What Mat[t]hieu Really Did: Rejoinder to Chemin, 2012a[1]

Introduction

Matthieu Chemin (2012a - MC1), in the longer version of his reply, Chemin (2012b - MC2), to our paper published in JDS (Duvendack and Palmer-Jones, 2012a - DPJ), which is in part a replication of Chemin (2008 - Chemin), clarifies a number of things that were not documented in Chemin, finds issues in our paper which he contests, and presents an interpretation of our email exchanges with him which explains why he did not provide all the replication code[2]. In this paper we address the main discrepancies identified by MC between our work and Chemin, and we attempt to use MC's recently provided data and code to replicate Chemin tables 1, 2 & 3.

MC1 writes that our "paper represents unfair criticism of [his] work" (1), in the sense that it appears that while we write that we could not replicate his work, code that MC provides demonstrates it can be. To be purely replicable a paper should provide information that enables the replication; for this type of study this normally requires that the raw data are available and sufficient information is provided in the paper or ancillary materials (appendixes on-line or available on application to the author) that enable pure replication. We could not find sufficient information in Chemin or on application to him, but the information now provided in MC1 & 2 together with the code on MC's www site, goes a long way to enabling this. However, the code actually provided does not replicate Chemin although it appears that identifiable mistakes in the coding and minor changes in the routines may account for the remaining discrepancies. We do not find that the results using the Stata user written "psmatch2" code replicates very closely the results Chemin obtains with his own routines. If one requires that, for a paper to be capable of pure replication it should be replicable from the description given in the published material (including any online appendices, and available code), then Chemin was not. There are many key features of the analysis that were not given and some material given was misleading.

MC also suggests that we failed to communicate with him appropriately about our intention to replicate Chemin, and adduces the email trails between us. He says we "claim that I [MC] would not provide them [DPJ] with my data files…" (MC1:1). While we did not explicitly ask for replication materials, the term replication is used in the opening email to him from us, and three times in all, and all our references to replication were in the context of his paper, providing one would think, ample material to induce the reader to understand our interest in replication. However, it is entirely possible this was overlooked. At no point do we claim MC "*would* not provide" (emphasis added) us with his files; DPJ only state that "Chemin *did not* provide …" (DPJa, galley proof page 6, emphasis added). We discuss this further below, as well as MC's view of replication.

MC provides a great deal of helpful material that allows a detailed understanding of his variable construction, sample selection, and estimation methods, and consequently a fairly close replication of his results. This shows that his paper is largely replicable in a pure sense, although his code needs some changes[3], and these new materials allow identification of some

---

[1] By M. Duvendack and R. Palmer-Jones, School of International Development, University of East Anglia; m.duvendack@uea.ac.uk; r.palmer-jones@uea.ac.uk. The ironic reference is to the existence of a sequel to the last scene of Jean-Paul Sartre's "Roads of Freedom" trilogy.

[2] We use MC to denote both MC1 & MC2. DPJa is our reply to both MC and Mark Pitt, published in the same issue in JDS. MPa and MPb are Mark Pitt's JDS reply and its longer version. PnK is Pitt and Khandker, 1998; RnM, [date], are Roodman and Morduch's "Impact …" documents.

[3] We do not have an exhaustive list of the modifications needed, but, for example, read the "analysis original data.log" in the replication package "microfinance.zip" available on Chemin's www site and at line 273 you

differences in our variable constructions, and, together with MP's comments, mistakes in our coding.

In this paper, we suggest that the email exchanges make it clear that we hoped to replicate his paper. Moreover, we suggest, his work does not stand up to statistical replication, which makes more appropriate choices about how to construct variables and use and interpret the propensity score matching (PSM) results. In particular we continue to show, as we did in DPJ, that all the PSM results seem to be highly vulnerable to unobserved confounding variables which are highly likely to have been present, rendering any interpretation of the results that does not acknowledge this, misleading at best. Finally, we use data and code recently made available by MC to replicate his (MC's) replication of Chemin.

When does replicate mean replicate?

Let us immediately dispose of the issue of interpreting our email exchanges (we return to this again at the end of the paper). Our interest in replication is mentioned in the final paragraph of the opening email from RPJ to MC, and the word "replication" is used three times in all by us in the emails (which MC posts, and we also provide). Perhaps our interests could have been clearer, but we prefer to differ with MC over this. We would argue that our decision to stop communicating with MC after the responses we received on 9/5/2009 was understandable; these responses provided .do files which did not enable a complete replication, and were hard to understand without others that showed how the variable construction code fitted into the overall data construction (as is now clear from the full code provided by MC); also, they did not fulfill the suggestion made by MC to RPJ on 30/3/2009. Our impression was that MC had lost interest in further communication with us, and we feel that this impression is reinforced by the code that MC has now provided, which does not do what he claims it does – it does not in its present form replicate Chemin.

There are no protocols for independent replication of economic studies; we felt there was no need to approach MC when we were ready to submit our paper for publication, though it might have been a good idea; we apologise for any misunderstandings that have arisen as a consequence.

Now we address the points raised in each of MC's headings (the numbering in brackets corresponds to the numbering in his response: MC1). While we think it best to agree to disagree over interpretation of the emails, we appreciate the care that MC has put into clarifying his study and making his code available, and we hope he will provide an updated version which fully replicates his results in due course. It would also be good if there were new results taking account of the issues discussed here.

Sample (1)

We did not restrict our sample to that used by Chemin, because we could never understand how he arrived at his sample. MC raises two issues; the first is the land restriction on the sample, and the second is the estimation of the propensity score for individuals in treatment villages only.

Without the estimation data set or code, we could not know that it was a sample of those who live in households which **operate** less than 0.5 acres. Chemin writes that "[T]he sample is restricted to individuals with less than 0.5 acres" (472). This is unclear since "with" could

---

will see that all the village dummies are dropped because of collinearity. This occurs because they have all been set to zero at line 23 of "analysis.do". It is simple to alter this, but even with village covariates the logits for specification 2 and specification 1 are not that close to those published. Much of the further output depending on specification 3, which is not affected by this problem, is affected by another problem, namely the failure to remove "systematic differences across villages" (Chemin: 475); see further below.

have a number of definitions in the context of land holding in Bangladesh[4], although in the present context the natural interpretation, we suggest, is to use "owned" land, because this corresponds to the common definition of eligibility for MF in Bangladesh at the time of the surveys (Hossain, 1998). At no point in Chemin are the words operated or cultivated used in the context of the sample or any other. It is surprising that operated land should be used, since the eligibility criterion set out by Bangladesh MFIs is well known to have been (at least in principle) 0.5 owned acres (or productive assets to an equivalent value); this is stated clearly by PnK, and by Morduch (1998). Given that the definition is unclear it is natural for us to use owned land since this was the rule operated in principle by MFIs in Bangladesh at this time (as MC acknowledges – see fn. 6). MC continues to use the term "ownership" to define his sample[5] while making it clear that he uses operated land (flopt[6]). There is a large difference between areas owned and operated, and the simple correlation coefficient between them is less than 0.5 in these data.

Because it never occurred to us that Chemin might have used operated land to restrict his sample, and because nowhere does he indicate this, we could never replicate his estimation sample. We find Chemin's use of flopt explicable only if Chemin interpreted "Operational" as owned, or was unaware of the likely difference between owned and operated land in Bangladesh, and its significance (for example, reverse tenancy, smaller land owners renting out land to larger land owners, is not uncommon).

While we did in fact estimate our preferred propensity score (ps2 in our code) using only the treatment villages, in some estimations this restriction was dropped[7]. As discussed more extensively in our response to MP, there are issues in the selection of control groups not only because control villages did not have the possibility of treatment, but also because some households have more than one MF client, sometimes of the same gender, but sometimes of different genders.

Differences in definition of variables (2)

There are indeed many differences in definitions of variables. Matthieu Chemin did send us a file "databaseR1.do", in June 2009; however this did not make much sense to us in the

---

[4] Thus it is common to differentiate between owned and operated land because there is much tenancy and mortgaging of land; also, operated land can vary within a year as land that is cultivated in some seasons is rented out, fallow, or returned to the owner, in others.

[5] "Because microfinance in this sample was limited to individuals owning less than 0.5 acres of land, I restricted the sample of my paper to only those individuals owning less than 0.5 acres of land (as described in Chemin, 2008:472)" (MC1:1).

[6] The PnK dataset description file (codes.pdf) is defined as "Total Operational Land" (p18 0 heading page 53). In the Bangladesh context this is likely to be equivalent to cultivated land, although different surveys might use the term differently. Operational land would generally be owned land + rented or mortgaged in land less rented or mortgaged out and fallow land. Chemin uses mainly the value of flopt in the first round in which the household appears, while we used the average flopt over rounds. The Pearson correlation coefficient between these two variables is 0.81; that between flopt and halab is less than 0.5. Chemin could have identified the discrepancy between halab/halab (defined as "Land Assets" in Codes.pdf:57) checked his understanding by computing operated land using the definition given above.

[7] Thus, in the code we sent MC on 24 May 2012 to replicate DPJ, we define at line 101:

```
global control1 = thanaid <= 24" // all treated vs non-part in treatvill
```

and we use `$control1` to estimate ps2 (our "preferred" pscore) at lines 110—1, ergo:

```
logit elig_defacto_treatpp $Chemin3 $treatvilldumm if control1
predict ps2 [I checked what exactly we sent Chemin on 24 may and hence adjusted
this footnote]
```

However this restriction was inappropriately and inadvertently dropped in some other estimations.

absence of a "ReadMe" file explaining how it fitted into the overall data construction In particular that Chemin restricts himself largely to values of variables for round 1 only, using values for other rounds only for persons who first occur in the later rounds. None of the variables constructed in databaseR1.do yield the sample size or, consequently, descriptives reported in Chemin, until one imposes the "land operated" restriction. Since we could not conceive of using this restriction we could not replicate Chemin's descriptives (to an acceptable accuracy).

Participation

On page 3 MC1 writes that DPJ "exclude participants with more than 0.5 acres". We identify two categories of participants – those who own less than 0.5 acres and those who own more; the sample of all participants is termed "*de facto*", and the sample of those owning less than 0.5 acres "*de jure*" (following the terminology used by Morduch, 1998). The problem MC identifies would lie in the control group if we used the *de jure* sample; but we do not; we overlooked that Chemin restricted his sample to those with less than 0.5 acres. There are other problems with the control group which we inherited from Chemin in that non-borrowing members in households with borrowers can be in the control group for an outcome that is defined at household level. Pitt (MP) identifies this error.

Land (p3, para 2)

As briefly mentioned above, the major problem we had strictly replicating Chemin seems to be in the use of operated rather than owned "land"; while databaseR1.do (available on MC's www site) does define "HHland=flopt", in the absence of explanation and lacking the estimation code which would have used HHland to restrict the sample, it did not occur to us, given our knowledge of the literature on eligibility for MFI membership, and realities of rural Bangladesh at that time, what PnK and Morduch (1998) (and later RnM), wrote and computed, that this could be used to identify eligibility for MFI membership. We used halab (land owned before access to microfinance), which requires halaa for non-participating households, although our construction of halab in DPJ did not realise this. We **never** used operated land for any purpose, although it occurs in our data set.

Age

It is true that we did not use age in years and decimal months; adding months to age makes no significant difference and has minimal effects on descriptive statistics or results. Difficulties in replicating Chemin's descriptive for age is due to the inconsistent use of all individuals for some descriptives, and those ages > = 15 for others. It turns out that Chemin's sample is not restricted by age (i.e. it includes everyone whatever their age).

Gender

Chemin actually means the variable "no adult present"; we do have such a variable in our data set (adultmalepres adultfemalepres) but we did use the number of adult males/females present (no_of_adultfemales, no_of_adultmales) in estimations. The descriptive in Chemin's Table 1 for 'No adult male in HH' (0.024) should have made his use clear to us, although it differs from PnK's figure (0.035). This difference is due the latter being a weighted average of an indicator taking the value 1 if there was no adult male reported in round 1, for the PnK sample[8]. We did use the number of adult males in our propensity score estimations. The impact effects we report below use the dummy variables for whether any adult males of females were present below.

[8] This excludes 41 households owning more than 5 acres.

Highest grade

Chemin nowhere describes his calculation and use of a dummy variable taking the value 1 if the individual had any schooling. The mean of "Highest grade completed" in Table 1 is 2.255 and so cannot be the mean of a dummy variable. The "Education" variable mean is 0.551 and hence is likely to have been the mean of a 0/1 variable if just over 50% of the sample had some education, and is included in Chemin's "preferred" estimation (Spec. 3). Such a categorical variable is computed in the .do file Chemin sent us in June 2009, but it was never clear to us what this file did and what relation to the results reported in Chemin had to it, since it did not seem to replicate them in any straightforward way.

We calculated two variables for education: "Education" and "highest grade". The former correctly uses the highest level achieved by people who are not currently in school, and the level **last** year for those currently in school. However, the "highest grade" variable used in our descriptive and estimations was calculated erroneously (and inadvertently) as a five valued categorical variable in which those still in school were (erroneously) given the highest level. We report results using the "correct" categorizations below.

Other variables

For most of the remaining differences in variables, which MC could identify from the data construction code we sent him, there was no way we could readily know what definition Chemin used, since the paper does not describe the constructions, and we could not use "databaseR1.do" he sent us in 2009 to interpret them for reasons we have already given. We expand on this below; bear in mind we only had Chemin and databaseR1.do to base our interpretations on.

Savings and non-farm enterprise

The descriptive for savings in Chemin's Table 1 reports a mean of 1128.9 (sd 4021); given that we did not know what the sample was we were unlikely to know whether Chemin's savings variable was household or individual level. Similarly for "having a non-farm enterprise"; this is an individual level variable in the data. Since observations in Chemin are at individual level it seems natural to use individual level variables where they are available rather than the household level variable.

Wages

Our calculation of agricultural and non-agricultural wages was flawed by experimentation - to try to match his variables - which later we failed to correct. The term wages is ambiguous since it can, usually, refer to wage rates, or to wage earnings. There are several ways to compute wage rates (and wage earnings), especially where wages are earned from different occupations under different contractual terms, as they are in Bangladesh.

The raw data provide various variables from which wages can be calculated; nowhere in either Chemin or databaseR1.do are the wage variables that Chemin uses described or computed. Chemin's wage variables are computed in MC's "database.do" file (available on his www site but never provided to us previously) which uses the output of databaseR1.do and other files to compute the estimation data file (which is used in MC's analysis.do). These files were not made available to us until MC responded to DPJ in 2012. In DPJ we computed a "wage earnings" variable for both sectors, including all sources of wages provided in the data, but we now compute variables which correspond to "wage rates"[9]. Again, Chemin's wages variables are not documented in the Chemin, and are not computed in databaseR1.do.

[9] See also our response to MP.

Conclusion with regard to differences in variables

Most of the differences in definitions are due to differences in interpretation of the information in the absence of full documentation in Chemin of the variable and or in the code that was provided in June 2009. Our interpretations are plausible, with the exception of the inadvertent error in computing agricultural wage earnings rather than wage rates, and the education dummy variable. Even the wage constructions were partly due to absence of documentation because they arose during the experimentation trying to match Chemin's under-documented descriptives[10].

Differences in propensity score estimation (3)

MC writes that we do not use the control variables he used in Table 1, which were provided in "microfinancefinal.do", which was sent by MC in March 2009. According to MC's email, he uses "microfinancial.do" for teaching PSM, and it was not clear to us how it related to Chemin); we did use the variables described to the best of our ability at the time, and while they are provided in "microfinancefinal.do", those that we (DPJ) did not use are not included in Table 1. As noted already, we did not pay much attention to this file because of earlier failure to replicate the descriptives and other results in Chemin[11].

We do in fact use largely the same variables in our estimation of Chemin's Spec 3 (Table 1)[12], although the constructions differ. We can find no reference to any augmented specification in Chemin, other than in footnotes to Table 1, although one is given in microfinancefinal.do. This is used to calculate "ps2", but this is not identified as the main or preferred propensity score. Indeed, rather confusingly it appears after a simpler specification, and is prefaced by the comment "now we keep only the significant variables and add some more guided by economic theory". We could have paid more attention to the second specification, although the earlier one, which includes all and only the variables of Spec 3, attracted our attention.

We did use Thana variables not village ones. For our preferred propensity score we use Chemin's Spec 3 (not the augmented version).

Difference in data sets

Chemin used a data set provided by MP in 2003; this data set seems to have been a version of that provided on the World Bank www site that we used. According to MC1, the data set Pitt provided "'includes corrections that are not included in the data posted on the World Bank

---

[10] RnM comment: "Reconstructing a complex econometric study from fragmentary evidence about how it was done is an act of science in itself. Hypotheses are generated, then tested against the evidence. Sometimes different bits of evidence appear to contradict each other. No piece of evidence is unimpeachable because all are produced by fallible human beings." (2011b:6) (RnM, 2011. Comment on Pitt's Responses to Roodman & Morduch (2009). Available at: http://www.cgdev.org/doc/RM/R&M%20response%20to%20Pitt.pdf.)

[11] The sample for the data set accompanying microfinancial.do also did not match samples that we could construct, since we failed to grasp his use of operated rather than owned land.

[12] DPJ use:

```
global Chemin3 "male agey agehhhh no_of_adultmales maxed cssv nonfarm livestockvalue hhsize
wageag wagenonag agesq age4";
```

in the estimation of our ps2 we include thana dummies and restrict the sample to treated Thana. (We did not include hhheduc or age^3, the latter not being in Table 1 - we use scohab – spouse (not) present in household – although this is not mentioned in Table 1).

But this is not what MC sent us. His estimation of ps1 includes a dummy for education as shown in the file used in teaching sent to us (MC microfinancefinal.do)

```
……..
logit part HGC sex age landHHpar landHHbro landHHsis landHHSPpar landHHSPbro landHHSPsis
HHland HGChead sexhead agehead adultmale adultfemale scohab village11-village243
predict ps1
logit part HGCdummy savings nfeown livevalue agrincome HHland hhsize nonagrwage agrwage fed
med mar age2 age3 age4 mliv sex age agehead adultmale village11-village243 if thana~=25 &
thana~=26 & thana~=27 & thana~=28 & thana~=29
predict ps2
```

web site'" (MC1:6, fn. 1). If true it is surprising that the relevant corrections were not on the World Bank www site. Some of the corrections listed by MC1 seem to have been corrected in the data set we downloaded, but not others[13]. In any case there seem to be quite a number of differences between the two sets as well as inconsistencies between waves within sets.

Differences in the data sets appear not to make substantive differences to estimation results according to the results of the code provided by MC and our own experimentations, although of course they give rise to troubling differences in descriptive and estimation statistics. One never knows until the results are in whether differences in data will make substantive differences in results, so the presence of known errors in some versions of the data unnecessarily multiplies the difficulties of replication. The discrepancy between Chemin's data set and the World Bank data set, which was known to MC when conducting his analysis, is relevant but undocumented in Chemin, making Chemin strictly non-replicable.

Results (6)

MC writes that Chemin is more or less strictly replicable, and recently provided code and log files on his www site to demonstrate this with the original data he received from Pitt ("original final data") and the World Bank data set. Estimation data sets are not provided, but can be constructed from the data construction code provided. We attempt to check whether the newly provided variable construction and estimation code replicate the results reported in Chemin. To do this we ran MC's data preparation code, which ran without errors, and then ran his analysis code. We checked the log file from our running of the analysis code with the log file provided by MC (for the original data). We only discuss the replication using the "original final data" since this seemingly corresponds to that used in Chemin. After the initial estimations at least some of the results in MC's log files do not correspond to those reported in Chemin and some obvious coding issues are apparent. As a result much time has been spent trying to understand what has gone wrong; we do not resolve all the issues. Because of the time taken we only attempt to assess the main issues at stake.

To be (purely) replicable the results of the (pure) replication should match closely those in the original study. We find several problems in the analysis code, and these are clear that the "analysis original data.log"[14] file reporting the analysis using Chemin's original data does not provide complete pure replication. In this section we report the problems encountered in MC's own replication of Chemin, attempt some corrections, and comment on the implications of both the putative and attempted replications. Some of the minor differences between MC's results estimated using his original data set and those reported may well be due to the improvements in code that MC reports, but without the original code we cannot verify this.

Firstly, it appears that MC's data preparation code reproduces the estimation data set used in Chemin. Our difficulties all relate to the "analysis.do" file. The first point to note is that code does not proceed step by step in a linear fashion to reproduce the tables and figures in

---

[13] Using his footnote 2-4: (1) the edslg observations are not corrected; (2) the flopt missing values are not corrected; (3) the "agehead" corrections seem to have been made, although not exactly as reported by Chemin. The cases can be matched using:

```
gen str2 pidstring=string(pid)
gen str6 nhstring=string(nh)
gen str8 idstring=nhstring+pidstring
gen id=real(idstring)
```

(4) There are 13differences between the WB and MC original loan sizes in rounds 2 & 1 and rounds 2 & 3, and in a few cases the sources of loan also differs.

[14] We use his log file rather than one we generated using MC's code in order to base our comments on the evidence base that MC presumably used in writing his response to DPJ. However, there appear to be only very minor discrepancies between MC's log file and that we produced. Obviously, when we alter his code the results begin to differ.

Chemin. Also, the results are not tabulated; rather they are provided as the full output and the statistics selected for reporting in Chemin have to be found in the output in the log file.

Table 1

We proceed to examine MC's replication of the main results reported in Chemin, starting with Table 1. The problems in the code and log file begin almost immediately because it does not follow the natural sequence of estimating column 1 of Table 1. After some preliminary sample counting and variable construction, the code estimates spec 3 (column 4) of Table 1, which it does to at least 4 decimal places for all reported RHS variables. The code then sets all village dummy variables to zero and predicts the propensity score for treatment and control villages, which are used in Table 2 for the comparison of borrowers from MFIs with individuals in control villages.

However, from this point the problems begin (line 32 of "analysis.do" in the "final original data" folder - hereafter we always refer to this analysis unless otherwise indicated). The code (in analysis.do) proceeds to compute descriptives, and then specification 1 and specification 2 logits (Table 1 columns 2 & 3). The output in the "analysis original data.log" file for specifications 1 & 2 do not match those reported in Table 1 columns 2 & 3.

Column 1 Descriptives

The descriptives reported in MC's log file (lines 171 – 231), unhelpfully not ordered in or restricted to the rows in Chemin Table 1, are close but not equal to those in Table 1 column 1. This appears to be because Chemin's Table 1 reports means for the whole population (including those younger than 15 years old) and those in control as well as treatment villages, while the replication code restricts the sample to those "with" (operating) less than 50 decimals (see Table 1).

Table 1: Chemin Table 1 Descriptives

	Chemin Table 1		Whole population treatment villages only			Whole population all villages		
	mean	sd	a_N	a_mean	a_sd	N	mean	Sd
HGC	2.255	3.173	6334	2.311	3.210	7573	2.253	3.175
Sex	0.512	0.500	7888	0.512	0.500	9397	0.513	0.500
Age	22.327	17.422	7888	22.296	17.384	9397	22.306	17.388
Agehead	42.313	12.383	7884	42.369	12.473	9393	42.315	12.382
Adultmale	0.024	0.153	7890	0.022	0.146	9399	0.024	0.154
landHHpar	0.246	0.560	7736	0.241	0.549	9196	0.247	0.561
landHHbro	0.714	1.224	7742	0.747	1.249	9202	0.712	1.223
HGCdummy	0.551	0.497	7890	0.592	0.492	9399	0.586	0.493
Savings	1128.900	4201.37	7890	1113.072	3259.348	9399	1128.899	4201.371
Nfeown	0.468	0.499	7890	0.496	0.500	9399	0.468	0.499
Livevalue	3273.150	5533.9	7890	3435.461	5700.366	9399	3273.150	5533.854
Hhsize	6.232	2.632	7888	6.244	2.617	9397	6.232	2.632
Nonagrwage	4.023	16.303	7890	3.904	15.937	9399	4.030	16.303
Agrwage	2.987	9.755	7890	2.840	9.623	9399	2.987	9.755
age2	802.000	1109	7888	799.283	1106.758	9397	799.851	1104.293
age4	1874542	5029988	7888	1863610	50164	9397	1859094	4959828

Source: Authors calculations from estimation data set compiled with MC code.

Columns 2 – 4 (logit specifications 1-3)

While specification 3 is exactly replicated by MC's code, specifications 1 & 2 are not. Specification 3 is estimated with a sample size of 5037 people, including individuals less than 15 years. One reason for the failure of MC's code to replicate specifications 1 & 2 is easy to identify, in that the dummy variables for villages are all eliminated from the logit estimation due to collinearity, which occurs because they have not been returned to their original values

after the prediction of the propensity score at lines 27-31 (of analysis.do). Returning these dummies to their original values and re-running the estimation produces results that approximate those in Chemin Table 1 columns 2 & 3, but not as closely as those in column 4, and not as closely as is desirable for replication. At the moment we do not have an explanation for the discrepancies between our replications and Chemin's Table 1 columns 2 & 3 (see Table 2)[15].

One evident problem is that we have different numbers of observations. Without further information from MC we do not pursue replication of Specs 1 & 2 because they play not further role in Chemin.

Table 2: Replication of Chemin Table 1 cols 2-4

			Logit results: dep var "participation"		
	N	Mean(sd)	Spec. 1	Spec. 2	Spec. 3
HGC	7573	2.253	0.0635**	0.0401	
		(3.175)	(0.0254)	(0.0310)	
sex	9397	0.513	-0.862***	-1.376***	-1.136***
		(0.500)	(0.108)	(0.163)	(0.128)
age	9397	22.306	0.0507***	1.270***	1.065***
		(17.388)	(0.00353)	(0.225)	(0.159)
agehead	9393	42.315	0.181	0.122	
		(12.382)	(0.119)	(0.136)	
adultmale	9399	0.024	-0.00291	-0.0698	
		(0.154)	(0.0563)	(0.0633)	
landHHpar	9196	0.247			0.336***
		(0.561)			(0.113)
landHHbro	9202	0.712	-0.0410***	-0.0218***	-0.0136**
		(1.223)	(0.00559)	(0.00796)	(0.00553)
HGChead	5977	1.997	1.906	2.766*	0.832***
		(3.012)	(1.239)	(1.516)	(0.308)
HGCdummy	9399	0.586		0.000155***	0.000162***
		(0.493)		(0.0000365)	(0.0000300)
savings	9399	1128.899		0.712***	0.630***
		(4201.371)		(0.159)	(0.111)
nfeown	9399	0.468		0.0000268	0.0000545***
		(0.499)		(0.0000278)	(0.0000181)
livevalue	9399	3273.150		-0.0879**	-0.147***
		(5533.854)		(0.0368)	(0.0285)
hhsize	9397	6.232		-0.00398	-0.00552*
		(2.632)		(0.00378)	(0.00296)
nonagrwage	9399	4.030		0.0143**	0.00969**
		(16.303)		(0.00603)	(0.00468)
agrwage	9399	2.987		-0.0334***	-0.0279***
		(9.755)		(0.00815)	(0.00597)
age2	9397	799.851		-0.00000138**	-0.00000116**
		(1104.293)		(0.000000643)	(0.000000501)
age4	9397	1859094.081		0.0401	-1.136***
		(4959828.165)		(0.0310)	(0.128)
N			3450	3450	5037

Standard errors in parentheses. * p<0.10 ** p<0.05 * p<0.01

Source: Authors calculations.
Note: the results for Specifications 1 & 2 do not correspond to those reported in MC's log file because those results are estimated without village level fixed effects (lines 315-340; 432-485). Results for Spec 3 are close.

[15] We do not report the values in MC's log file because of their evidently misleading nature.

Distribution of propensity scores

Figures 2 to 4 in Chemin report the distributions of propensity scores. MC's code does not produce these density plots, and the x axis of the plots in Chemin is not meaningfully labeled. We cannot compare our graphs exactly with Chemin, but it is evident that the match is not as close as one would like. We note particularly the apparently higher concentration of cases around 0.5 on the x-axis in the graph for treated individuals.

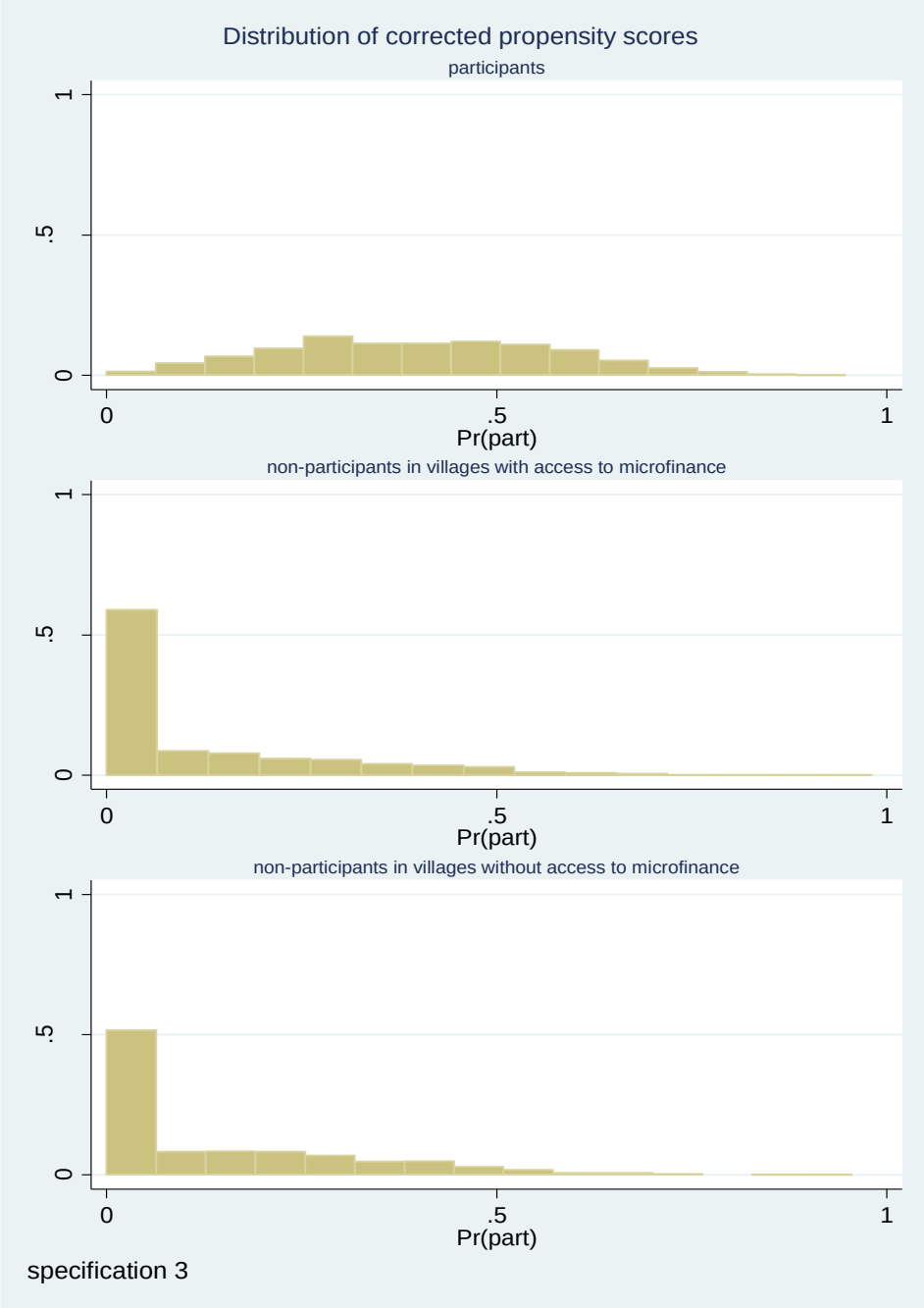

Figure 1: Propensity scores by access to microfinance and participation[16]

These propensity score estimates are surprising in that the distributions for non-participants in treatment villages and in control villages are significantly different to that of the participants in treatment villages (recall that the sample is restricted to those with less than 0.5 acres).

[16] Figure 1 is drawn to approximate Figure 2-4 in Chemin, using Chemin's sample (HHland <= 50).

Propensity score matching results

MC points out that the PSM results of DPJ do not match those in Chemin for reasons discussed in MC and above. Also, DPJ used the STATA psmatch2.ado command for matching, although it is now clear that Chemin used his own purpose written programmes. This was not documented, and we assumed Chemin used one of the available Stata routines which were available in the early 2000s[17]; furthermore, MC used psmatch2 in the teaching code he sent us in March 2009 (microfinancefinal.do).

Here we are concerned with MC's replication code and the results generated with it. We focus only on the kernel estimation results reported using psmatch2[18], those reported in MC's log file, and those in Chemin. Our results using uncorrected expenditure as the dependent variable (which is natural when restricting the comparison to treatment villages only) for the treatment effect match that reported in Chemin and in MC's log file (see Table 3) (if not to 3 decimal places or significance level); sample sizes are the same. However, the results using the corrected "pure" expenditure measure[19] diverge considerably from those in Chemin, Table 1 and in MC's log file.

Table 3: Estimated Impact of Microfinance: replication of Chemin's Kernel (bw = 0.05) estimates

	Chemin	Log file psmatch2 results	Including individuals from treated households in control group		Excluding non-participating individuals in borrowing households from control group	
			Expenditure			
Expenditure			Un-corrected	corrected	Un-corrected	Corrected
Control group	(1)	(2)	(3)	(4)	(5)	(6)
Treatment villages	-0.039*	-0.040[a]	-0.041 (0.050)	-0.024 (0.047)	-0.027 (0.048)	-0.004 (0.061))
Control villages	0.028***	0.111*[b]	na[c]	na	0.108** (0.050)	-0.152*** (0.051)

Notes: bootstrapped standard errors in parentheses. * p<= 0.10, ** p<= 0.05, *** p<= 0.01
Table 1 Row 1 Kernel bw = 0.05
a: line 825 - the result of MC matching programme = -0.033 (line 516)
b: line 940 and the result of MC matching programme = .117*** (line517)
c. no people in control villages are member s of households with MF borrowing

Thus, some of the results produced by psmatch2 in MC's analysis.do file (original data) are close to those published in Chemin (see the "analysis original data.log" files provided on MC's www site). For example at line 825 of this file the "r(att)" comparing treated with controls in the treatment villages is reported as -0.040 for a bootstrapped kernel matching with bandwidth 0.05, and the corresponding figure in Chemin, Table 1 Row 1, is -0.039. Other results are not so close; comparing treated individuals with those in control villages MC's .log file reports a figure at line 940 of 0.111 (p< 0.086), while Chemin gives 0.028 (p<0.01) (Chemin Table 1).

[17] pscore was developed in 2002 (Becker and Ichino, 2002; psmatch2 in 2003 (Leuven and Sianesi); nnmatch (ver. 1.3.1) in 2004 (Abadie and Imbens, 2004).
[18] We report stratification estimations in our rejoinder to MP.
[19] "The first row [of table 2] contains results comparing expenditure of participants to expenditure of matched non-participants in treated villages. To make sure that results do not come from systematic differences across villages, the logarithm of per capita expenditure is removed from village effects by regressing this quantity on village dummies from both the programme and control villages only, and then estimating the residual arising from this regression. This quantity is termed the 'pure' logarithm of per capita expenditure since it is now freed from any village level effects. (Chemin, 2008:475).

However, these results are estimated for the "uncorrected" expenditure level[20], and include individuals who do not themselves borrow living in treated households in the treated village control group. Chemin appears not to have realised that his control group included individuals from treated households (any more than we did until MP pointed it out). In Table 3 we report results for the "corrected" expenditure variable and when individuals in treated households are dropped from control groups.

When we estimate the figure given in line 825 with the "corrected" expenditure levels, but including untreated individuals in treated households, we find an ATT of -0.024 rather than the -0.039* reported by Chemin (our figure with the "uncorrected" expenditure is -0.041). For the comparison with matched individuals in the control villages the ATT reported by psmatch2 with the "corrected" expenditure level is -0.152** rather than 0.028*** reported by Chemin (see Table 2 col 3). Thus, removing village level fixed effects from the log of per capita expenditure gives a negative impact relative to control villages, while Chemin reports a positive impact (col. 1).

Columns (5) and (6) report the results dropping people from the control group in households which have a MF loan who do not borrow; this comparison is only relevant for those in treatment villages. Neither for the uncorrected nor the corrected expenditure variables are the estimated treatment effects significantly different from zero[21].

Sensitivity analysis

Since two of the results reported in Table 3 are statistically significant, albeit with opposite signs for the different consumption outcome variables ("uncorrected" and "corrected"), we conduct sensitivity analysis (Rosenbaum, 2002; Ichino et al., 2006). We cannot conduct sensitivity analysis following kernel estimation, but we can perform it following nearest neighbour estimation. To better match kernel estimation we match treatment cases to the 20 nearest neighbours.

[20] We can see no code in which the log of per capita expenditures is computed. Chemin describes his calculation as "freed from any village level effects" (Chemin: 475). In the various analysis.do files on MC's www site, the expenditure dependent variables he uses is calculated as the log of per capita expenditure, and then renamed as "e" around lines 38-44 (log file lines 491 -494). It is easy to correct this and the results with "corrected" expenditure are also reported here.

[21] Using the land ownership (< 50 decimals) sample inclusion criterion and excluding non-participating members of participating households from the control group yields positive and significant impacts within treatment villages for both corrected and uncorrected expenditure. However, for the land ownership sample, the impact of MF compared to individuals in control villages for uncorrected expenditure is positive but not significant while that for corrected expenditure is negative and highly significant.

Table 4: Estimated Impact of Microfinance: sensitivity analysis				
		Kernel bw 0.05	nn (20)	gamma[a]
		(1)	(2)	(3)
Sample	Expenditure	att (se)	att (se)	
Operated land	uncorrected	0.1082** (0.0580)	0.137** (0.062)	< 1.20
	Corrected	-0.1525*** (0.0523)	-0.128** (0.056)	< 1.20
Owned land	uncorrected	0.0555 (0.0532)	0.050 (0.0578)	< 1.5
	Corrected	-0.1014** (0.0493)	-0.157*** (0.052)	< 2.4

Source: Authors calculations.

Notes: a. gamma is estimated using rsens.ado[22]

The results of these estimations are reported in Table 4; for both the land operated and the land owned samples the estimates are of similar sign, order and significance to the equivalent results for kernel estimation reported in Table 3, however, the land operated estimations are highly vulnerable to hidden bias, while the owned land estimate with corrected expenditure is quite insensitive. However, as in Table 3 it has the opposite sign to that reported in Chemin; MF borrowers seem to be worse off than matched households in control villages, when matching and impact estimate is performed by the methods suggested by Chemin.

Replication Studies (7.1)[23]

To sum up our experience with replicating Chemin, so far: There was insufficient evidence provided in Chemin and in our email exchanges, to undertake a successful pure replication. Further, code (and data) provided by MC, still do not allow pure replication. However, the explanations provided by MC together with the code allow us to purely replicate some of Chemin. We can confirm many of the discrepancies in variable and sample construction reported by MC, but correcting this still does not result in full pure replication of the key PSM results reported in Chemin.

The most significant differences in sample construction is due to Chemin's use of operated rather than owned land as the criterion for inclusion; this undocumented use of operated rather than owned land to define the sample complicated pure replication making it nigh on impossible to reproduce the main descriptive and logit results. Replication was further complicated in that the descriptives Chemin reported were for the whole population in all villages, although Chemin implies that they were for those aged over 15 in treatment villages only. Despite using the term "land owned" MC actually uses "land operated" (flopt) and this is nowhere even hinted at in Chemin and is counter to general knowledge about the operations of MFIs in Bangladesh at the time and apparent usage by PnK. The inclusion of all people in descriptives and logit estimations makes little difference to PSM results compared to restricting the sample to adults, because children do not match people who have loans (who are all over the age of 15). A further major unreported difference between the replication text and Chemin is the additional covariates for Spec 3 used in Chemin. It is true that the expanded variable set was present in the do file for student use provided in March

[22] rsens can be used with up to 20 nearest neighbours. The more commonly used rbounds.ado can only be used after single nearest neighbour estimation. rbounds after single nearest neighbour shows sensitivity to a confounding variable at gamma < 1/4
Gamma is the ratio of the odds of having a characteristic that completely confounds the association between treatment and outcome among the treated to its odds among the untreated (Rosenbaum, 2002).

[23] These and related issues are discussed in Duvendack and Palmer-Jones (forthcoming).

2009, but we have already explained why we found this hard to make use of (because we could not replicate the descriptives and logit results in Chemin, Table 1).

Failing to purely replicate Chemin's initial descriptives and logit results we moved from pure to statistical and scientific replication. We did not pursue communication with Chemin after our failures to replicate the descriptives and the rest of Table 1 because by that time we gained the impression he was not interested to provide further assistance (vide the provision of only incomplete code in June 2009). Our communications are discussed further below to make some additional points related to protocols, behavior and communication between replicators and replicatees.

Chemin does not claim to be a replication or PnK, nor does it discuss differences in results other than to mention that he finds "positive, but lower than previously thought, effect of microfinance" (p.463). This is a rather modest account of the differences between PnK and Chemin, which are in fact rather more extensive; indeed the results in row 1 of Table 2 show that compared to controls within the same village, the comparison suggested for PSM by MP, participants are implied to be worse off[24]. Compared to individuals in the control villages, a comparison MP does not comment on in his reply to DPJ, Chemin implies that participants are better off. However, our replication results (Table 3) do not support this conclusion.

There are of course a number of problems with Chemin's approach which could invalidate these comparisons, including the inclusion of potentially endogenous RHS variables[25] in the estimation of the propensity score[26] (as reiterated by MP), and the use of land operated rather than land owned as a selection criterion. Neither of these constructions is appropriate for PSM, although DPJ followed Chemin in the former, and in using potentially endogenous variables in the propensity score estimate (again, as pointed out by MP). DPJ made further errors, including failure to use village rather than Thana fixed effects, and in not "correcting" the propensity scores when extrapolating to control villages, or expenditure levels for village fixed effects (but as we have seen it is not clear that MC does the latter).

Communication with Chemin

MC1 writes that that replications which are not "rigorous, consistent, and done correctly" (8) "can seriously undermine the work of the original researchers" (ibid), and that DPJ had not "followed an appropriate protocol" in conducting their replication. He suggests we were not "upfront" about our intentions, and that had we been he would have "gladly provided all [his] do files … to replicate his results" (9). As mentioned above, we believe that we were adequately open as to what we were planning, and that for reasons unknown to us Chemin did not perceive this.

MC writes that we did not follow an appropriate protocol in our replication, a view strongly shared by MP. It is certainly the case that had we had the full set of code that MC has now made available we could have saved ourselves a large amount of time and effort, and so would have been in our interest. We were strongly deterged from further communication by the delayed and brief emails and limited materials MC sent us in March and June 2009, as we think others would have been. However, we would still have had to deal with those constructions MC made which we find debatable – the use of flopt instead of halab, for example.

---

[24] Chemin [482] comments that this could be due to within village externalities, but it is hard to see how this could make participants actually worse off than controls rather than reducing the size of impacts.

[25] Savings, ownership of a non-farm enterprise (nfeown), value of livestock (livevalue), and agricultural income (agrincome).

[26] "only variables that are unaffected by participation (or the anticipation of it) should be included in the model" (Caliendo and Kopeinig, 2005:6).

Given the potential difficulties in communication between replicators and replicatees (see below), it interesting to read MC's understandings of our communications. Looking at our email trails the point already noted is that RPJ uses the work "replication" in his first message admittedly in regard to "the original papers" (both PnK and Chemin "your recent JDS paper", are mentioned). MD uses "replication" twice in her message of 5/6/2009, once in regard to PnK and the other time in regard to Chemin; MC quotes MD's use of the word replication, but sees this as restricted entirely to the outcome variables. A second point we notice is that we did not explicitly make a link between the inquiries of RPJ and MD, although since we came from the same institution and showed very similar interests in PnK and Chemin, it would not have been difficult to infer that we maybe were working together.

Both the opening emails from us were straightforward and expressed our interests. MD was RPJ's PhD student, and was considering a replication of PnK and Chemin, and RPJ was preparing an Impact Evaluation workshop and considering using PnK as an example. We are surprised that MC appears not to have put two and two together[27], and that he seemed not to recall his suggestion in his email to RPJ of 30/3/2009 that he "send you [RPJ] my [MC] do file if you wish, on how to merge all this data in a si[n]gle master database, but these .do files are extremely long and hard to understand...". RPJ interpreted the files that MC actually sent as those that MC could or would send, especially as he had sent the same to MD (MD received microfinancial.do on 17/03/2009 and databaseR1.do on 09/06/2009) despite her twice identifying replication of Chemin as an objective. We felt it would be inappropriate to press further, especially in view of the generally significant delays (by email standards) in getting responses from MC[28]. As the work evolved, and problems seemed to emerge, it appeared to be worthwhile not just to undertake a pure replication of Chemin and then exploring the effects of gender of borrower, but to change to reporting the failure to purely replicate the paper, and undertaking something closer to statistical and scientific replication.

In the absence of a formal request for "complete replication files" it is hard to know how much more explicit we could have been, and such a request might well have been seen as impertinent – it would certainly have been unusual. Furthermore, there is virtue in trying to replicate independently using the description given in the original paper, since this allows unintended errors to emerge[29]. On the other hand, clearly it allows the replicators to make their own mistakes or impose their own variable constructions which are at odds with sometimes un-declared constructions of original authors, and lead to misunderstandings and misrepresentations.

Original authors tend to feel threatened by replication, in part because they perceive this as a potential threat to their reputation; replication which correctly identifies flaws in their work, and replication which mistakenly claims to identify such flaws when they are not present, can, of course, harm reputation. There seems less reason to complain about the latter than the former, unless the (mistaken) replication is widely publicised (and even then it may not be the replicators who are responsible for the publicity). This is because journals generally allow original authors to respond to replications, either by invitation in the same issue, or by readily accepting replies in a subsequent issue. In such replies original authors can set the record straight, if anything enhancing their reputation.

On the other hand, deference to original authors can lead to inappropriate suppression or delay in publication of replications which correctly identify flaws in original studies. Original authors can be extremely combative, and through their responses to replications seek not only

[27] The frequent use of "we" by RPJ does not reflect dreams of royalty.
[28] MC did not respond to the request in the PS to the message of 10/6/2009.
[29] For example RPJ found an unintended error, and a debatable variable construction, in replicating the analysis of the DISE data in Jensen and Oster (2009) by writing completely independent code even though he had access to the original code.

to undermine the credibility of the replication, but to deter would be replicators – see Duvendack and Palmer-Jones, forthcoming). The primary intentions of replicators are often not hostile (*pace* Hamermesh, 2007). Rather, it seems many, probably most, replications (if one focuses on those which are done rather than only those which are published) are undertaken to understand better the original work as a basis for extending the original work (or for pedagogic purposes). However, the work may begin to appear as if it could be threatening if problems appear during the replication process or if replicatees respond in ways which indicate lack of cooperation. Encountering or anticipating these reactions (from replicatees), replicators may be cautious in communicating their intentions or progress in their work as it may induce either withdrawal of cooperation or retaliation (see DPJb).

Overall, extensive communication between replicators and replicatees is not always desirable, and can distort outcomes. Replicatees may frequently feel under duress by replication (Dewald et al., 1986; Hamermesh, 2007), and aspersions are cast on the motives of replicators (op. cit.), although no empirical support for this claim is offered. There are no protocols, but they are needed for both replicators and replicatees. Thus there are also no protocols for original authors to enable replication; the currently suggested way in which they can facilitate replication (mandated by the AER) is to deposit estimation data and code at the time of publication.

At the present time it is optional to deposit original raw data and variable construction code. JDS did not and does not have a data and code policy, though it asks whether authors are in principle prepared to make data and code available. Nor did JPE have such a data and code policy at the time PnK was published, although earlier it had had a "Confirmations and Contradictions" section (from 1975 – 1987), and it developed a voluntary data and code archive policy in the last decade. Neither Chemin nor PnK deposited estimation data sets or code; they could expect requests for code (or clarification if the code is unavailable or does not do the job) to bring their conduct into line with current prescriptions (such as the AER data policy).

While deposit of data and code that do replicate the results does not perhaps completely absolving original authors from further communications this would seem to go a long way to meeting their obligations. Etiquette for dealing with further inquiries from replicators would depend largely on whether the deposited code and data do in fact produce the results reported in the original paper. For example, there may be points of clarification, or places where errors seem have occurred, at least in the minds of the replicators. In some cases this may be due to the replicators' limitations, but in other cases it may be because of "real" errors, or debatable differences of opinion, or (real) difficulties in interpreting the code. Again, there are no protocols for communication in these circumstances. Replicators may be being importunate, or they may be correct; each case may be judged on its own merits.

Future directions (7.2)

We agree with MC that sophisticated methods of analysis cannot compensate for the failure of '"the assumption that selection is based on observables"' (Chemin: 464 cited in MC1:10), but we disagree that the RCT approach will provide adequate answers to selection on unobservables in real life for many, probably most, important development interventions (Shaffer, 2011; Deaton, 2010; Harrison, 2011). Hence, methods which do continue to address this issue with observational data must continue to be an important part of the development economists' toolkit. We do need to develop appropriate protocols for original authors original publications, replicators and replicates (original authors when faced with replication) to facilitate replication of computational studies in economics.

Conclusions

DPJ made a number of errors in their replication of Chemin, but these may in part be excused by incomplete reporting of the methods used in Chemin, and unsatisfactory communication about the replication work. However, various problems have been uncovered which are useful for the interpretation of Chemin, DPJ and PnK; these are discussed more extensively in DPJb). Leaving aside the incomplete documentation of variable construction, Chemin's undocumented use of operated land rather than owned land was a major obstacle to replication (and indeed to understanding Chemin).

In the light of MC we have been able to clarify much, if not all of Chemin, and DPJ. The results suggest that Chemin's published results are not completely reproduced by the code Chemin has made available, and that there are significant errors in the models used in the original paper (such as the inclusion of no-borrowing members of borrowing households in control groups, and the inexplicable use of samples defined by land cultivated rather than land owned). In particular, it is likely that that the logit model used by Chemin is inappropriate in using several potentially endogenous variables to estimate the propensity scores (Caliendo and Kopeinig, 2005:5-6), and the low explanatory power of the logit suggests that important variables determining participation may have been excluded giving rise to biased results (ibid). Using a perhaps more appropriate specification of the propensity score estimating model gives very different results (see MPa&b and DPJb). Also, the choice of methods to determine the control groups by Chemin were unsatisfactory, since the comparison with eligible households in treatment villages only almost certainly was vulnerable to unobserved confounding variables (the reasons eligible non-participating households did not become members was in some part likely to have been due to unobservable energies and abilities), while the method comparing treated households with matched households in control villages was liable to placement biases which were effectively uncontrolled by Chemin's method. A methodology for dealing with this problem, that allows pooling all non-participating households in the data, is to use village level covariates rather than village fixed effects. We report our results doing this elsewhere (DPJb).

Communications between replicators and replicatees are subject to misunderstanding (and their behaviors may less than edifying); hence protocols and appropriate norms of behaviour for constructive communication for all parties to the replication need to be developed.

References

Abadie, A. et al., 2004. Implementing Matching Estimators for Average Treatment Effects in STATA. *The STATA Journal*, 4(3):290-311.

Becker, S.O. & Ichino, A., 2002. Estimation of Average Treatment Effects Based on Propensity Scores. *The STATA Journal*, 2(4):358-377.

Caliendo, M. & Kopeinig, S., 2005. Some Practical Guidance for the Implementation of Propensity Score Matching. *Forschungsinstitut zur Zukunft der Arbeit (IZA) Discussion Paper No. 1588, May*.

Chemin, M., 2008. The Benefits and Costs of Microfinance: Evidence from Bangladesh. *Journal of Development Studies*, 44(4):463-484 (Chemin).

Chemin, M., 2012a. Response to "High Noon for Microfinance Impact Evaluation", *Journal of Development Studies*, 48(12) (MC2).

Chemin, M., 2012b. Response to "High Noon for Microfinance Impact Evaluation", online (MC1).

Deaton, A., 2010. Instruments, Randomization and Learning about Development. *Journal of Economic Literature*, 48:424–55.

Dewald, W.G., Thursby, J.G. & Anderson, R.G., 1986. Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, 76(4):587-603.

Duvendack M. & Palmer-Jones, R., 2012. High Noon for Microfinance Impact Evaluations: Re-investigating the Evidence in Bangladesh. *Journal of Development Studies*, 48(12) (DPJ).

Duvendack and Palmer-Jones, 2012a. Reply to Chemin and Pitt. *Journal of Development Studies*, 48(12) (DPJa).

Duvendack and Palmer-Jones, 2012b. Wyatt Earp's High Noon? Rejoinder to Pitt, 2012b. *forthcoming online* (DPJb).

Duvendack, M. & Palmer-Jones, R., forthcoming. Replication of quantitative work in development studies: experiences and suggestions. *Progress in Development Studies.*

Hamermesh, D. S., 2007. Viewpoint: Replication in Economics. *Canadian Journal of Economics*, 40 (3):715-733.

Harrison, G.W., 2011. Randomisation and Its Discontents. *Journal of African Economies*, 20(4):626–652.

Hossain, M., 1988. Credit for Alleviation of Rural Poverty: The Grameen Bank in Bangladesh. IFPRI Research Report 65. International Food Policy Institute: Washington D.C. Available at: http://www.ifpri.org/publication/credit-alleviation-rural-poverty.

Ichino, A., Mealli, F. & Nannicini, T. 2006. From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? Forschungsinstitut zur Zukunft der Arbeit (IZA) Discussion Paper No. 2149, May.

Jensen, R. & Oster, E., 2009. The Power of TV: Cable Television and Women's Status in Rural India. *Quarterly Journal of Economics*, 124(3):1057-94.

Leuven, E. & Sianesi, B., 2003. PSMATCH2: STATA Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Matching. *Available at: http://ideas.repec.org/c/boc/bocode/s432001.html*.

Morduch, J., 1998. Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh. Unpublished mimeo.

Pitt, M.M., 2012a. Gunfight and the NOT OK Corral: Reply to "High Noon for Microfinance" by Duvendack and Palmer-Jones, *Journal of Development Studies*, 48(12) (MPa).

Pitt, M.M., 2012b. Gunfight and the NOT OK Corral: Reply to "High Noon for Microfinance", online (MPb).

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

Roodman, D. & Morduch, J., 2009 (revised 14/12/2011). The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence. *Available at: http://www.cgdev.org/files/1422302_file_Roodman_Morduch_Bangladesh_2.pdf.*

Rosenbaum, P.R., 2002. *Observational Studies.* New York: Springer.

Shaffer, P., 2011. Against Excessive Rhetoric in Impact Assessment: Overstating the Case for Randomised Controlled Experiments. *Journal of Development Studies*, 47(11):1619–1635.