	0.05	0.40
	Other formal only	7.15	3.13	10.66	3.24
	Total formal sources	7.77	4.70	11.31	11.41
Informal					
	Any informal source	8.67	5.50	10.51	5.48
Utility					
	Water or electicity company	4.07	15.68	6.38	17.34
Other					
	None of the above or not specified	14.92	1.66	11.66	1.44

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

Figure 1: Changes in access to credit sources for panel households – households surveyed both at baseline and endline (N= 4,118)



Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

headed the influential national MFI association at the time) calling on its fellow financial institutions to minimise the contamination of the experiment during the RCT.

Third, according to the survey data, utility credit is by far the most prevalent credit source in both treatment and control groups. The stability of access to utility credit over time is based on the extrapolation made by CDDP that all "other" sources of credit were "utility" credit. We show that this could not be true in a significant proportion of cases where there are patent contradictions with available information (Table 5). If we reject the automatic reclassification of "other" credits as "utility" credits when no such specification was given by respondents, then variations in "utility" credit are substantial between the baseline and endline.

These observations challenge the very meaning of the experimentation put forward by CDDP. What has been evaluated: is it the impact of the replacement of other formal sources with AAA in the treatment group? Is it credit rationing in the control group? Or is it the variation in utility credit?

5.2 Outcome measures and controls

5.2.1 Incomplete or inconsistent data

We discuss here just two examples of the many survey data inconsistencies we found.

In 1,382 cases at baseline and 307 cases at endline, households declared having agricultural assets of some kind, but the number of items is missing. These assets were therefore not taken into account in the total. This goes for all types of agricultural assets from tractors, reapers, cars and trucks to shovels, axes and sickles. The same problem concerns livestock assets and business assets.

Two sections of the questionnaire focused on the assessment of business (non-farm) activities. Section D on "Household activities" records (only) self-employment activities, and G gathers all information, including financial, on these activities outside of agriculture (questionnaire section E) and stockbreeding (questionnaire section F). We find that of the 746 households with business activities registered in D at endline, 41 (5.5%) have no business activity or only part of their business activities documented in G. On the other hand, of the 751 with business activities documented in G, 46 (6.1%) have no business activities or only part of their business activities registered in D.

These inconsistencies cannot be corrected with the available information. However, they call into question the quality of the underlying data of this RCT, and hence its internal validity.

5.2.2 'Tractors' *and* 'reapers' *removed from asset appraisal at endline*

At baseline, CDDP included all types of assets to calculate the total value of the assets owned by all households. However, an examination of the code used to compute endline data (see Appendix A.2.4) shows that two types of assets have been removed from the sum of asset values calculated for each household: tractors and reapers. The code between endline and baseline preparation do files is overall the same, suggesting that it was copy-pasted. This specific change was therefore made intentionally, but is not mentioned in the published article. It was probably motivated by the fact that the appraisal method used by CDDP produces inaccurate prices, which are particularly erratic for those two assets (see Section 5.2.3). This is however inadequate, because this RCT aims at evaluating the impact on assets, among other outcomes, and tractors and reapers are the most valuable assets that households possess.

Including tractors and reapers in the asset appraisal at endline increases average asset value in the sample from 1,377 to 5,111. 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. However, it becomes 3,041** (1,402), which is larger and significant, when we included the total access to credit as corrected in Section 5.1.2 and Section 5.1.3. This estimation is further modified when we correct the price calculation used for asset valuation, as explained in the following section.

5.2.3 *Assets, sales and consumption appraised with inconsistent prices*

The survey suffers from a classic problem with price imputation every time assets, sales, consumption of own production and in-kind savings have to be evaluated (Deaton 1997: 28-29, 35-39). A price has to be imputed for each item to account for its value. Yet in most cases, no transaction price is available for that particular item, either because there was no transaction (assets purchased more than a year ago, consumption of own production or savings) or because the transaction price was not registered (new assets and sales). In these cases, the median price of all observed transactions by other households for this item was imputed. The problem is that for some items, the number of transactions for which a price is available is very small, exposing the median to being skewed by outliers or implausible prices reported by the households. Table 12 presents some illustrations of median prices imputed to agricultural assets.

Table 12: Median prices imputed to agricultural assets at baseline and endline

Asset	Total number of items owned by households at baseline	Number of times a value was reported at baseline	Median value at baseline	Total number of items owned by households at endline	Number of times a value was reported at endline	Median value at endline	% variation of median value
Tractor	114	11	21,000	193	14	85,000	305
Reaper	30	4	5,500	31	2	187,500	3,309
Traditional laborer	1,880	80	200	2,638	42	200	0
Cart	312	7	500	452	3	350	-30
Rake	2,094	44	25	3,250	20	55	120
Shovel	3,088	103	25	5,221	48	40	60
Ax	2,341	69	60	4,736	29	50	-17
Wheelbarow	1,381	41	250	2,897	23	250	0
Sickle	4,796	222	35	7,415	83	35	0
Car or truck	101	8	52,750	79	2	70,000	33
Oil Mill	92	1	1,000	72	0	NA	NA
Other 1	5	20	60	137	16	80	33
Other 2	36	4	60	9	2	2,650	4,317
Other 3	4	2	65	3	0	NA	NA
Other 4	0	0	NA	0	0	NA	NA
Other 5	0	0	NA	0	0	NA	NA

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

The median reaper value increased by 3,309% between the baseline and endline. The example of agricultural assets presented in Table 12 shows that imputing a median price where only a small number of transactions have been made in the last year gives rise to erratic assets valuations. With scarcely recorded transactions, it is clearly preferable to compute a median price that takes into account transactions observed both at baseline and endline. This hinders the capture of genuine price variations (inflation), but seems like a reasonable trade-off considering the absurd price variations observed above. According to this approach, tractor should have been appraised at MAD 60,000 at endline, which is the median value of the 25 transactions registered at both baseline and endline.

Reaper values should be appraised at MAD 10,200 at endline, which is the median value of the 6 transactions registered at both baseline and endline, etc.

This correction further modifies the estimation of the experiment impact introduced in Section 5.2.2. The result on total assets at baseline becomes 1,628* (969), which is slightly marginal, when we keep the same control variables as CDDP. It becomes 2,520** (1,102), which is larger and significant, when we replace the total access to credit by the rectified value corrected in Section 5.1.2 and Section 5.1.3.

Moreover, a recurring problem can be seen in Table 8 with all items owned, sold or bought computed by CDDP. The "other" category is always valued at the same median price, despite its covering highly heterogeneous items. This problem with the undefined "other" category is found with the business, livestock and agricultural assets, and also with vegetable, cereals and tree sales. For instance, a tiller, a handheld sprayer and a pruning shear are valued at the same median price as soon as they come under the same "other" category.

5.2.4 Other measurement and coding errors on outcome and control variables

A series of other errors have been identified. A disputable amortisation procedure led to divide the value of some agricultural investments by 10. A Stata coding error added units of livestock assets that do not exist. Several confusions were made between prices before, during or after harvest when appraising agriculture sales and consumption. Control variables referring to household composition are altered in some observations: no members, several heads, missing ages, etc. These errors affect a limited number of observations, or affect observations with a limited magnitude. They only yield a marginal incidence on the estimated results, so we present them in Appendix 3.

5.3 Results with partial corrections

We now recompute the regression presented by CDDP (Table 3), correcting the coding and measurement errors that can be corrected: account of borrowing at baseline including credit from other MFIs (see Section 5.1.2); borrowing at baseline factoring in all outstanding loans in the past 12 months, instead of just outstanding loans (see Section 5.1.3); appraisal of agricultural assets at baseline including tractors and reapers (see Section 5.1.4); livestock assets excluding non-existent units (see Appendix 3.2); business earnings including all business sales (see Appendix 3.3); prices before, during and after harvest suitably assigned to corresponding sales or consumption (see Section 5.2.4); and investment in agricultural assets not amortised by an arbitrary procedure (see Appendix 4.1). All in all, these corrections affect 3,866 of the 4,934 observations (78.35%) used by CDDP (Table 3) for their ATE estimation on self-employment activities.

Table 13 shows that the standalone correction of some coding errors reduces and cancels out the magnitude and significance of the estimated impacts, as shown for instance the inclusion of credits from other MFIs and all credits outstanding in the 12 previous months. But the correction of other errors considerably reinforces the estimated impacts. Taken together, these rectifiable errors appear relatively well balanced between treatment and control groups and their correction does not,

Table 13: Replicated impact estimates correcting some coding and measurement 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)
Some error corrections and trim at 0.5%	1,251* (653)	7,556*** (2,706)	3,115* (1,827)	-220 (223)	4,441** (1,935)

Source: Our replication of CDDP Table 3 with R, using the same data but 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; inconsistent amortisation rules for agricultural investments. Same specifications as CDDP Table 3: Sample includes 4,934 households classified as high probability-to-borrow and surveyed at endline, after trimming 0.5 percent of observations. 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 activities, 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.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

in itself, disqualify the conclusions of the first part of the published article. We notice at this stage that estimated impacts on assets and expenses are smaller and less significant, and that estimated impacts on outcomes and profits are larger and more significant.

One should bear in mind that what we have here is only a partial correction, since measurement errors remain: there are still missing and absurd values (see sections 5.1.4, 5.2.1 and Appendix 3); consumption of own production and in-kind savings are still valued at erratic median prices wherever there were not enough registered transactions to obtain reliable estimates (Section 5.2.3), etc. Besides, the measurement errors observed on credit variables do raise major concerns about the reliability of the externality tests and the local average treatment effects, which are the second part of the CDDP paper, not reproduced here.

6 Sampling errors

CDDP describe their sampling procedure as follows. From a pilot survey including 1,300 households in seven pairs of villages, 24 variables were identified as "good predictors" for a household to borrow from AAA. A logit model was built to assess borrowing propensity based on these 24 variables. One village per pair was randomly selected from 81 pairs of similar villages to receive microcredit services from AAA. Prior to the opening of an AAA branch in the village, a short preparatory survey was administered to a sample of 100 households in each village, or to the entire village where the population was less than 100 households. The 24 variables previously mentioned were included in the short questionnaire, and they were used to compute a borrowing propensity score for each of the 15,145 households surveyed in this preparatory phase. In each village, all the households surveyed during the short preparatory survey that ended up in the top borrowing propensity quartile were included in the sample. Five other households that were surveyed during the short preparatory survey but that did not end up in the top quartile were also randomly selected. A total of 4,465 households were interviewed at baseline, of which 92% were successfully re-interviewed at endline. The propensity score to borrow was then re-estimated before the endline for all households interviewed during the short preparatory survey, based on the take-up observed by the AAA information system in the 81 treatment villages. According to this new score, 1,433 households that had not been selected to be interviewed at baseline were considered as having a very high propensity to borrow and were added to the endline sample.

In the following section, we call the latter "households added at endline," as opposed to "panel households" interviewed at both baseline and endline, and "attrition households" those that were only interviewed at baseline.

6.1 Household differences between preparatory and baseline surveys (and endline for those added at endline)

We first seek to assess whether the information collected about the households at baseline is consistent with the information collected on those same households by the preparatory survey. We focus on household size, which should not have changed substantially in a short period. We flag the households whose number of members varied by more than 30% and by more than two people (to avoid a false positive with small households) between the preparatory survey and the baseline survey. We also examine three variables used to compute the borrowing propensity score that determined household inclusion in the sample: the household owns land ('*yes*' or '*no*'), the household has olive or argan trees ('*yes*' or '*no*'), and one or more household members receive a pension ('*yes*' or '*no*'). These three variables are chosen from the 24 included in the propensity score, because they were collected in an identical way in the preparatory and baseline survey questionnaires. Variations on the same households in a short period of time should therefore be limited. In addition to the households surveyed at baseline, we also run the same analysis for households added at endline.

We observe in Table 14 that in 985 cases (22.06%), the number of household members is compatible between the preparatory survey and baseline survey, but the selected propensity score criteria are inconsistent. In 431 additional cases (9.65%), the selected propensity score criteria are consistent between the preparatory and baseline surveys, but the number of household members

changes significantly. In 104 other cases (2.33%), both the number of household members and the selected propensity score criteria are inconsistent. In total, we observe a mismatch on these key variables between the preparatory and baseline surveys for 1,520 households of the 4,465 households sampled at baseline (34.04%).

For households added at endline, substantial changes in household composition can happen considering the time lapse (delay between preparatory and baseline survey plus two years), but not to such an extent. In total, a mismatch is observed on these key variables between the preparatory and endline surveys for 724 households of the 1,433 households added at endline (50.52%).

We do not try to correct these observed inconsistencies, as it would imply removing a large number of observation from the sample, hampering its statistical power. However, we notice here a major concern regarding the way households have been selected for their inclusion into the sample.

Table 14: Differences in household characteristics between preparatory survey and baseline

Significant difference in number of household members[1]	One or more of the 3 selected propensity criteria[2]	Selected at baseline	%	Added at endline	%
No	No	2,869	64.3	705	49.2
No	Yes	985	22.1	442	30.8
No	NA	14	0.3	1	0.1
Yes	No	431	9.7	170	11.9
Yes	Yes	104	2.3	112	7.8
Yes	NA	5	0.1	2	0.1
NA	NA	57	1.3	1	0.1
	Total	4,465	100.0	1,433	100.0

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.
[1]Number of members varied by more than 30% and by more than two people;
[2]The household owns land, the household has olive or argan trees, and one or more household members receive a pension.

6.2 Inconsistencies in household composition between baseline and endline

For panel households, the same households should have been interviewed at both baseline and endline. A consistent definition of household composition is also needed to make reliable comparisons, as household composition determines all living standards parameters such as income, consumption, poverty and food security (Deaton 1997: 204–268). The literature on the informal economy in developing countries also establishes that household composition is the defining criterion to be able to assess all parameters relating to self-employment activities (Cling et al. 2014).

At baseline and endline, the respondent was asked to list and describe the key characteristics of all household members. We use the information to analyse whether the composition of each

household is consistent between baseline and endline surveys. The average household size was 5.17 at baseline and 6.13 at endline. This clearly points to a problem, as the number of members per household is not consistent between baseline and endline. For comparable figures, national population censuses establish that rural household size in Morocco was 6.59 in 1994, 6.03 in 2004 and 5.35 in 2014 (Direction de la statistique 2005a: 14; Direction de la statistique 2015: 3).

We created an algorithm to compare household composition between baseline and endline. The algorithm checks for each household member at baseline whether there is a corresponding household member at endline of the same gender at a compatible age. The endline survey was conducted two years after the baseline survey, so we consider for each household member that a compatible age at endline would be the person's age at baseline, plus 1 to 3 years. To check the sensitivity of our matching analysis, we also broaden the range of compatible age from 0 to 5 five years' difference between endline age and baseline age. Benefit of the doubt is accorded in the case of missing information, i.e. when age is not documented. We therefore consider the presence of a household member of the same gender, but with no registered age, as a possible match. All possible combinations between all members at baseline and all members at endline are checked and the configuration with the highest number of matches is retained for each household. We then compute a score that classifies each household according to the proportion of matches in its composition between baseline and endline:

- Identical: all household members match between baseline and endline;
- Slightly different: one-tenth or less of household members do not match between baseline and endline;
- Different: one-tenth to one-quarter of household members do not match between baseline and endline;
- Very different: half to one-quarter of household members do not match between baseline and endline;
- Mostly inconsistent: more than half of the members do not match between baseline and endline;
- No match: none of the members at endline matches the members at baseline;
- Too many members/check manually: the algorithm checks all possible permutations of household same-gender members between endline and baseline. It therefore becomes computationally overwhelming if there are more than ten same-gender members at baseline and/or endline. This only occurs in 16 cases, which we discard from the analysis.

Table 15 shows that the composition of 834 households (655 + 179, i.e. 20.25% of panel households) is entirely or mostly incompatible between baseline and endline. As illustrated in appendix 4, it is not plausible in cases presenting such a magnitude of discrepancy that the same households could have been re-interviewed. The full list of mismatched households is published with the supplementary material[9] to this paper in a `.csv` file. In these cases, it seems plausible that the interviewer failed to reach at endline the household that had been interviewed at baseline and

[9] www.iree.eu

Table 15: Number of households according to the proportion of members whose gender and age match between baseline and endline

Status	Threshold 1 to 3 years	%	Threshold 0 to 5 years	%
Identical	1,783	43.3	2,295	55.7
Slightly different	233	5.7	281	6.8
Different	603	14.6	579	14.1
Very different	649	15.8	471	11.4
Mostly incompatible	655	15.9	362	8.8
No match	179	4.3	114	2.8
Too many members: check manually	16	0.4	16	0.4
Total	4,118	100.0	4,118	100.0

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

interviewed another household instead.

Removing the observations corresponding to these mismatched households translates into slightly different estimates. But this removal has to be combined with the inclusion of the observations initially discarded by CDDP as "low borrowing propensity households", as explained in Section 6.3. We present the incidence of the overall resampling in Section 6.4.

6.3 Contradictions in propensity scores used as sampling criteria

The cornerstone of this RCT protocol and the corresponding article's identification strategy is the household propensity to borrow, which was evaluated by scores. Attentive readers of the article will understand that two scores were used to assess the household borrowing propensity. Examination of the do-files reveals that there were actually four scores:

- **Score 1**: as we explained at the beginning of Section 6, the households sampled at baseline in each village were selected based on a score predicting their propensity to borrow (Score 1). This score was calculated before the baseline using variables collected on all surveyed households by the preparatory survey. In each village, the top quartile of households was classified a "*high borrowing propensity*" group and sampled. Five households randomly selected from the rest of the village were also included in the sample and classified a "*low borrowing propensity*" group;
- **Scores 2 and 3**: at the beginning of the endline survey, given the low take-up observed since the beginning of the RCT, CDDP re-calculated a second score (Score 2) supposed to be more accurate than the previous one. Matching the preparatory survey with current AAA administrative data, the new score was computed to better identify potential borrowers that were not sampled at baseline in order to include them in the endline survey. They then recalculated a third score (Score 3) supposed to be even more accurate – based on the same procedure,

but using an updated version of the AAA client register – to select the households for the last phases of the endline survey. Households added based on both scores were classified a "*very high borrowing propensity*" group.

- **Final score**: CDDP computed a last propensity score, based on the ex-post information contained in the AAA client register. Section 5.1.1 already points out that this administrative data was substantially inconsistent with the information collected by the survey.

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. The final score is the variable used by CDDP (Table 8-panel C) to segment the sample. The values in this Table 8-panel C are the main argument used to justify the instrumental variable regression (using treatment/control classification as an instrument) conducted by CDDP (Table 9).

6.3.1 Scores contradict one another

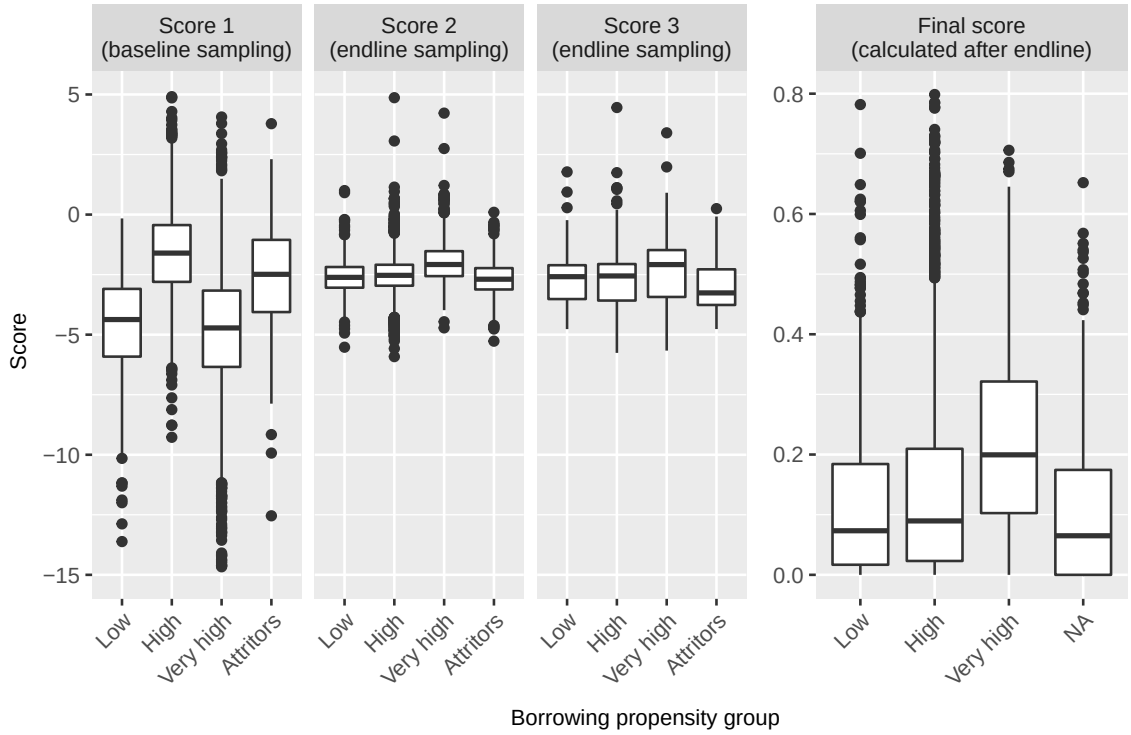
We analyse whether households classified in different borrowing propensity groups do indeed have consistent scores across the subsequent estimations made by CDDP. We do so by charting the distribution of observations for each score, separating out "low propensity", "high propensity" and "very high propensity" observations each time (Figure 2).

If the scores were reliable, Figure 2 would present a difference between the distributions: the "high propensity" group would be well above the "low propensity" group, not only for score 1 (on which the classification was based), but also for scores 2 and 3 and the final score. This is not the case. For instance, the "low propensity" group and "high propensity" group have very similar score 2 and 3 distributions. Moreover, the "very high propensity" group displays a score 1 distribution that is similar to the "low propensity" group.

The low association between scores is puzzling, as they are supposed to reflect, at least in part, the same phenomenon. It can be deciphered by observing the scoring factors that were attributed each variable to compute scores 1, 2 and 3 and the final score, as presented in Appendix 5.

Observation of Table 22 indicates that the coefficients attributed to each scoring variable drastically change from one score to the next, denoting a lack of estimation robustness. Some of them become non-significantly different to 0 and vice versa. Moreover, some coefficients change signs for opposite values, from positive to negative and vice versa. For instance, owning land was attributed a negative factor for propensity scores 1 and 3, but a positive factor for score 2 and the final score. Having a fibre mat corresponded to a positive coefficient for scores 1 and 2, but negative for score 3 and the final score. Doing more than three self-employment activities was associated with a significant positive coefficient for score 1, negative for score 2, and was not retained as a scoring variable for score 3 and the final score. And so on and so forth. We observe such contradictions for most of the variables used to compute the borrowing propensities, suggesting that these scores suffer from a major lack of robustness.

Figure 2: Contradiction between borrowing propensity scores

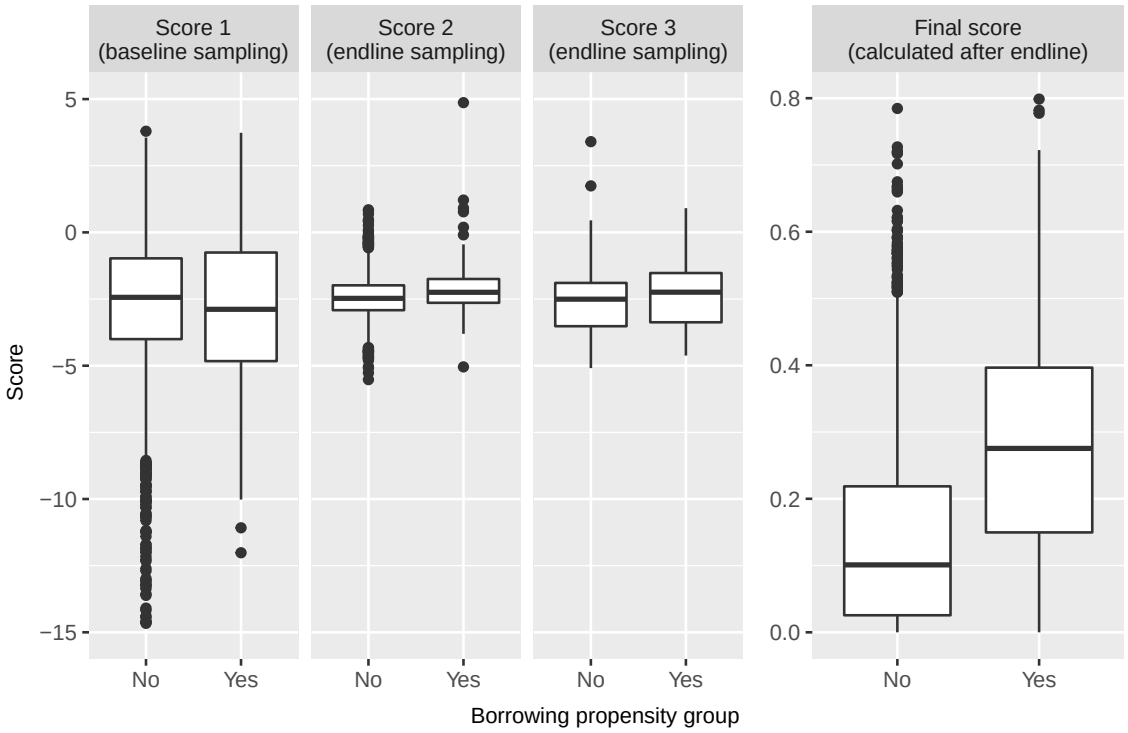


Source: Our analysis using CDDP microdata retrieved from endline survey.

6.3.2 Borrowing propensity scores fail to predict borrowing

To be considered as a propensity score for an event, a variable must predict the occurrence of such an event. We analyse whether the borrowing propensity scores are able to predict borrowing. Figure 3, which presents score distribution based on the borrowing status of households in treatment villages, suggests that the power to predict differences in access to credit is more than limited.
Table 16 presents association tests between scores and borrowing. For score 1, p-values above 0.05 mean that the null hypothesis cannot be rejected. We can conclude that score 1 is not associated with borrowing. We can reject the null hypothesis for the other scores, i.e. there is some association between the score and borrowing.

Figure 3: Borrowing propensity scores fail to predict who borrows treatment villages



Source: Our analysis using CDDP microdata retrieved from endline survey.

Table 16: Score 1 is not associated with borrowing

Score	T test	p value T test	Wilcoxon	p value Wilcox
Score 1	-1.3903	0.1653	848,223.5	0.0703
Score 2	6.6552	0.0000	1,099,727.5	0.0000
Score 4	3.4959	0.0005	1,009,108.0	0.0002
Final score	13.0926	0.0000	1,299,019.0	0.0000

Source: Our analysis using CDDP microdata retrieved from endline survey. Association between scores and reported borrowing in variable 'i3'.

6.4 *Results with a consistent panel sample and correcting some coding and measurement errors*

To tackle the sampling issues listed above, we recompute the impact estimates with resampling. We include the households classified by CDDP as "low borrowing propensity", because they were selected based on a score that does not reflect their borrowing propensity (see 6.3.2) and their actual

borrowing propensity is not different from the households classified as "high borrowing propensity" (see 6.3.1). We also restrict the analysis to households with compatible baseline-endline compositions, which means discarding households classified as "mostly inconsistent" or "no match" (see Section 6.2).

Here too, this only partially corrects the sampling errors. For instance, it does not discard households whose characteristics used as sampling criteria differed between the preparatory survey and the baseline. Rectified estimates with resampling and coding errors corrected in 5.3 are in Table 17.

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 endline 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.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

With a consistent panel sample, we have 3,268 observations. We see that focusing the analysis on this consistent panel yields different results: the impact estimate on sales is smaller and less significant, the impact estimates on expenses is smaller, and the impact estimate on profits is not significant anymore.

In Table 18 we check for imbalances at baseline for this resampling, as well as the impact at endline for a series of outcomes.

Table 18: Balance tests at baseline and impact estimates at endline, correcting some measurement and sampling errors

	Balance at baseline						Impact at endline		
	N	Control group			Treatment - Control			ATE estimates	
Variable	Obs.	Obs.	Mean	SD	Coeff.[1]	p-value	Obs.	Correcting some errors[2]	Adding controls[3]
Outcomes on self-employment activities									
Assets	3,268	1,686	14058	29234	629	0.505	3,268	1,277* (767)	1,320 (809)
Sales and home consumption	3,268	1,686	35924	153162	-7511*	0.099	3,268	5,990** (2,680)	6,070** (2,733)
Expenses	3,268	1,686	17773	52531	4021	0.241	3,268	3,815** (1,893)	4,084** (1,884)
Of which: Investment	3,268	1,686	721	5469	216	0.317	3,268	-5.46 (148)	-113 (147)
Profit	3,268	1,686	18181	148487	-11528**	0.032	3,268	2,175 (1,722)	1,987 (1,714)
Household characteristics									
Male head	3,268	1,686	0.941	0.235	0.017***	0.001	3,268	0.012** (0.005)	0 (0.004)
Head is a public servant	3,268	1,686	1.11	0.415	-0.014	0.205	3,268	-0.007 (0.014)	-0.012 (0.013)
Head born in the same village	3,268	1,686	0.857	0.422	-0.029***	0.005	3,268	-0.003 (0.01)	-0.001 (0.009)
Head without education	3,268	1,686	0.594	0.498	-0.008	0.605	3,268	-0.037** (0.015)	-0.027* (0.015)
Members left in the last 5 years	3,268	1,686	0.095	0.346	0.013	0.288	3,268	0.051* (0.026)	0.059** (0.027)
Household head spoken language									
Darija	3,268	1,686	0.859	0.399	-0.006	0.434	3,268	0.012 (0.008)	0.01 (0.009)
Berber	3,268	1,686	0.398	0.513	-0.017	0.242	3,268	-0.017 (0.018)	-0.021 (0.018)
Classical Arabic	3,268	1,686	0.189	0.401	0.006	0.597	3,268	0.045*** (0.012)	0.037*** (0.012)
French	3,268	1,686	0.065	0.247	0.006	0.38	3,268	0.017** (0.007)	0.014* (0.007)
Household assets									
Number of color TVs	3,268	1,686	0.447	0.515	0.033*	0.085	3,268	0.044** (0.019)	0.028* (0.016)
Owns land	3,268	1,686	0.609	0.488	0.005	0.789	3,268	-0.01 (0.015)	-0.012 (0.015)
Area of owned land	3,268	1,686	2.52	7.14	0.459	0.119	3,268	-0.193 (0.315)	-0.255 (0.343)
Access to basic utilities									
Electricity from grid	3,268	1,686	0.619	0.486	0.056**	0.035	3,268	0.005 (0.016)	-0.014 (0.014)
Sewage network	3,268	1,686	0.016	0.124	-0.008*	0.097	3,268	-0.013* (0.007)	-0.013* (0.007)
Septic tank	3,268	1,686	0.326	0.469	-0.033**	0.014	3,268	0.041*** (0.014)	0.035** (0.014)
Private connection to piped water	3,268	1,686	0.345	0.475	-0.014	0.653	3,268	-0.034 (0.023)	-0.038 (0.023)
Shared connection to public tap	3,268	1,686	0.143	0.351	0.033*	0.058	3,268	0.026** (0.01)	0.025*** (0.009)
Respondent considers that women should not:									
Go to the souk alone	3,268	1,686	0.702	0.458	-0.025*	0.08	3,268	0 (0.017)	-0.001 (0.017)
Take the bus alone	3,268	1,686	0.676	0.468	-0.03**	0.046	3,268	0.005 (0.017)	0.005 (0.016)

[1]Exact same specifications as in Table 1.
[2]Same specifications as in Table 17.
[3]Same specifications as in Table 18, adding as controls the baseline values of sales, profits, head was born in the same village, household has a connexion to the electricity grid, to the sewage network, to a septic tank, access to a public tap, respondent considers that women should not go to souk alone and that women should not take the bus alone. Sample includes 3,268 households interviewed both at baseline and endline and which member gender and age composition is compatible between baseline and endline.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

Table 18 confirms that even after correcting some measurement and coding errors and focusing on a consistent sample, we still find important imbalances at baseline, on sales and profits, household head gender and origin, access to electricity, water and sanitation, or opinion on women's empowerment. When applying the same corrections of coding, measurement and sampling errors at endline, we find that the impact on assets and profits are not significant, and that the main results are to be found in increasing turnover from self-employment activity. However, we also observe disconcerting estimates on other outcomes. Microcredit would then increase household head education, foster members to leave the household, increase the knowledge of Arabic and French, impede households' access to public sewage and incentivise the use of septic tanks, as well as access to public taps. We also see that household buy more TVs, while a prominent conclusion of CDDP was that it reduced nonessential expenditures. Such outcomes are hardly plausible and we interpret them as an indication of a lack of quality of the data and of alterations in the protocol and the survey sampling.

7 External validity: what might the results be representative of?

If the sampled households do not represent high borrowing propensity rural households, then what do they represent? The inconsistent scoring system explained in 6.3 skewed the representativeness of the baseline sample towards a population subset. Score 1 tended towards the sampling of households owning less land, with fewer cows and more non-agricultural self-employment activities. Yet scores 2 and 3 used to add new households at the endline tended more towards the inclusion of agricultural households.

We can compare some of this population's characteristics with other Moroccan data taken from such sources as major national surveys or censuses of the rural population. For instance, CDDP report a monthly consumption average of MAD 2,272 per household at baseline (data collected from April 2006 to December 2007) compared with the MAD 3,611 found in Morocco's rural population by the 2007 National Living Standards Survey (Direction de la statistique 2007, data collected between December 2006 and November 2007). This means either that the study population was 37% poorer than the average population or that there are inconsistencies between the household expenditure estimation method used by this RCT survey and the national household survey. Pamies Sumner (2015: 72-74) pointed out, for instance, that the questionnaire designed by CDDP deviated considerably from the living standards measurement survey questionnaire and procedures developed by Moroccan statisticians for domestic surveys.

CDDP also report that household heads are men in 93.5% of cases as opposed to the 87.4% average for rural households found in the 2004 population and housing censuses (Direction de la statistique 2005b). Section 6.2 also saw that the average household size in the RCT sample stood at 5.17 members at baseline and 6.13 members at endline. Moroccan rural households had an average of 6.03 members in 2004 and 5.35 members in 2014, displaying a decreasing pattern contrary to the experiment's observations.

In short, the RCT sample covered households with lower income and different demographic characteristics to the average Moroccan rural population, and with converse household size variation trends. So what are they representative of?

8 Conclusion

This replication was made possible by the fact that the authors and the journal shared the data and codes used to produce the published results. This is commendable and should be further encouraged, as it will enhance the reproducibility and credibility of empirical research, in particular in development economics. The replication of this RCT on microcredit in Morocco identifies a number of shortcomings that challenge the conclusions drawn by CDDP.

The trimming procedure used on the data by the authors is debatable and the impact estimations rely heavily on the trimming threshold selected. Trimming at slightly different thresholds returns different or statistically non-significant results. We also find out that the sample was significantly imbalanced at baseline on the main outcomes, as well as several other important variables. We apply the same regressions as in the original paper, but controlling for these imbalances at baseline and find that the impacts on profits do not hold and that the increases in expenses and outputs were underestimated. We also find impacts on variables that are unlikely to be influenced by microcredit. This suggests there are issues in the quality of the underlying data or issues with the integrity of the experiment.

We identify numerous sampling errors and measurement errors. The measurement errors are due to inconsistent survey data, faulty variable recoding and a number of coding errors. In particular, the authors collected information from the microcredit institution's information system and appended it to the survey data. Their demonstration relies essentially on this administrative data, which proves to be largely inconsistent with the borrowing information collected by the surveys. The authors' explanations for the differences between survey data and administrative data are implausible in most cases. Handling the coding errors and measurement errors that can be addressed using the available data alters the average treatment effect coefficients and significance tests. However, these rectifiable errors are relatively well balanced between treatment and control groups and their correction does not, in itself, disqualify the main conclusions of the first part of the published article. Yet the measurement errors do raise major concerns about the reliability of the second part (externalities and LATE), which is based on inconsistent administrative data.

The conclusions of the published article are further called into question when sampling errors are also taken into consideration. Households were sampled based on their answers to a short preparatory survey, but data collected from the same households on the same variables at baseline differs considerably. The borrowing propensity score used as the sampling criterion at baseline fails to predict borrowing and is at odds with the revised borrowing propensity scores used as sampling criteria in a second stage to add new households at endline. The average number of household members grew from 5.17 to 6.13 between the baseline and endline surveys. The gender and age composition of one fifth of the households interviewed at baseline and re-interviewed at endline differs to such an extent that it is not plausible that the same units were re-interviewed in these cases. These sampling errors undermine both the internal and external validity of the RCT. They also cast doubt over what was tested; whether it was increased access to microcredit in the treatment group, credit rationing in the control group or substantial variations in other credit sources.

We conclude that this RCT lacks both internal and external validity.

Our understanding of these shortcomings is that they are largely due to poor quality survey data. Data quality and sampling integrity are systematically analysed for standard surveys (such as Demographic and Health Surveys and Living Standards Measurement Surveys) and are reported in the survey reports' appendices. This does not appear to be common practice for most RCT ad-hoc surveys and was not the case with CDDP. It would seem appropriate to align survey methods and practices used for RCTs with the quality standards established for household surveys conducted by national statistical systems (Deaton 1997; United Nations Statistical Division 2005). This implies adopting sound unit definitions (household, economic activity, etc.), drawing on nationally tried-and-tested questionnaire examples, working with professional statisticians with experience of quality surveys in the same country (ideally nationals), properly training and closely supervising survey interviewers and data entry clerks, and analysing and reporting measurement and sampling errors.

This would also entail taking seriously the question of local context and imperfect RCT implementation process. In their article, CDDP cite 17 references: nine RCTs, four on econometric methodology, three non-RCT empirical studies from India and one economic theory paper. No reference is made to other studies on Morocco, microfinance particularities or challenges encountered with this particular RCT. This is especially surprising in the case in hand, since this RCT was a subject of debate and a number of published papers, including in well-regarded journals, prior to the article by CDDP, all seeking to constructively comment on and contextualise this Moroccan RCT (Bernard, Delarue, and Naudet 2012; Doligez et al. 2013; Morvant-Roux et al. 2014; Pamies-Sumner 2014). Morvant-Roux et al. (2014), in particular, built on an extensive literature review on borrowing in rural Morocco and their own qualitative empirical data to improve our understanding of microcredit take-up patterns in treatment and control villages. Among other criteria, they found strong collinearity at village level in terms of agro-ecological settings, land ownership structures and the socio-political relationship with Moroccan Kingdom institutions. It would be particularly interesting to conduct a reanalysis of CDDP based on compound variables that classify the villages along these criteria.

9 References

Angelucci, Manuela, Dean Karlan, and Jonathan Zinman (2015). "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco." *American Economic Journal: Applied Economics* 7 (1): 151–82.

Attanasio, Orazio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart (2015). "The Impacts of Microfinance: Evidence from Joint-Liability Lending in Mongolia." *American Economic Journal: Applied Economics* 7 (1): 90–122.

Augsburg, Britta, Ralph De Haas, Heike Harmgart, and Costas Meghir (2015). "The Impacts of Microcredit: Evidence from Bosnia and Herzegovina." *American Economic Journal: Applied Economics* 7 (1): 183–203.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan (2015). "The Miracle of Microfinance? Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 7 (1): 22–53.

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; Guérin, Isabelle; Morvant-Roux, Solène; Roubaud, François (2019). "Estimating microcredit impact with low take-up, contamination and inconsistent data: Replication study code and data." Version: 1. *International Journal for Re-Views in Empirical Economics*. Dataset. doi: 10.15456/iree.2019071.090421.

Bernard, Tanguy, Jocelyne Delarue, and Jean-David Naudet (2012). "Impact Evaluations: A Tool for Accountability? Lessons from Experience at Agence Française de Développement." *Journal of Development Effectiveness* 4 (2): 314–27. doi: 10.1080/19439342.2012.686047.

Bédécarrats, Florent, Isabelle Guérin, and François Roubaud (2017). "All That Glitters Is Not Gold. the Political Economy of Randomized Evaluations in Development." *Development and Change*. doi: 10.1111/dech.12378.

Chen, Greg, Stephen Rasmussen, and Xavier Reille (2010). *Growth and Vulnerabilities in Microfinance*. Focus Note. Washington DC: CGAP. www.cgap.org/gm/document-1.9.42393/FN61.pdf.

Clemens, Michael A (2017). "The Meaning of Failed Replications: A Review and Proposal." *Journal of Economic Surveys* 31 (1): 326–42.

Cling, Jean-Pierre, Stéphane Lagrée, Mireille Razafindrakoto, and François Roubaud (2014). *The Informal Economy in Developing Countries*. Vol. 112. Routledge.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté (2015). "Estimating the Impact of Microcredit on Those Who Take It up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics* 7 (1): 123–50. doi: 10.1257/app.20130535.

Deaton, Angus (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore, MD: World Bank Publications.

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

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. doi: 10.1257/pol.4.4.68.

Direction de la statistique (2005a). *Recensement Général de La Population et de L'habitat de 2004: Population Légale*. Rabat: Haut Commissariat au Plan du Maroc.

——— (2005b. *Recensement Général de La Population et de L'habitat de 2004: Caractéristiques Démographiques et Socio-économiques de La Population*. Rabat: Haut Commissariat au Plan du Maroc.

——— (2007). *Enquete Nationale Sur Les Revenus et Les Niveaux de Vie Des Ménages 2006/2007 : Rapport de Synthèse*. Rabat: Haut Commissariat au Plan du Maroc.

——— (2015). *Recensement Général de La Population et de L'habitat de 2014 : Population Légale*. Rabat: Haut Commissariat au Plan du Maroc.

Doligez, François, Florent Bédécarrats, Emmanuelle Bouquet, Cécile Lapenu, and Betty Wampfler (2013). "Évaluer L'impact de La Microfinance : Sortir de La 'Double Impasse'." *Revue Tiers Monde*, no. 213: 161–78. doi: 10.3917/rtm.213.0161.

Duvendack, Maren, Richard Palmer-Jones, and W. Robert Reed (2017). "What Is Meant by' Replication' and Why Does It Encounter Resistance in Economics?" *American Economic Review* 107 (5): 46–51.

D'Espallier, Bert, Marc Labie, and Philippe Louis (2015). "Microcredit Crises and Unsustainable Growth: A Management Perspective." In *The Crises of Microcredit*, edited by Isabelle Guérin and Marc Labie. London: Zed Books.

Hejjaji, El Mehdi (2010). *Analyse Des Retards de Remboursement : Cas Al Amana Microfinance-Maroc*. Dakar: Centre Africain d'Etudes Supérieurs de Gestion. bibliotheque.cesag.sn/gestion/documents_numeriques/M0020MAM12.PDF.

Jatteau, Arthur (2016. *Faire Preuve Par Le Chiffre ? Le Cas Des Expérimentations Aléatoires En économie*. Paris: ENS.

Loiseau, Justin, and Claire Walsh (2015). *Where Credit Is Due*. J-PAL and IPA Policy Bulletin. povertyactionlab.org/sites/default/files/publications/where-credit-is-due.

Morvant-Roux, Solène, Isabelle Guérin, Marc Roesch, and Jean-Yves Moisseron (2014). "Adding Value to Randomization with Qualitative Analysis: The Case of Microcredit in Rural Morocco." *World Development* 56 (April): 302–12. doi: 10.1016/j.worlddev.2013.03.002.

Ogden, Timothy N., ed (2017). *Experimental Conversations: Perspectives on Randomized Trials in Development Economics*. Cambridge, Massachusetts: The MIT Press.

Pamies Sumner, Stéphanie (2015). *Development Impact Evaluations*. Paris: AFD. www.afd.fr/en/development-impact-evaluations-state-play-and-new-challenges.

Pamies-Sumner, Stéphanie (2014). *Les évaluations d'impact Dans Le Domaine Du Développement - Etat Des Lieux et Nouveaux Enjeux*. A Savoir 27. AFD. www.afd.fr/webdav/site/afd/shared/PUBLICATIONS/RECHERCHE/Scientifiques/A-savoir/27-A-Savoir.pdf.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org.

Rozas, Daniel, Karine Pinget, Mohammad Khaled, and Sarah El Yaalaoui. (2014). *Ending the Microfinance Crisis in Morocco: Acting Early, Acting Right*. Washington DC: IFC. www.ifc.org/wps/wcm/connect/5e1e5a0047850bdba0d4f5299ede9589/IFC+Morocco+MicroFinance+Crisis+report.English.pdf?MOD=AJPERES.

RStudio Team (2018). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc. www.rstudio.com/.

Selst, Mark Van, and Pierre Jolicoeur (1994). "A Solution to the Effect of Sample Size on Outlier Elimination." *The Quarterly Journal of Experimental Psychology* 47 (3): 631–50.

Sukhtankar, Sandip (2017). "Replications in Development Economics." *American Economic Review* 107 (5): 32–36.

Tarozzi, Alessandro, Jaikishan Desai, and Kristin Johnson (2015). "The Impacts of Microcredit: Evidence from Ethiopia." *American Economic Journal: Applied Economics* 7 (1): 54–89.

United Nations Statistical Division (2005). *Household Surveys in Developing and Transition Countries*. New York: United Nations Publications.

Appendix

Appendix 1 : Reclassification of utility credit

In the questionnaire, the '*Other, specify*:' option was followed by a field where the respondent was supposed to give the name of this unspecified source. We present in Table 19 below the occurrences encountered in this complementary variable and their corresponding frequencies.

Table 19: Reclassification of "other" credits that had all been reclassified as "Utility" by CDDP

Collected as	Recoded by CDDP as	Must instead be recoded as	Baseline	Endline	Respondent specified
Other	Utility	Informal	1	2	boucher, epicerie, souk
Other	Utility	Other	572	26	No specification, or: afni, ascam, credit, e2oom, ecd091, en scolaire, hebouss, kayadat, macon, maison de vente, proprietaire ferme, remboursement pour la retraite, societe tene shems, societe tenne shems, trysol
Other	Utility	Other formal	NA	16	ecdam, ecdom, eddom, eedam, eqdom, eqdon, ikdem, ikdom, ikdon, wafsalof
Other	Utility	Utility	225	223	arsilaf electricite, barnchement electricite, branchemenet electricite, branchemenr electricite, branchement d ectricite, branchement d electricite, branchement electricite, branchement electrique, ectricil, elect one, elec one, elect one, electrcite, electricel, electricit, electricit one, electricite, electricite one, eletrul, elictricite, energie, energie solaire, office nationale electricite, o n e, one, one electricite, onep, safac credit, tema sol, temasol, temsol, tenasol, tenesol

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

We see in Table 19 that, at baseline for instance, a specification corresponding to a utility company was provided in 29% of the cases, but in the others, the specifications corresponded to other

types of sources (local stores, consumer lending, real estate purchase, etc.) or were missing. This indicates that, both at baseline and endline, credits registered as '*other*' should not have been systematically reclassified as '*utility credit*'.

Appendix 2: Code excertps of the coding errors explained in Section 3

A.2.1 Credit from other MFIs was omitted at baseline

The Stata code section the authors used to compute total access to credit and borrowed amount was, at baseline:

- For active loans: `egen aloans_total = rowtotal(aloans_alamana aloans_oformal aloans_informal aloans_branching);` (BL:52)
- For loans that matured in the last 12 months: `egen ploans_total = rowtotal(ploans_alamana ploans_oformal ploans_informal ploans_branching);` (BL:113)

At endline, the same script section became:

- For active loans: `egen aloans_total = rowtotal(aloans_alamana aloans_oamc aloans_oformal aloans_informal aloans_branching);` (EL:138)
- For loans that matured in the last 12 months: `egen ploans_total = rowtotal(ploans_alamana ploans_oamc ploans_oformal ploans_informal ploans_branching);` (EL:158)

A comparison of baseline and endline codes reveals that, at baseline, the variables 'aloans_oamc' (i.e. household's number of outstanding loans from other MFIs) and 'ploans_oamc' (i.e. household's number of loans from other MFIs that matured in the last 12 months) were omitted when creating 'aloans_total' and 'ploans_total' variables, which were in turn summed into 'loans_total' (i.e. the total number of loans taken by each household). This means that the loans from other MFIs were not taken into account when reporting access to credit and assessing the balance between treatment and control groups at baseline.

The same mistake was made for analogous 'aloansamt_total' and 'ploansamt_total' variables, which correspond to the total amount borrowed by each household.

A.2.2 Only outstanding loans were taken into account at baseline

Section 5.1.3 discusses the code used by CDDP to count the number of loans taken out by each household from different source categories: AAA, other MFIs, other formal sources, informal sources and utility companies.
First counted were loans outstanding at the time of the survey ('aloans_[SOURCE]', where [SOURCE] corresponds to each type of source). Second counted were loans not outstanding at the time of the survey, but outstanding in the past 12 months ('ploan_[SOURCE]'). Third, the two previous categories (aloans_[SOURCE] and ploans_[SOURCE]) were summed up to obtain the total number of loans outstanding in the past 12 months (loans_[SOURCE]). Yet it is not the total number of loans that was taken into account in the analysis.

What was taken into account by CDDP is a dummy version of the loan count. In other words, a new variable (named 'borrowed_[SOURCE]') was created for each source category. This variable takes the value '*0*' if the household had no loan from the source category in the last 12 months. It takes the value '*1*' if the household had one or more loan from the source category in the last 12 months. There is, however, an error in the way this variable was computed at baseline.

This is the code used to produce the 'borrowed_[SOURCE]' variables at baseline (BL: 171-176):

```
*** DUMMY of loans over the period ***;
foreach var in alamana oamc oformal informal branching total oformal2{;
    gen borrowed_`var'=0 if loans_`var'!=.;
        replace borrowed_`var'=1 if aloans_`var'>=1 & aloans_`var'!=.;
    };
```

The reader will notice that what is transformed into 1 or 0 are the variables 'aloans_[SOURCE]' (starting with "a"), that is, only the loans that were *outstanding at the time of the survey*.

On the other hand, this is the code that was used to produce the 'borrowed_[SOURCE]' variables at endline (EL: 216-221):

```
*** DUMMY equal to 1 if had a loan over the period ***;
foreach var in alamana oformal informal branching oamc total oformal2{;
    gen borrowed_`var'=0 if loans_`var'!=.;
        replace borrowed_`var'=1 if loans_`var'>=1 & loans_`var'!=.;
    };
```

The reader will notice that what is transformed into 1 or 0 are the variables 'loans_[SOURCE]' (not starting with "a"), that is, the loans that were outstanding at the time of the survey and also the loans that were not outstanding at the time of the survey, but were outstanding in the previous 12 months. In other words, it includes all loans that were *outstanding in the 12 past months*.

A.2.3 Recoding of "other" credit

When recoding the credit variables presented in 3.1.4, CDDP used the following script in both baseline and endline do-files:

`gen branching 'j' = (i3_'j' == 16 | i3_'j' == 17);` (BL:43, EL:92)

This means that all sources registered as '*Other, specify:*' were reclassified as '*Utilities credit*'.

A.2.4 'Tractors' and 'reapers' removed from asset appraisal at endline

The Stata script used by CDDP to compute agricultural assets at baseline includes the following code:

- `egen asset_agri=rsum(ag_1-ag_16);` (BL:269)

The script segment used for the same measure at endline is written as follows:

- `egen asset_agri=rsum(ag_3-ag_16);` (EL:371)

The fact that `ag_1` was replaced by `ag_3` means that, at endline, the assets indexed number 1 and 2 in the survey questionnaire (i.e. tractors and reapers) have disappeared from the sum of agricultural asset values calculated for each household.

Appendix 3: List of coding errors with minor incidence on impact estimates

This appendix presents a series of measurement and coding errors that were only mentioned in the paper. The coding errors could be corrected and did not substantially nor significantly alter the estimated impacts. The measurement errors were limited and magnitude, but provide an additional illustration of the reliability of the data used by CDDP.

A.3.1 Debatable amortisation rule for asset expenses

CDDP computed a series of variables to capture investments in different activity categories: investment in livestock activities, in agricultural activities and in business activities. They also total these investments in activity categories in a variable 'inv_total', which is one of the main outcome variables on which impact is estimated. All investments correspond precisely to expenses that are also included in the 'expense_livestock', 'expense_agri' and 'expense_business' variables expense impact estimations. There is, however, one notable exception: at endline (EL:736-740), purchases of agricultural assets for an amount over MAD 10,000 (all corresponding to tractors, reapers, cars and trucks) are divided by 10. For instance, a tractor purchased for MAD 60,000 is counted as MAD 6,000. This is no mention of it in the paper, but this presumably corresponds to amortisations. However, this is inconsistent for four reasons:

- No such rule was applied to compute baseline expenses, as reported by CDDP (Table 1);
- No amortisation rule was defined for any other investment in any durable assets;
- Other investments for amounts over MAD 10,000 in business assets (cars and trucks) were not amortised;
- One-tenth of all assets with a value over MAD 10,000 purchased in the last nine years should also have been counted in expenses, but this could not be the case as the recall period for asset purchase was only 12 months.

A.3.2 Miscalculation of livestock assets

The following segment of code is used to appraise livestock assets, both at baseline (BL:307-314) and endline (EL:412-422):

```
* Value of stock of livestock assets;
    foreach j of numlist 1(1)3 { ;
        gen assetlive`j'=0;
        gen unitprice`j' = f4_`j' / f2_`j' if f4_`j'>0 & f4_`j'!=. &
         f2_`j'>0 & f2_`j'!=.;
        sum unitprice`j' if unitprice`j'>0, detail;
        replace assetlive`j'=r(p50)*f2_`j' if f2_`j'>0 & f2_`j'!=.;
        };
        gen assetlive4 = f4_4 if f4_4>0 & f4_4!=.;
        gen assetlive5 = f4_5 if f4_5>0 & f4_5!=.;
    egen asset_livestock = rsum(assetlive1-assetlive5);
```

This script creates three variables in the form of 'unitprice1', 'unitprice2' and 'unitprice3' to compute median prices of each asset type and inserts them between 'assetlive3' and 'assetlive4'. The last line of the script sums all variables according to their location in the Stata dataset, starting with 'assetlive1' and ending with 'assetlive5'. We understand that the authors' intention was to sum only the variables 'assetlive1', 'assetlive2', 'assetlive3', 'assetlive4' and 'assetlive5', but they unintentionally included 'unitprice1', 'unitprice2' and 'unitprice3' in this total. In other words, they mistakenly added one unit price to each asset type when appraising the value of livestock assets owned by households.

A.3.3 Subset of business income not taken into account at endline

The following script is used to capture service sales at endline:

```
foreach j of numlist 1(1)6 {;
foreach i of numlist 1(1)4 {;
replace sale_business = sale_business + g35_`j'_`i'*12 if g35_`j'_`i' !=.
 & g35_`j'_`i' >=0;
};
};
```

It loops over service sales (`g35`) activities (`j`) 1 to 6 and over items (`i`) 1 to 4. However, there are as many as six items in the database. Items 5 and 6 are not accounted for.

A.3.4 Confusion between prices before, during and after harvest

For each agricultural product (cereals, fruit tree production and vegetables), median prices before, during (for fruit) and after harvest were computed and imputed for all production for which the transaction price was not registered. However, there are a number of errors at endline:

- Prices after harvest were mistakenly imputed to cereal sales before harvest at endline (EL:581);
- Prices before harvest were mistakenly imputed to cereal savings, i.e. cereal kept after harvest (EL:837); and
- Prices before harvest were mistakenly imputed to tree sales during and after harvest (EL:610-618).

The errors listed above concern only the endline preparation do-file. The sections on sales and savings of cereals, fruit tree production and vegetables for baseline preparation are also plagued by errors: some item types are mysteriously not taken into account (e.g. BL:477 for cereals: only half of the cereal types are included) and the evaluation of (frequent cases of) items with missing transaction prices is inconsistent (sometimes not accounted for and sometimes with a price before or after harvest). These errors at baseline have an effect on the balance test between treatment and control villages put forward by CDDP (Table 1).

A.3.5 Incomplete and inconsistent information on control variables

The control variables include variables from the baseline survey: number of household members, number of adults, age of household head, household does animal husbandry ('*yes*' or '*no*'), household does other non-agricultural activities ('*yes*' or '*no*'), and household had an outstanding loan in the last 12 months ('*yes*' or '*no*').

They also include dummies for whether the spouse responded to the survey, and whether another household member (excluding the household head) responded to the survey. Missing values for all variables are converted into 0 and dummy variables are created for each of these variables where a value is missing.

Some missing and absurd values are found for the controls, albeit in small numbers. For instance:

- 48 households are registered as having more than one head at baseline;
- The age of the household head is missing in 28 cases at baseline and 20 household heads are registered as being under 10 years old;
- Four households have no members at baseline, and six at endline.

These faulty or missing variables could be considered as relatively low considering the sample size. However, this does illustrate that no serious data cleaning was undertaken, even for the most basic variables. In ordinary surveys, and especially in high-quality surveys, such minimal requirements are systematically met.

A.3.6 Coding errors on control variables

Due to the coding error described in 3.1.2, the "had an outstanding loan over the previous 12 months at baseline" variable does not include credit from other MFIs.

There are 28 missing values for the household 'head_age' variable, four for the 'members_resid' variable and nine for 'nadults_resid'. In principle, the variables corresponding to whether these values are missing – respectively 'head_age_d', 'members_resid_d' and 'nadults_resid' – should take the value 1. But the code in AN failed to flag them properly.

Appendix 4: Illustration of household composition mismatch between baseline and endline

Tables 20 and 21 below provide a simple illustration of the first five lines in the original dataset classified as totally or mostly inconsistent by the algorithm presented in Section 5.1. The left side columns present the ages and gender of those household members at baseline and the right side columns present the age and gender at endline of the (in principle) same household. The reader can observe the discrepancies: no plausible narrative could explain such transformation in household composition.

Table 20 and Table 21 report the ages of all members at endline and baseline, the total number of members at baseline, the number of mismatches and the matching category in which the household was classified. A couple of lines are sufficient for readers to be able to assess the consistency of the computation we ran. The third case in Table 20 has no members at baseline. This is one of six occurrences in the entire dataset where no information on members was entered at baseline (see Section 5.2.1).

Table 20: First occurrences of household compositions classified as mostly inconsistent

Household identifier	Age of female members at BL	Age of female members at EL	Age of male members at BL	Age of male members at EL	Number of members at BL	Number of inconsistencies between BL and EL	Matching status
001065	48	33, 4	52, 10	44, 12, 9	3	6	Mostly inconsistent
001067	58	60		24	1	1	Mostly inconsistent
001070	26	28, 45, 28, 26, 22	30	33, 59, 11, 1	2	6	Mostly inconsistent
001089	25	58, 28, 10, 7, 32, 13	37, 3	68, 41, 2, 35, NA	3	9	Mostly inconsistent
002015	55, 23, 12	59, NA, NA, 30, 2	60, 22, 21	63, 24, 23, 35, 12, 6	6	6	Mostly inconsistent

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

Table 21: First occurrences of household compositions classified as no match

Household identifier	Age of female members at BL	Age of female members at EL	Age of male members at BL	Age of male members at EL	Number of members at BL*	Number of inconsistencies between BL and EL	Matching status
001091			30		1	1	No match
005008	36, 3		39		3	3	No match
007053	50, 28, 25, 18, 1	32, 1	50, 18	32, 5, 4	7	8	No match
013081	28, 3, 2, 1		36, 6		6	6	No match
014044	42, 9	46, 22	45, 15, 12, 7	50, 20, 12	6	10	No match

Source: Our analysis using CDDP microdata retrieved from baseline and endline surveys.

Appendix 5: Scoring factors that were attributed each variable to compute borrowing propensity scores

The scoring factors presented below correspond to the regression coefficients of the borrowing propensity models built for the subsequent scores. CDDP only provide the coefficients for score 1 in their article (Crépon et al. 2015: Table A1). However, knowing the different scores for each observation and the variables included in the models, we rerun the regressions for the four scores. As the recomputed scores are a perfect fit with the initial scores at individual level, we are quite confident that the models used by CDDP are similar to ours. The results are presented in Table 22. Levels of significance for the coefficients are reported as p-values are equal or very close to 0 ($<$ 0.001%).

Observation of Table 22 indicates that the coefficients attributed to each scoring variable drastically change from one score to the next, denoting a lack of estimation robustness. Some of them become non-significantly different to 0 and vice versa.

Table 22: Contradictory coefficients attibuted to propensity determinants for each score

Variable	Score 1	Score 2	Score 3	Final score
does trading as self-employment activity	0.846	1.159	-0.25	0.029
owns land	-1.588	0.598	0.017	0.011
rents land	-1.992		-0.178	-0.054
have not bought agriculture productive assets over the past 12 months	-1.048	0.161	-0.271	0.013
# of cows bought over the past 12 months	-2.01	-0.001	0.03	
gets a pension	2.021	-0.33	0.484	0.065
has a radio	1.066	-0.203	0.168	0.008
has a fiber mat	1.574	0.358	-0.61	-0.015
phone expenses over the past month (in MAD)	-0.019	0.002		
clothes expenses over the past month (in MAD)	0.001	-0.002		
had an outstanding formal loan over the past 12 months	0.869	0.175	0.129	0.01
would be ready to form a 4-person group and guarantee a loan mutually	0.57	0.333	-0.115	0.008
would uptake a loan of 3,000 MAD to be repaid in 9 monthly installments of 40	0.593	0.25	0.491	0.043
share # of members with trading, services or handicraft as main activity to #	3.125			
Number of people whose principal activity is craft, commerce, or services?		-0.195	0.109	-0.004
does more than 3 self-employment activities	2.365			
How many activities does this household pursue?		-0.201	-0.042	0.019
ln(amount that would be able to reimburse monthly (in MAD))	0.25			
What is the maximum monthly installment you can afford?		0.001	0.001	
ln(# of olive and argan trees)	0.518			
How many olive or argan trees do you own?			-0.001	
uses sickle & rake (in agriculture)	-0.979			
Do you use the following for agricultural activities: Sickle?		0.191	-0.35	-0.036
Do you use the following for agricultural activities: Rake?		-0.094	0.268	
Number of people living in this household?		0.135	0.056	0.008
Number of people 18yrs or older?		-0.103	0.035	-0.002
What is the distance from your house to the souk (km)?		-0.045	-0.019	-0.001
Do you operate any land that belongs to someone else?		1.106	0.6	0.043

Source: Our analysis using CDDP microdata retrieved from endline survey.

The coefficients reported for score 1, score 2 and score 3 are the result of the three corresponding linear regressions on the CDDP entire endline dataset. The scores computed by CDDP were included in the dataset and we use each one of these variables as the dependent variable for the three subsequent regressions. We use the variables indicated in the first column as independent variables and obtain p-values equal to 0 for each variable and a R-squared equal to 1 for each regression. These coefficients therefore correspond exactly to the factors that were applied by CDDP to the corresponding variables to compute the borrowing propensity scores used to sample households at baseline (score 1) and add new households at endline (score 2 and score 3). The final score was not included in the CDDP dataset. We computed it with a logit regression using the same specification as CDDP in their code (AN: 1537-57). The coefficients reported above are the result of a linear regression using the final score as dependent variable and the variables indicated in the first column as independent variables and controlling for strata dummies (paired villages). The coefficients reported for the final score have a p-value < 0.001 % but not equal to 0 and the R-squared is 0.90.