Wyatt Earp's High Noon? Rejoinder to Pitt

Maren Duvendack and Richard Palmer-Jones[1]

Abstract

Mark Pitt (Pitt, 2012a - MPa; see also Pitt 2012b – MPb, jointly MP) does indeed find many problems with our paper ("High Noon for Microfinance Impact Evaluations", Duvendack and Palmer-Jones, 2012a - DPJ), and provides what he terms a replication of our replication of Chemin (2008 - Chemin) and Pitt and Khandker (1998 - PnK). MP also makes a number of comments of a combative and belligerent nature. While owning up to many, but not all, of the problems identified in DPJ, we find significant problems with MP's replication and show that it provides little support for PnK's headline finding that microfinance benefits the poor especially when provided through women. MP's derogatory texts are not conducive to understanding the issues involved and coming to reflective appreciation of what can be learnt from the various studies. These replicatory exchanges, together with those with Matthieu Chemin (DPJb), have provided significant evidence on the lack of validity of Chemin, cast further doubt on the conclusions of PnK, and do not contradict DPJ's conclusion that PnK suffers from a weak research design which does not appear to be remediated by the use of complex analytical methods.

Introduction

Mark Pitt' reply (MPa) to our paper "High Noon" (Duvendack and Palmer-Jones, 2012a - DPJ) is the original a version of that published in JDS (Pitt, 2012b - MPb) after editing by one of the JDS editors[2]. In this paper we provide a more expansive rejoinder than that published in JDS (DPJc). We reply to the "strident" (Hamermesh, 2007) as well as the substantive issues raised by MP, including MP's application of propensity score matching.

We accept we made many errors in our paper, but many of them derive from Chemin and others from the poor state of the PnK data. We show that propensity score matching (PSM), when "properly" applied to replicate PnK may "indicate positive and highly significant results of women's participation in microfinance in every one of the matching estimators", but when subjected to sensitivity analysis[3] they are highly vulnerable to unobserved confounding variables, which are highly likely to be present - i.e. we cannot assume ignorability or unconfoundedness (Caliendo and Kopeinig, 2008). This vulnerability, together with the theoretical reasons (de Aghion and Morduch, 2005) and ethnographic evidence (Fernando, 1997) for expecting such confounding unobservables to be present, would seem to suggest that these results do not provide robust support for the headline findings of PnK. We provide a rather more convincing PSM analysis of the PnK data

[1] School of International Development, University of East Anglia, Norwich, NR4 7TJ, UK. r.palmer-jones@uea.ac.uk; m.duvendack@uea.ac.uk

[2] Not RPJ. The review process of DPJ and the various replies was handled entirely independently by one of the other JDS managing editors.

[3] By sensitivity analysis we mean the application of Rosenbaum bounds; since this currently can only be done for n nearest neighbour matching our conclusion applies to PSM results obtained with this method, although we have no reason to believe that they are not generally applicable. See further below.

(rectifying some problems in MP's work) and conduct sensitivity analysis (SA) , showing that MP's (and PnK's) headline conclusion, which contradicts that in Chemin (2008 – Chemin[4]), is not robust. Our main reasons for asserting this are firstly the SA results, and secondly the debatable mature of a key variable in PnK's identification strategy – the "non-target" dummy variable (see further below). Other critiques of PnK using the same estimation strategy as PnK (Roodman and Morduch, 2011 - RnM; see also below) seem to support the implications we draw from this oeuvre (RnM, 2011; Roodman, 2012), namely that the weak research design is not overcome by the application of sophisticated methods in such a way as to provide credible evidence of the impacts claimed.

Asides

Several of MP's footnotes are of an unpleasant character. MPb footnote 2 is both denigratory ("claims"), and illogical – why should our supposed replicatory failures mean that our assessment of the evaluations of MF are of poor quality? Pitt's orchestration of a character assassination of RPJ to 3ie (see MPa footnote 3, and email communication from MP to 3ie, 14/7/2012, and from Robert Jensen and Emily Oster, on the same date[5]) is further evidence of Pitt's unhelpful behaviour. One wonders what the state of knowledge would have been had Pitt not written in such a combative and belligerent way in response to Morduch (1998) (see Pitt, 1999).

Replication is a reputable activity (Bernanke, 2004) that is crucial to the scientific status of a computation study (Peng, 2011). PnK is widely regarded in policy circles as the capstone study legitimating the belief that MF benefits the poor especially when targeted through women, and thereby given confidence to development practitioners and investors to support and promote this type of MF. However, ethnographic evidence casts doubt on fundamental features of the PnK analysis (and data), for example, documenting widespread mis-targeting, and lack of sustainable benefits for the poorest from MF (see, for example, Fernando, 1997; Goetz and Gupta, 1996; and others). Furthermore, analytical methods of the type used by PnK were and are widely held to be fragile (Leamer, 1983; Manski, 2011). It is crucial that the validity of this work is established both because it is still claimed that it provides robust evidence in support of its headline findings (Pitt and Khandker, 2012), which are still widely embodied in development projects, and because it is an exemplary case of the influence of complex analyses in evidence based policy making. A string of other results based on these data and mostly similar methods have been produced by Pitt and his associates (mainly his students[6]) in support of a number of other interventions (for example about effects of microfinance on contraceptive use, fertility, nutritional status, women's empowerment, etc.). The PnK oeuvre can also be seen as exemplars of rigorous impact evaluation (Khandker et al., 2010). It is in this context that we respond to the criticisms of MP (and Shahidur Khandker (Pitt and Khandker, 2012)), of DPJ, and are not deterred by MP's belligerent texts and correspondence critical of our work, especially when the resources for replication of the original authors (primarily from Chemin but also from PnK) were (and are) in such poor state.

---

[4] Many of the errors in DPJ were first perpetrated in Chemin, a point on which MP does not remark on.

[5] Copies of this correspondence can be provided on request.

[6] Mark Pitt has worked with many of his students and some other collaborators on the PnK data; none that we contacted were, or felt, able to provide us with the data or code that they used. We have no opinions of their actual situations in this regard.

This paper proceeds as follows. First we deal with some of the discourteous comments made in MP; many readers can skip this section. Then we respond to some of the specific comments. We respond to all (we hope) of the remaining substantive comments made by MP in Appendix 1; this too can be skipped since several are effectively the same and we either accept them, generally with explanations, or reject them in what we hope is an open a way.

Following this we replicate MP's replication of our replication; we show that, while providing some helpful material, the data set MP uses has errors, and MP makes at least one mistake very similar to one he criticises us for. Most importantly we argue that it is hardly surprising that the comparison MP makes shows apparent benefits but when subjected to an appropriate type of sensitivity analysis MP's results are highly vulnerable to hidden bias, which is highly likely to have been present. We also provide PSM estimates using control villages and non-borrowing households in all treatment villages using first Chemin's method with village fixed effects, and then one substituting village fixed effects with village covariates which allows us to pool all (treatment and control) villages. These also turn out to show significant positive effects of female borrowing which SA shows are highly vulnerable to hidden bias. We repeat all these estimations for male borrowing households, finding similar, if smaller, positive impacts highly vulnerable to hidden bias. Thus, claims about the robustness of the headline findings of PnK finds little if any support in this application of PSM to these data, contrary to the assertions on MP. Following this we address a crucial issue in both PnK and in MP's replication of DPJ – the use of the "non-target" dummy variable. We suggest that there is no good reason to include such a variable if its justification is that it differentiates not eligible (non-target) households from households which do borrow and have similar quantities of cultivable land. When either dropped, or substituted with a data driven variable the headline results using PSM either disappear or are similarly vulnerable to hidden bias, and therefore provide evidence of limited credibility.

"High Noon"

Our use of this title was ironic, polysemic and, perhaps, mischievous. Pitt adopts only one interpretation – that DPJ purports to be definitive (which it does not) with regard to the impacts of microfinance (especially when lent to women…….). This interpretation is neither in our text – it would have been truly hubristic - nor in the meaning we would attach to the title. At best we would see our contribution as showing something about the specific data set which is in itself relevant only to a specific locale and time. Clearly this is not how Pitt sees the meaning of the episode; his valuation of his work can be seen in his commentary on RnM (Pitt, 2011a).

There are other possible interpretations based on the text of the film. Our use of this title could equally be about an ageing "marshal" undertaking one last attempt to dispose of something that has been threatening to pursue him into retirement. As we made clear in our paper we are concerned with what we consider the disproportionate influence of complex analytical methods in development policy discourse rather than a decisive judgement on the merits of MF. We do not consider High Noon (hereafter this refers –to DPJ) – or indeed our systematic review (SR) - definitive with regard to the impact of MF. Perhaps, hopefully, our critique of Chemin (and less directly of PnK) contributes to a clearer assessment of the ability of the PnK data set to provide a clear judgment about the impacts of MFIs in that data set analysed by PSM. Our SR, which was informed by our and other critiques of PnK, does, we believe reflect helpfully the state of evaluations of MF (the findings of our SR are supported by other SRs - see Stewart et al, 2010 & 2012 – more on this further below). That it would seem that Pitt regards our aim as definitive reflects perhaps his opinion of the merits

of PnK, an opinion widely shared, and of great influence in justifying belief in the merits of the models of MF which underlay the massive expansion of MF in the last fifteen years or so. However, there has been much criticism of this phenomenon (Bateman, 2010; Roy, 2010; Karim, 2011; Sinclair, 2012), and also of the particular MFIs in Bangladesh around the same time and subsequently (Fernando, 1997; Karim, 2011). Our own view is rather more agnostic, finding that the evidence "can neither support nor deny the notion that microfinance is pro-poor and pro-women" (Duvendack et al., 2011a:74).

The messages of the "Gunfight at the OK Corral" are rather more ambiguous; while the popular mythology, from the Hollywood film community and the like, is of conflict between just lawmen (the Earps and Doc Holliday) and bad outlaws (the Clantons and McLaurys**)**, the reality was perhaps not so clear. It seems that not everyone approved of the Earps, who were seen as tending "to protect the interests of the town's business owners and residents" (Wikipedia, Gunfight …, 22/9/2012). There was a feud, a fight which may or may not have been "fair", an attempted and a successful assassination, presumably by the outlaws, and a murderous extra-legal vendetta by Wyatt Earp perhaps financed by the interests that the Earps sought to favour. High Noon is entirely fiction, but is often taken to be an allegory for the House Committee on Un-American Activities (HUAC), to which the (Hollywood) townsfolk were unwilling to stand up. Readers can imagine their own castings and interpretations.

Systematic review

Pitt elaborates further on "ethical" aspects of our work and behaviour referring to our Systematic Review (Duvendack et al., 2011a; see also Duvendack and Palmer- Jones (2012c – DPJc). DPJc point out both a non-sequitur (errors we made in the analysis in DPJ do not mean that our writing in Duvendack et al., 2011a was equally compromised), and a singular interpretation of the appropriate ethics. MP writes that DPJ is the "keystone" paper which backs up our "claim that other studies, particularly those by authors Pitt, Pitt's students, or Khandker, are deficient" (MPa:1 fn 1), and that the reader should judge the quality of "the DFID report and its conclusions" by the "quality of the keystone" paper (ibid). Of course the SR does no such thing. The SR draws its conclusions from a much wider range of studies than the Pitt and associates oeuvre, although it follows the literature in attributing some of this oeuvre prominence in providing evidence in support of the beneficent effects on MF. Rather, we, again following the literature, classify quasi-experimental research design analysed with complex econometric methods as being moderately to highly vulnerable to bias. We will adduce further evidence in the present paper that the PnK data set does not provide robust evidence in support of the key conclusions of PnK. Also, our claims in both DPJ and the SR with respect to PnK are far more modest; we report that the evidence is neither for nor against the headline conclusions of PnK.

Pitt's view of the importance of his and his co-authors work based on the PnK data set is also evident in footnote 1 (MPa:1)[7], where Pitt argues that his critique of our replication will undermine the conclusion in the SR that "impact evaluations of microfinance suffer from weak methodologies and

[7] For his views of the significance of criticisms of PnK see also the first few pages of Pitt's opening critique of RnM (Pitt, 2011a).

inadequate data" (4). But, at best, it would only undermine DPJ; it would not change the status of PnK as a quasi-experimental design. So far, Pitt only shows deficiencies in one paper, but we show here that our conclusions are not contradicted by his criticisms.

Even complete demolition of our PSM results might not undermine this specific conclusion of the SR, which is based on quite widely held views about the ability of observational, or quasi-experimental, data, even when analysed with complex methods, to provide strongly credible evidence of impact (c.f. the title of Manski, 2011, referring to "incredible certitude"). Nor does it dispose of the critique of RnM of the specific analysis in PnK[8]. Thus the aspersions Pitt casts on the SR is both a non-sequitur and misleading in misrepresenting the basis of the conclusions drawn in it. Our approach was to follow to some extent the evaluation of studies by a "hierarchy of methods" as in the medical literature (e.g. Concato et al., 2000), and that broadly speaking it is not possible to remediate by analysis data that do not have a strong research design (Light et al., 1990; Meyer and Fienberg, 1992; Rosenbaum, 2002).

Ethics

The general assertions we make, that replicability and replication can contribute to ethical practice, seem to be widely held, but not widely practiced. Pure replicability (that the estimation data set and the deposited analytical code do indeed generate the published tables and figures) of papers in the American Economic Review (AER) may have improved greatly in recent years (Glandon, 2010), but this is a very limited view of replication. Replication should ask rather more searching questions, including whether the estimation data set is constructed appropriately from the raw data, whether the appropriate models and estimation methods are used in appropriate ways, and whether similar analyses would produce equivalent results with different (but appropriate) data sets from different (relevant) contexts. These rather more taxing statistical and scientific replications (Hamermesh, 2007) are also legitimate.

Pitt's main concern seems to be about the process of our replication, and specifically that we did not communicate with him, that we did not provide variable construction code when he wrote his reply, that we reported inaccurately what we actually did, and that we made mistakes in what we actually did. There are concerns for the reputations of original authors, and about incompetent work lacking transparency.

We can easily dispose of the last concern; the estimation data set and code were made available promptly and fully replicate the published results, as far as we are aware. They enabled Pitt to identify our egregious errors. Pitt had some concerns about variable construction, but, while this is important, and we undertook to make our variable construction code available when the paper was published, this is not mandated by current codes (e.g. the AER data policy), and was not provided to Pitt for good reason. We clarified adequately his specific requests about variables, although he

[8] Late in the writing of this rejoinder we became aware of the response of Pitt and Khandker (2012). Their response does not dispose of the issue of the effect of outliers. We make further criticisms of PnK's analysis using their methodology in this paper.

seems not to have understood our emails[9]. The good reasons are (1) that the code was long and had grown "like Topsey" in an attempt to understand the underlying data and variable constructions of Chemin (and of PnK), and contained much unnecessary and messy code; (2) it constructs many more variables than used in High Noon and therefore was still our intellectual property.

Pitt has never disclosed his variable construction code, provided a misleading and barely usable data set to Roodman, apparently sat on known errors in the data set available at the World Bank (Chemin, 2012a), allowed the World Bank data set to remain poorly documented and incomplete, and has latterly released a data set that contains only the minimal set of variables required to purely replicate PnK rather than one that would allow a broader replication[10]. Although he continued to publish using these data up to at least 2006, Pitt explained in 2008 that no more data (or any code) could be provided to Roodman because the backup CD created around 1997 was found to be unreadable (see personal communications between Roodman and Pitt, reported by Roodman, 2011a).

In regard to communications with original authors we can see no ethical obligation to do this when data and results are in the public domain, although, as noted, it might be a good idea; however, there is a downside to communication with original authors, in that it gives them considerable power to delay or prevent publication by raising not only appropriate concerns, but also further issues which can tie up the replicators almost endlessly. Damage to reputation by mischievous replication (for example deliberate fabricating or misreporting results), is of course unethical, but erroneous replication can be readily countered by reply in the same journal and, nowadays, through electronic media, and poses little or no risk to reputation. One might be concerned about reproduction of unsatisfactory replications or misquotations (and quotations taken out of context) in the media, but this would hold academics responsible for all the unintended consequences of their work, and is not a generally held ethical principal, as far as we know.

Requiring extensive communication with original authors can stifle free speech and legitimate debate. Of course, one should do one's best to not provide support for inappropriate promulgation of doubtful arguments both in academic arena, and in the media, by suitable qualification (and transparency, replicability and so on), and we think our work in both High Noon, and in subsequent communications about it, and the SR meet this standard.

Who is really at fault in this matter? Publishing papers with mistaken (rather than falsified) results is a fault but not a crime. One of our (anonymous) reviewers was adamant that it is the original authors who failed to provide data and code and to adequately explain their methods, who are "always" at fault. Replication is a valuable and legitimate activity. Replicators of computational papers which do not provide full documentation (including data and code) are always playing catch-up; when they encounter reluctant communication with original authors there is no mandated procedure for fair communication or publication.

---

[9] He seems to have had somewhat similar problems with Roodman's data construction (see Pitt and Khandker, 2012) which is also inexplicable.

[10] We discuss below the apparently undocumented provenance of the crucial "nontrgth" variable, which, when we eventually understood it, we found lacking in credibility.

We would argue that it has been clear for a long time (at least since Morduch, 1998, if not Fernando, 1997) that there are legitimate concerns about the conclusions of PnK. It is arguable that in the light of these concerns it would have been appropriate for the original authors (Pitt and Khandker, and, authors and co-authors of the other papers in this oeuvre) to make data construction and estimation code[11] available to allow other researchers to gain a better understanding of the strengths, or otherwise, of PnK (and the other works).

Replication

As DPJb explain, and is clear from DPJ, we initially set out to replicate Chemin as a way of extending his analysis to see if the gender of borrower was confounding his results. His results are not strongly supportive of PnK in that Chemin finds mixed impacts, including a negative impact when comparing treated individuals with matched individuals in the same villages, points not mentioned by MP.

Chemin looks at all borrowers combined, so it is possible that, following PnK's results, the absence of an effect of male borrowing is masking the positive effect of females borrowing. In DPJ, we did not set out to replicate PnK directly; our replication would use a different method of estimation to PnK with the "same" data, and thus constitute a "statistical replication" in the categorisation suggested by Hamermesh, 2007. Thus, DPJ is primarily an attempted replication of Chemin, not PnK. As DPJ reported, we could not replicate Chemin, and, as DPJb, report we now understand many of the reasons why. The most significant reason for our failure was perhaps the unusual use by Chemin of land operated to define eligibility for MF rather than land owned. This, together with incomplete documentation of variables in Chemin, and our inability to infer what had been done largely because we did not think to construct our sample using land cultivated, meant we abandoned pure replication for an approach shaped by our interpretation of his work, but aiming mainly at obtaining the impacts of gender specific borrowing. MC also sees us as communicating inadequately. This and other issues are dealt with in DPJb & c; some issues are also discussed here.

Data construction

Pitt chides us in several places for not providing our data construction code (e.g. MPa:3 fn 3). However, although we provided both estimation data and estimation code as currently mandated by the AER data availability policy, Pitt has not provided his data construction code, has not provided estimation code for PnK, and has in the past provided an incomplete and unhelpfully labelled estimation data set which certainly made difficulties for replicators (Roodman, 2011a). As it happens MD and RPJ disagreed about whether to provide the data construction code, but we complied with the AER mandated code of practice, which recommends rather than mandates posting variable construction procedures. We also responded promptly to queries from MP; we don't think it is our fault that he did not understand our emails.

Furthermore, we reported in DPJ, and pointed out in emails to Pitt, that almost all our variables were corroborated with Roodman's construction, and we provided a table of the main

[11] Pitt and Khandker (2012) make considerable play of the fact that the data were available to Morduch (1998) even before publication of PnK, and has been available more widely for a considerable time. It seems likely, based on our own reasons, that the failure of others to attempt replication of PnK have much to do with the character of Pitt's (1999) reply to Morduch (1998).

disagreements. Pitt could have availed himself of Roodman's SQL database, which provides explicit code on data manipulations and construction, but Pitt seems to have been unable to do so[12]. We certainly make good use of Roodman's SQL database, communicated quite extensively with Roodman, and for the most part understand how Roodman constructed his variables from the World Bank data set, and other files which he obtained from Pitt and others, and has made available.

Contrary to MP, we did clarify the construction of variables he inquired about (wages) in two emails from 6 & 8 July 2012 (see Appendix 4), but it seems we could not adequately explain what had happened, or what we intended. This particular variable does not occur in the PnK data set; Chemin does not define the two wage variables he reports, but MC and accompanying code (not the incomplete code he originally provided to us) makes clear what he intended. What he intended was wage rates (per day); we could have inferred this, but we could not replicate his descriptives because we could not conceive that he had used operated land to define his sample, and this made it difficult to be confident why we could not replicate his variables. We then also made errors in our code in part because of extensive experimentation[13] with the code trying to replicate Chemin's descriptives[14]. Clearly, there are dangers if replication follows the original study too closely (or inattentively!).

As noted above, there were broadly two reasons for not providing variable construction code to Pitt, and we do not rehearse them further here.

Process

Pitt chastises RPJ for the process of replication employed by DPJ, in contrast to the values he attributes to RPJ in his quotations from the "As well as the subject" www site[15]. DPJb explain that we did not contact Pitt because of the experiences of others "replicating" PnK. Pitt also confuses good practice for replication by independent researchers with good practice for replications funded by 3ie. There seems no reason why independent replicators should contact original authors, although they might be well advised to do so when helpful responses might be expected. For replicatees who are still active, the market place in ideas is available to them to provide clarification, robust riposte, and so on, as the history of replication makes clear. There is no reason in ethics why DPJ should have contacted Pitt over "High Noon". There are disadvantages to pre-publication exchanges between

[12] Pitt (see also Pitt and Khandker, 2012) apparently could not use Roodman's data base. One needs to install MS SQL (which is free) read the SQL commands and understand Roodman's variable constructions In any case, Roodman makes estimation data sets available.

[13] Roodman points out that: "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." (RnM:6).

[14] To provide an example of the problems for re-constructing some of the descriptives: Wage rates in an economy such as rural Bangladesh are not easy to define; they vary by task, season, person, and employer (at least). One could argue that the appropriate wage rate to use is total wage earnings divided by total numbers of days worked, computed as the sums over the three rounds of the survey. Chemin's code makes clear that he used the specific wage rates reported in the first round in which the person appeared in the survey; this was not documented in Chemin or in any code he made available to us until late 2012.

[15] http://www.uea.ac.uk/international-development/dev-co/research-networks/dimensions-in-research-ethics

replicators and replicatees, as noted already, as this can lead to inordinate delays in publication as the latter engage in activities which can be characterized as "muddying the waters" (McCullough, personal communication; see also Hamermesh, 2007; McCullough and McKitrick, 2009; Duvendack and Palmer-Jones, forthcoming – DPJ, forthcoming). As we argue elsewhere, replication raises concerns among replicatees about the possible effects of the replication on their reputations, especially if the replication is botched, or somehow unfair. But, as we have pointed out, there is the marketplace in ideas to provide adequate restitution, and replicatees may have more to fear from any replicatory failings than original authors.

On the other hand, especially in view of the fairly common animosity generated by replications (DPJ, forthcoming), an organisation engaging in funding replications might want to require controlled communication between replicators and original authors in order to avoid, if possible, rancorous, ill-tempered, and perhaps in the end unhelpful exchanges that have occurred in some cases (see Hoxby-Rothstein, Acemoglu-Albouy, various years[16]). Nevertheless, a fairly robust attitude towards original authors even though replication could affect their reputation, seems warranted, to avoid allowing replicatees undue control over the replication and publication processes.

It is not our responsibility to justify the editorial processes of JDS as to whether to offer original authors either opportunity to peer-review or to reply. In any case, there is no necessary reason why Pitt should have become aware of High Noon, which was cast as a replication of Chemin, especially, perhaps, as Pitt had seemingly shown no interest in Chemin. To do so might imply that Pitt should be consulted over every paper addressing PnK (or is it just those which contest PnK's main findings?), or even Chemin for that matter[17]. As it was primarily concerned with Chemin it would have been perhaps sensible to offer him (Chemin) the opportunity to reply or ask him to be one referee. Whether this should have been extended to Pitt is moot; High Noon is less directly concerned with PnK, of which it is a "statistical" rather than pure replication, using a different method which Pitt had not previously used, as far as we know.

Propensity Score Matching

Pitt hints that he does not have a high regard for PSM, and notes that all parties to this discussion share scepticism about its merits with these data. However, PSM is being widely, and probably somewhat indiscriminatingly used (Shadish, 2012). We (DPJ) have been strongly critical of applications of PSM, especially when not used in conjunction SA. We used SA to report that our findings, whatever their merits, were unlikely to be robust to confounding unobservables, which we have good reason to believe are likely to have been present. This is especially likely in the comparison Pitt (MPa) uses, namely that between women borrowers and women in the same village in households which are eligible for MF but are not borrowers. One only has to ask why these (control) women did not become members of the available MFI, and, perhaps, to have some ethnographic insights, as well as knowledge of the microfinance literature on selection by peers and MFI agents, to realize that there are likely to be unobserved and unobservable characteristics of these women likely to confound any estimated impacts. Reporting PSM without SA in such a context

[16] In both these cases the original authors are in our view the belligerent parties.

[17] Which as we show below shares several of the problems Pitt takes us to task for.

is just the type of selective reporting that those concerned with ethical publication practices worry about (see http://publicationethics.org/category/keywords/selective-reporting[18]).

MP's Replication

MP conducts a replication of our replication, and concludes that "[T]he results… indicate positive and highly significant results of women's participation in microfinance… . The results are also very robust ….[19]..." (MPa:15). Elsewhere Pitt and Khandker, write that "the Duvendack and Palmer-Jones results provide no credible evidence on the validity of PK or Chemin or on the effectiveness of microfinance" (Pitt and Khandker, 2012:29-30). This may be the case, but our reply in JDS (DPJ, 2012c) and also the replication of MP's replication conducted here do in our view vindicate our conclusions. We find a number of points in MP's replication that are wrong or misleading, we replicate his replication, conduct SA, and conclude that his conclusion is far from robust; furthermore, when corrected, our conclusion that "policymakers would have been well advised to have placed less reliance on PnK" (DPJ: 13) is amply supported. These conclusions do not contradict the concerns that have been expressed about the headline findings of PnK, raise further questions[20] as to the legitimacy of the belligerent and hostile nature of the texts produced by MP when controverting the attempted replications of PnK (or Chemin).

We now replicate MP's replication using code which he provided to us on request, before presenting the SA results. The interested reader might also want to read Appendix 1 which addresses specific points MP raises about DPJ, these points and our responses inform the replication of MP's replication.

Replicating the replication of the replication

In this section we explore whether Pitt provides adequate evidence in support of his claim that "[T]he results… indicate positive and highly significant results of women's participation in microfinance… . The results are also very robust ……." (MPa:15). Although MP corrects some of our errors, and suggests some improvements, we find a number of points that are wrong or misleading in Pitt's replication (and indeed in PnK). We replicate his replication of DPJ and suggest that MP's conclusion derived from his PSM is far from robust.

We also provide further evidence of problems with this analysis of the PnK data related to the use of the "non-target variable" and justification for not including a "mis-target" variable. More than 20% of participating households have cultivated land areas greater than the putative cut-off of 50 decimals of cultivable land. The use of the "non-target" variable in the PSM analysis resulted in excluding "non-target" households from matching with participating households whether "mis-targeted" or not. We provide evidence that there is little to distinguish the non-target households from mis-targeted households at least as far as the unit values of land are concerned, which Pitt (1999) uses to justify classifying households which participate in microfinance owning more than 50

[18] See especially: http://publicationethics.org/case/publication-misleading-information-and-publication

[19] "to changes in bandwidth, trimming, and the number of strata, as well as dropping the last two rounds" (MPa:15).

[20] Further to those that arise because PnK's data and code archiving has not been of the standard mandated for the former and suggested for the latter by professional best practice (e.g. the AER data policy).

decimals of cultivable land as not mis-targeted. We then drop the non-target dummy variable from the PSM analysis, and demonstrate the effects of using alternative data driven classifications of non-target using PSM[21]. We also correct for other issues such as female borrowing within the survey period, mis-classifying at least one village as not provided with female borrowing choice when several households report female borrowing even in Round 1, and dropping another village that seems to have been inadvertently dropped by MP because a dummy variable representing presence of female borrowing from a MFI is missing. PSM including "non-target" households as potential matches, and these other changes, provides no better evidence for the PnK headline findings.[22]

There are a number of specific points to clarify. Firstly, Pitt uses his "household" level data set (PKexp.dta) rather than an individual level data set as used by Chemin and DPJ. It is possible to use individual level data, as we did following Chemin, correctly, but for the purposes of this reply we adopt the household level analysis. The data set that MP has provided has a number of problems and characteristics which need to be addressed. It is inadequately documented, and has errors such as the land owned of household 32111 in round 3 (see also the issues in the data sets reported by MC2:6, and RnM, 2009:14). It includes the crucial "non-target" dummy variable the provenance of which is undocumented. The specific code used by MP also inappropriately excludes some villages with female borrowing, and results in including households with male borrowing in both treatment and control groups introducing possibilities of bias.

Secondly, Pitt wants to restrict the comparison group to households which have no female MF borrowers and live in villages which have access to female MF groups. It is not surprising that this control group is worse off than MF borrowers, and making this comparison raises issues as to the meaning that can be attached to an estimation using a control group that is likely to have been strongly selected not to participate in MF. In particular what are the ethics of making a claim such as that the results "indicate positive and highly significant results … [which] are also very robust" (MPa:14-15) when using such an obviously flawed comparison[23]?

---

[21] We have also re-analysed the PnK data using Roodman's cmp module with the model specification provided by Pitt (2011a). However, the vulnerability of this approach to outliers identified by Roodman makes us wary or reporting any results using it.

[22] The WESML-LIML-FE method as currently implemented also provides no meaningful evidence about impacts of microfinance once non-target is redefined on an area or value of land basis, and the instability of the results to alternative definition of the target and non-target households is further testimony to the fragility of such methods, and the "incredible" belief required for them to serve a useful purpose in evidence based policy analysis.

[23] In addition to the likely selection of participants on unobservables among "target" households, we show below that it may be inappropriate to exclude the "non-target" households based on land unit values and owned cultivable land. Whether "non-target" households as defined by Pitt had non-land assets which would have made them non-eligible cannot be determined from the data set with much confidence as non-land assets are reported for the time of the survey rather than the period prior to borrowing. However, even by this criterion (value of assets greater than the value of 1 acre of medium quality land), some appear to be target at the time of the survey, and anyway have less than 0.5 acres, making them eligible for MF by the criteria of the time.

Furthermore, thirdly, why discard for control purposes eligible (however defined) households without male borrowers in villages with access to male only borrowing, or indeed no access to MF borrowing (female or male), e.g. those in the control villages. There is an apparently straight forward answer to this latter point in that including them prevents use of village-level fixed effects in the propensity score estimation. However, there are plausible ways around this either following Chemin's procedure of estimating the "corrected" propensity score model[24] , or using village level covariates to control for village characteristics instead of village-level fixed effects. In either case one would not be discarding potentially relevant information. One can wonder about the ethics of not conducting and (or not) reporting the results of a seemingly legitimate and easy to conduct analysis (and not reporting that Chemin had done this and commenting on his results), that may better exploit the research design than the particular results MP chooses to report?

Fourthly, as is well known, PSM can only control for bias associated with observed covariates, but not for "hidden bias". However, the possible vulnerability of the estimated impact may be estimated effects using SA. MP makes no use of the appropriate type of SA. Again, this raises ethical issues as to the selective analysis and reporting relative to what can plausibly be considered good practice (at least as far as PSM is concerned).

These points suggest there are a number of grounds for subjecting MP's replication of our replication of Chemin to replication, and to include the insights such as those just reported. Given the number of plausible objections to MP's criticisms and his own PSM, and the much more qualified conclusions it seems reasonable to reach, MP's belligerent tone seems not only unwarranted but also unhelpful in promoting collegiate debates from which perhaps all, including MP, could learn.

Propensity Score Analysis of the Impact of Microfinance

Now we report the results of re-estimating MP's propensity score analysis making corrections for the inappropriate inclusion of male borrowing households and the exclusion of some villages with female borrowing, and applying SA to the PSM analysis.

MP compares households with female borrowing with non-target (MP's definition) in villages with access to female MF borrowing and we replicate this but drop households with male borrowing and villages with no access to female borrowing. Later we report comparisons with control villages using Chemin's method, and then one based on propensity scores estimated using village level covariates. When doing this we first restrict our sample to (all) villages with female borrowing only and exclude households with male borrowing; we compare estimates using village fixed effects with those based on village covariates, and we compare the balancing properties of these estimations. The estimation using village covariates are apparently no worse than those using village fixed effects. Then we estimated the impact of female MF using all the villages. In each case we exclude households with male borrowing, and we conduct sensitivity analysis. Only limited results are reported, but we contend that they are sufficient to make the points we wish to emphasise.

[24] This estimates a propensity score model using only treatment villages, but predicts propensity scores for control villages using these coefficients after removing village fixed effects from the propensity score, and using "corrected" outcome variables from which village fixed effects have been removed(Chemin, 2008: 473).

Before proceeding to the PSM estimations we make some comments on SA of PSM results, as it is different to the sensitivity analysis (i.e. robustness checks) reported by MP which do not address the issue of vulnerability to unobservables.

Sensitivity Analysis

SA (which we DPJ undertook and reported) was developed by Rosenbaum (2002) to assess the vulnerability of estimations to hidden bias. This vulnerability could be assessed as the degree of association of an unobserved confounding variable with treatment and outcome that would render the observed impact statistically insignificant. It is generally recognized, as we reported (DPJ), that "sensitivity analysis should always accompany the presentation of matching estimates" (Ichino et al., 2006:19; see also Rosenbaum, 2002). Using data from the association of smoking with death from lung cancer Rosenbaum (2002) suggests we imagine that there is an unobserved variable that completely confounds the association of treatment with outcome, and suggests a number, gamma, that is the odds of having this characteristic that the treated have relative to the untreated having it. A gamma of 1 would mean there need be no difference; a value of 1.5 would imply a 50% greater odds, and so on. The medical literature suggests that a value of gamma of 5 would suggest it is highly unlikely that a characteristic with this difference in odds between treated and untreated would not have been observed (i.e. could be readily observed), while variable with a gamma of 1.5 or even 2 might have been readily overlooked. Some, economists, have suggested that this is a worst case interpretation, that there is a single variable with these characteristics. But, it can be more readily interpreted as a combination of variables (for example "energy and ability"). Nevertheless, the gamma required to explain the associations observed as early as the late 1950s between smoking and lung cancer were of the order of 5, which did not of course prevent the tobacco industry from continuing to deny this association suggesting instead that there was an unknown hormone that accounted for both smoking and lung cancer for several decades, and consequently contributing to many, many excess deaths.

MP does not report (this type of) SA, instead performing matching with different kernel bandwidths, and trimming of outliers, and stratification matching; these methods throw little light on the question of vulnerability of the estimated impacts to unobserved confounding.

Replication using PKexp

We now turn to the replication of MP's replication using his PKexp data set. Our replication makes the following modifications:

1. Don't drop villages 11 & 143: village 11 is clearly a village in which there has been access to female MF credit, although the situation in village 143 is not clear to us[25];
2. Correct the land owned variable for nh = 32111 (from 10123 to 123 in round 3);
3. Remove households borrowing for the first time in rounds 2 & 3 from the definition of participant[26].

[25] 5 households report female borrowing in village 143 in round 1. We treat the one household in village 242 which reported female borrowing in round 1 as an aberration even though several more households in this village (and some in others) reported female borrowing in round 2 & 3.

These estimates are based on treatment villages only and are inherently limited because it is highly likely that those eligible (or matched) households which have not become MFI clients are likely to have unobserved characteristics which are negatively associated with outcomes. We show that these estimates are highly vulnerable to unobservables (which are highly likely to have been present).

This approach also makes no use of the quasi-experimental design of the PnK study. Hence, we then proceed to test two alternative ways to utilize the information from the control villages (which, when we are concerned with female borrowing would include households in villages with male only borrowing which had not borrowed at the time of the survey). The first method is that suggested by Chemin, and the second substitutes village covariates for village fixed effects, which allows us to include all villages in the propensity score estimation and matching estimates.

First we repeat MP's kernel estimation (bw 0.08) and show that it includes households with male borrowing in both treatment and control villages (Table 1, row 1, cols 6 & 7). An equivalent table for households with male borrowers is given in Appendix 3, Table 11; is it interesting to note that MP's method shows that male borrowing also has a positive effect on household consumption, with similar effect size and level of significance to female borrowing – allowing for the smaller number of households with male borrowers. This is different to the lack of effect of male borrowing reported in PnK, and appears to be a case of selective analysis if not selective reporting.

A number of issues in addition to those raised above arise in MP's PSM:

(1) the propensity score estimation has "nospouse" entered twice;
(2) the "noadultf" variable (a dummy variable = 1 when no adult females are present) perfectly predicts non-membership and so causes 53 observations to be dropped,
(3) the "nontrgth" variable drops out because of collinearity (and also perfectly predicts non-membership) and causes 438 observations to be dropped.
(4) Households with female borrowing for the first time in rounds 2 & 3 also appear among the treatment and control groups.
(5) Households with male borrowers appear in both the treatment and control groups.

---

[26] The main problem is that these households are more likely to immediately consume some or all of the amounts they borrow. Excluding these households makes little difference to the results, so we do not report these results.

Table 1: ATT and sensitivity analysis for Female borrowing (treatment villages only) [27], [28]

Panel A								
kernel estimate, bw = 0.08							N households with any male (female first borrowing in rounds 2 & 3) borrowers	
		1	2	3	4	5	6	7
Sample	Method	n	diff	sediff	p <=	't-value'	Un-treated	treated
MP [a]	kernel, bw 0.08	3192	.0516	.0163	.0016	3.16	493 (20)	173 (34)
Excluded male credit households[b]	"	2592	.0537	.0183	.0034	2.930	0	0
Panel B: Sensitivity Analysis								
Nearest neighbour estimates							gamma for p > 0.10	
Excluded male credit households[b]	n(10)	2557	.0473	.0192	.0141	2.459	1.25[c]	
"	n(1)	2557	.0464	.0240	.0534	1.931	1.15[d]	

Notes: a. corrects area owned of hh 32111; b. also includes villages 11 & 143 (see text)
c. estimated using rsens; d. estimated using rbounds

Neither exclusion (due to nospouse or nontrgth) is reported by MP. Surprisingly, dropping noadultf (and 53 cases) causes meaningful changes to the results using kernel matching[29]. In the case of nontrgth the cases dropped are those which are not targets of MF. These cases are defined by the csg variable in the data rather than being defined by the eligibility cut-off of 0.5 acres; i.e. there are nontarget households with less than 0.5 acres of cultivable land – this is discussed further below. It seems most likely that this variable reflects the "choice based sample" in which three "non-target" households were selected following a villages census in which the population was stratified in those qualified and non-qualified "to join a program". Surprisingly neither PnK (1996) nor PnK (1998) discuss the criteria which lead to households being identified as non-target; some, as noted, surprisingly had less than 0.5 acres of cultivable land at the time of the survey, just as some "target" households had more than 0.5 acres of cultivable land prior to accessing MF. Those households which do borrow and have more than 0.5 acres (some with more than 5 acres which are excluded if they are "non-target") are included among the treatment cases. Later we will argue that there is no good reason to exclude the non-target households which are not distinguishable from mis-targeted MF participants in terms of the unit value of their land, which is the indicator which used by Pitt (1999) to justify their non-target status. We address the issues of mis-targeting and of alternative "non-target" criteria in a later section of this paper.

[27] Same specification as MP, i.e. including nontrgth & noadultf. Results dropping households which borrow in rounds 2 and 3 for the first time make no substantive difference (results available in code).

[28] An equivalent table for households with male borrowers is given in Appendix 3, Table 11.

[29] Not reported. MP has already dropped 41 households which don't borrow and have more than 5 acres, although several borrowing households report owning more than 5 acres (in total).

Results of sensitivity analysis (Rosenbaum bounds) applied to female borrowing only villages

There is as yet no theory of SA for kernel matching estimations. However, theory and methods for nearest neighbour matching are available, and we applied these to MP's estimations. We were not surprised to find that for the PSM estimate reported in Table 1 panel B (in this paper), the gamma at which vulnerability to an unobserved confounder of the association of female borrowing with the log of household per capita weekly expenditure in the various estimations, including those not reported, was around 1.2 to 1.4 - i.e. the headline result reported by PnK (and MP) of significant benefits from female MF borrowing, is rather vulnerable to one or a combination of unobserved confounding variables which are quite likely to have been present but be unobserved – for example entrepreneurial energies and abilities.

The robust conclusion confirming the headline findings of PnK in MP's text invites us to think that there is no significant likelihood of an unobserved variable which could account for the observed relationship between female borrowing and consumption in these data, thereby perhaps exaggerating the evidence for MF that these data analysed in these ways provide, and diverting attention from the pursuit of more robust evidence, and the consideration of alternative policies to MF to mitigate poverty. This - failing to present, and interpret, the results of the SA – is, perhaps, the antithesis of the denial of the tobacco industry. Instead of strongly asserting the existence of a highly unlikely confounding variable which could account for the observed association of smoking with death from lung cancer, contributing in no mean way to countless avoidable deaths of smokers and passive smokers in the intervening years, MP denies the potential of a confounding variable that is quite likely to exist, and consequently may have diverted attention and resources from alternative more beneficial interventions (or the search for more robust evidence).

PSM using control villages

As noted above, it is hardly surprising that households with female MF clients have higher per capita expenditures compared to households in the same villages which are matched to them, and that the estimate is vulnerable to unobservables. Also, this estimation does not take advantage of a key putative advantage of the research design, namely the presence of a control group of villages in which there was no access. However, these cannot readily be included in the PSM sample with village fixed effects since these are perfectly collinear with no credit in control villages. Chemin's solution is to "predict" propensity scores for control villages using the estimates for treatment villages but removing the village level fixed effects from the predictions (i.e. estimate the propensity score with village fixed effects for treatment villages only, set village dummies to zero, and the predict the propensity score for all villages). Another approach is to use village covariates instead of fixed effects. We discuss both approaches below. First we compare the village fixed effects model with the village covariates model for the treatment villages only; then we estimate the village covariates model for the treatment and control villages combined.

Village fixed effects

Chemin suggests that the problem of comparing treated households with households in control villages, avoiding the limitations of village fixed effects, is to use the treatment villages to estimate the propensity score model but predict propensity scores without using the village level dummies, and to remove village fixed effects from the dependent variable. While it is not clear that this is what Chemin actually does (DPJb), we implement this here. We continue to use MP's data set with the

limitations noted above removed (i.e. reinstating villages 11 and 143, and removing households with male borrowers from the sample). The results (Table 2 rows 1 – 3) show negative impacts of female MF when using only control villages to compare with treated households; they are different to those reported in DPJb Table 3 because the latter uses a different propensity score estimation, and different sample. When using all households without male borrowing (whether in treatment or control villages) (Table 2 rows 4-6) the results are positive but of low significance. We do not attach any precise significance levels to these results because Chemin's method is not convincing, even when implemented with MP's data and propensity score estimation.

In a later section (Implications for replicating PnK) we report the effects of dropping the variables noadultf (=1 if no adult female present) and nontrgth (= 1 if non-target household[30]), with the latter case corresponding closer to our suggestion (see below) that it may be appropriate to compare "mis-targeted" households with non-target households, since it is not possible to distinguish between them on the grounds or either area or unit values of cultivated land (see section "Land and land values" below).

Table 2: Impacts of Microfinance treated vs matched households in control villages (Chemin's method)[a]

Control villages	method/logit model	method	n	diff	sediff	pdiff	tdiff	Critical gamma
Control[b,c]	Chemin/ MP	Kernel	2146	-0.0263	0.0183	0.1499	-1.4406	.
"		N(10)	2146	-0.0280	0.0201	0.1639	-1.3924	1.45[a]
"		N(1	2146	-0.0510	0.0255	0.0451	-2.0047	1.10
All[b,c,d]		Kernel	3465	0.0221	0.0134	0.0991	1.6498	
"		N(10)	3465	0.0128	0.0144	0.3761	0.8853	1.05
"		N(1)	3516	0.0276	0.0199	0.1651	1.3884	1.10

Notes: a. We use MP's logit propensity score specification.
b. Male borrowing and non-target households excluded.
c. Controls are target households in control villages.
d. Includes all non-borrowing target households in treatment and control villages.

Village level covariates

Another way to avoid the limitations of village fixed effects is to use village covariates to estimate the propensity scores. In this case there is no need to restrict the sample for propensity score estimation to treatment villages only. However, it raises the question of whether using village level covariates is sufficient to account for relevant factors affecting the likely impact of MF. We address this issue by comparing the covariate balance between treatment and matched households.

The results are reported in Table 3 rows 3-7; in all rows the estimated impact is positive, and not much different to the ones presented in Table 2, but as with the results reported in Table 1 they are highly vulnerable to hidden bias (see column 6).

[30] We eventually found out that this variable takes the value 1 if the household was classified as non-target in the villages census conducted prior to the sample selection and survey. We compare these households with mis-targeted households there.

Table 3: Impacts of Microfinance using Village Covariates[a]

Control Villages	model	PSM method	N	Difference estimate	se	p	t	Critical gamma
			1	2	3	4	5	6
Control[b,c]	VFE	kernel	2578	0.0358	0.0169	0.0340	2.1217	.
	VFE	N(10)	2578	0.0223	0.0178	0.2113	1.2507	1.10
"	Vcov	kernel	2600	0.0402	0.0167	0.0160	2.4129	.
	Vcov	N(10)	2600	0.0368	0.0177	0.0377	2.0802	1.20
All	Vcov	kernel	3487	0.0384	0.0152	0.0118	2.5211	.
"	Vcov	n(10)	3487	0.0341	0.0162	0.0350	2.1103	1.20
"	Vcov	n(1)	3487	0.0223	0.0205	0.2762	1.0894	1.10

Note: excludes non-target households and non-borrowing households in treatment villages.

a, b, c. see notes to Table 2

Whether we use Chemin's method of comparing the matched households in the control villages with the treated, using village level fixed effects in treatment villages to estimate the propensity scores and extrapolate them to control villages (Table 2), or we use village level covariates to estimate the propensity scores for all villages (Table 3), we find modest impacts (which are negative using only control villages) highly vulnerable to hidden bias.

Balancing tests

In principle we would like to discriminate between the village-level fixed effects and village covariate PSM models on the basis of the degree to which they balance covariates when used on the same treatment village villages only sample (Gelman and Hill, 2008). Instead of a longer discussion we report the simple results of the convenient balancing tests provided by the Stata pstest.ado file. Table 4 and Figure 1 report the results indicating a slight improvement in overall balancing using village level covariates. Balancing is particularly improved for the age of the household head and the log of household owned land, and is also improved for other variables. From these tests it appears that there is little reason to not use village covariates rather than village-level fixed effects to estimate propensity scores for the treatment villages; whether this extends to the treatment and control villages combined seems to be untestable.

Table 4: Balancing Statistics- before and after matching[a]										
Control groups	Match method	Cont-rols	Mean bias		Median bias		Chi2 prob		R2	
			before	after	before	after	before	after	before	after
Control[b, c]	Kernel	Vfe	5.242	2.995	3.821	1.918	0.000	0.169	0.016	0.006
"	N(10)	"	5.242	3.444	3.821	3.812	0.000	0.091	0.016	0.006
"	Kernel	Vcov	5.086	2.412	4.275	1.456	0.000	0.563	0.015	0.004
"	N(10)	"	5.086	2.134	4.275	1.699	0.000	0.854	0.015	0.003
All[b, c, d]	Kernel	"	5.960	1.757	3.667	1.584	0.000	0.925	0.023	0.002
"	N(10)	"	5.960	1.614	3.667	1.248	0.000	0.951	0.023	0.002
"	N(1)	"	5.960	3.138	3.667	2.595	0.000	0.102	0.023	0.006
Note: Pitt's logits; kernel matching, bw=0.08. a, b, c. see notes to Table 2										

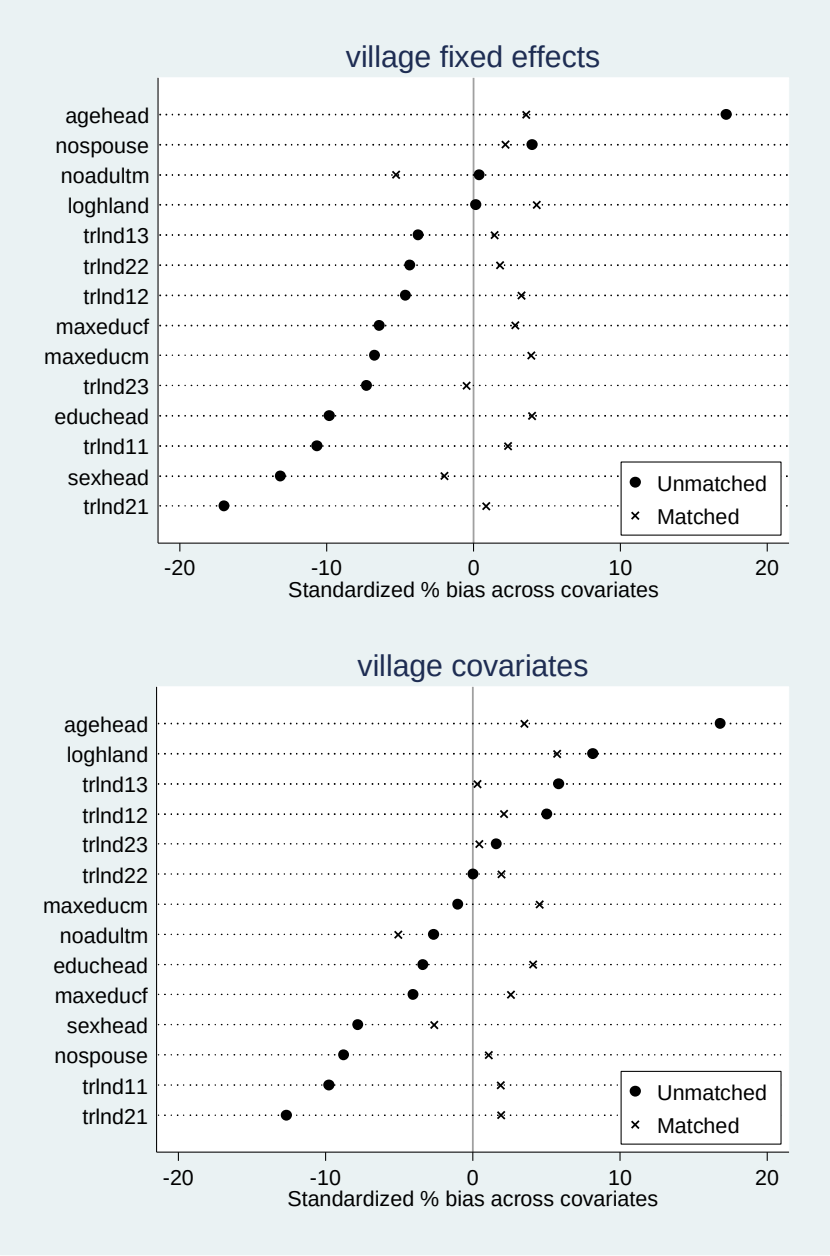

Figure 1: Balancing of RHS variables in the PSM models

Implications for replicating PnK

In this section we look at some implications of the discussion above for the replication of PnK; recall that originally our work aimed to extend Chemin, but in the process became something of statistical and scientific replication of Chemin. Indirectly this became also a statistical replication of PnK. Our replication of MP's replication of our replication was restricted to comparison within treatment villages. We extended this, continuing to include the valid points raised by MP, to the comparison with control villages following both Chemin's method and another using village covariates, which appeared to be at least as good or even better as the one using village fixed effects. All these results (reported in Tables 1-3) indicated positive impacts of female borrowing (except those using Chemin's method with control villages) which are highly vulnerable to hidden bias. Also, they all made use of the "nontrgth" variable which causes some 200 households to be dropped from the analysis (all villages), which does not appear appropriate.

As noted above, including the nontrgth variable causes a significant number of observations to be dropped because they are exactly identified as non-participants by this variable. Unless there are unobserved characteristics of these households that would confound the impact of MF, there does not seem any good reason to exclude them. The households identified by this characteristic are supposed to be "ineligible" for MF, but a basic analysis suggests that they may not have been correctly identified; in particular, it turns out that quite a number of them own less than the 0.5 acres of cultivable land which is commonly supposed to have been the cut-off below which households would have been eligible for MF at the time of the survey. Also, it turns out that quite a number of MF clients themselves had more than 0.5 acres of cultivable land, which should have made them ineligible for MF, and consequently may have been mis-targeted. Their inclusion may bias the results (we address Pitt 1999's objections to classifying them as mis-targeted below). Unless differentiated by some other characteristic (than area of cultivable land owned) it would appear to make sense to include the non-target households in the PSM estimations of impact. Is there any reason to believe that their inclusion in the propensity score estimation (along with the available covariates including cultivable land owned) introduces any more bias than occurs without them?

Hence, in this section we explore the rationale for identifying non-target households and the effects of including this variable (nontrgth) in the propensity score estimations. It turns out that the results (reported in tables Table 7, and Table 12 in Appendix 3) are very similar to those reported above; namely, that impacts of MF are modest, and highly vulnerable to unobserved confounding variables. A further reason for pursuing this analysis is that, if it is correct that at least some of the putatively non-target households are incorrectly classified, then there are implications for WESML-LIML-FE estimations (which we do not pursue here) given the crucial role[31] this variable plays in those estimations.

The Treated and the Controls

Which households and which values should be included in the analysis? MP's PSM drops a number of villages which have female borrowing, and includes households which borrow for the first time in

[31] and unexplained provenance.

rounds 2 or 3 of the survey, when it is unlikely that credit could have had an effect through the presumed causal pathway. In this section we discuss some of these issues[32].

From MP's comments on our sample we were lead to believe that he would compare households with female borrowers with households without borrowers in villages which had female MFI groups. However, as already noted, his code results in (a) dropping whole villages with female MF borrowers and (b) including households with male borrowers in both treatment and control groups. (a) occurs because, code at lines 19 and 20 of MP's kernel_commontrim0.do (for example) drops villages for which there is no variable in the set of dummies _Ww*[33] which should take a value of 1 for village 11[34]; 17 households village 11 in each round are reported as having female choice[35] (wchoice == 1), and this village had 9 households in round 1 with female credit[36]. All households in villages 102, 111, 113, 132, 143, and 242 are dropped, in the same way as those in village 11, even though they all reported as having female MF borrowers. None of these villages have any case with wchoice == 1, presumably because they first had female borrowers in rounds 2 or 3. However, villages 143 (5 households) and 242 (1 household) have female borrowers in round 1 (5 and 1 respectively). Perhaps the first time borrowers in rounds 2 & 3 should not be considered as "treated" in so far as impacts through the presumed causal pathway (borrowing -> investment -> net income -> consumption) are unlikely within less than one year, although they could have increased consumption in anticipation of, in order to, or necessitating, or as a result of use of the MF credit facilities.

Perhaps, also, one should exclude credit received in the six months prior to the first interview? In this context it is interesting to note that most of the MF loans and most of the loaned amount from MFIs occurred in the years immediately preceding the survey or during it. In order to demonstrate this point we need to use our own data construction as neither Pitt's nor Roodman's data sets include credit by date of receipt. Using our individual level file we report the distribution of loans

---

[32] 41 households which do not report borrowing from MFIs with more than 5 acres are dropped from the full sample of 1798 households and do not appear in PKexp.dta. 8 households in the estimation data set which report borrowing report owning more than 5 acres of cultivable land in round 1.

[33] We presume these dummies take the value 1 for "villages with female MF choice" and 0 elsewhere, although lack of documentation of Pitt's data set means we have to infer this.

[34] gen wvill1 = wvill > 0 // 0/1 dummy for MP's villages with female credit
tabstat wvill1 creditff creditmm if round == 1, by(vil) s(sum)

[35] i.e. they are eligible (either < 0.5 acres or are an existing MFI client) and have access to a female MFI credit programme. The _Ww* and _Mm* variables are sets of village level identifiers for households in villages which have access to female and male MFI credit. _Wwvill_11 and _Wwvill_143 are reinstated.

[36] write "tab vil creditff after executing code to line 19 (immediately after line 16 – "gen creditff = creditf > 0" Then write "tabstat wvill, by(vil)" after line 18 ("egen wvill=rowmean(_W*)"). Since the next line ("drop if wvill==0") drops villages with wvill == 0, village 11 is dropped even though it is a BRAC village with female (and male) borrowers
Try:
gen creditmm = creditm > 0 // male credit
tabstat wvill creditff creditmm, by(vil) s(max) // shows villages which MP's code indicates having female or male credit

and amount of loans received by source and year of receipt in Figure 2[37] & Figure 3. More than 55% of the total MF loan value reported in wave 1 was disbursed in 1990 & 1991; since loans are a significant proportion of annual household consumption expenditure (median loan size is more than 15% of median household annual expenditure[38]) it may not be surprising if expenditure is directly affected by some immediate consumption out of loans, or indeed by anticipation of receiving loans in the near future, or where loans are obtained to finance emergency or unusual expenditures in the preceding period. All of these circumstances speak to the possibility of reverse causality between loans and consumption, or direct consumption of loans, in at least some cases.

This emphasises the lack of an explicit causal model within the PnK estimation strategy. This is particularly unfortunate when consumption is the outcome variable as the well-known fungibility of credit allows credit received as cash to be consumed immediately rather than invested in a productive (in terms of income) enterprise (Hulme, 2000).

A further point to note is that in Pitt's data and code, MF credit acquired in rounds 2 & 3 is (or at least seems to be) included in the credit variables (amount of MF credit and the variables indicating whether any MF credit had been received) for rounds 2 & 3, even though villages in which female credit is received for the first time in rounds 2 & 3 are dropped from the estimation (because there is no corresponding wvill == 1 or Wwvill_nnn). None of this is reported by MP.

---

[37] Figure 2 includes all borrowing, including that in Rounds 2 &3. Figure 3 excludes round2 and 3 borrowing. A similar figure for number of borrowers excluding rounds 2 &3 does not change the points being made.

[38] median loan size 1990-1991 (unweighted) / (weighted median household per capita weekly expenditure * median household size * 52) -> 2950 / (74 * 5 * 52) = 0.15

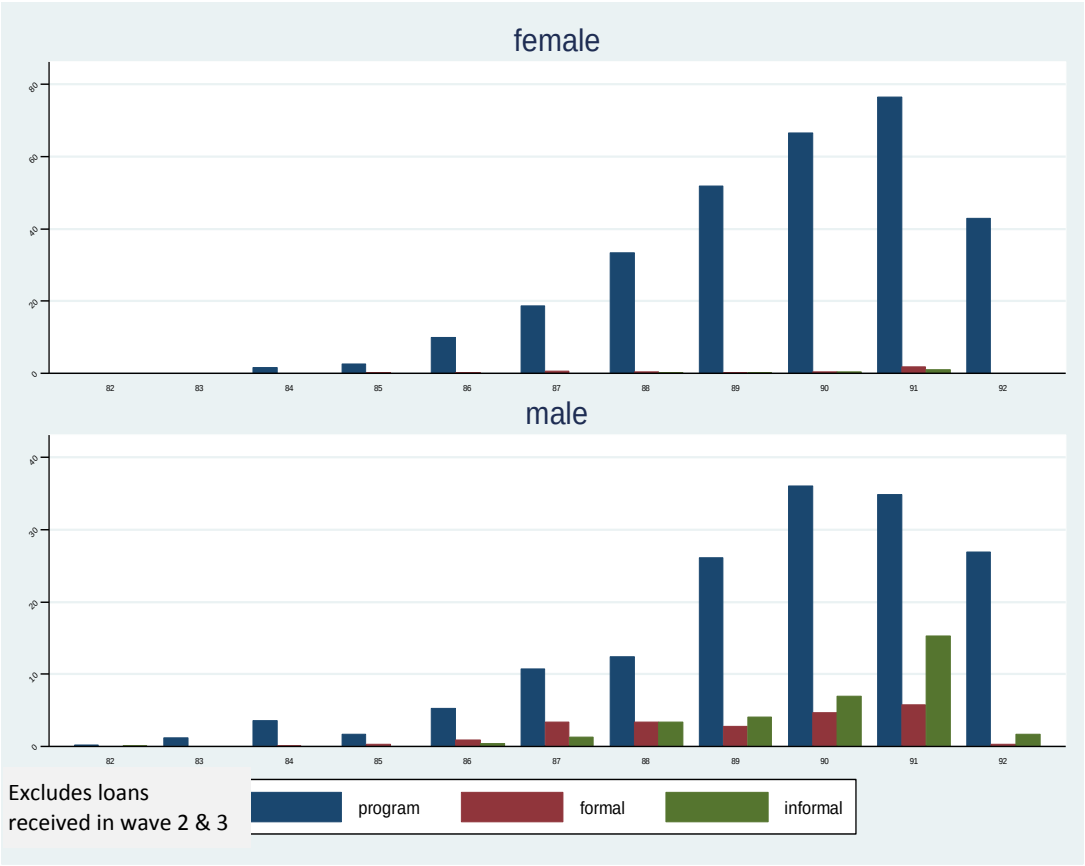

Figure 2: Number of MF loans by sources, date of receipt, and sex. (pkexp_analysis_rpj.do).

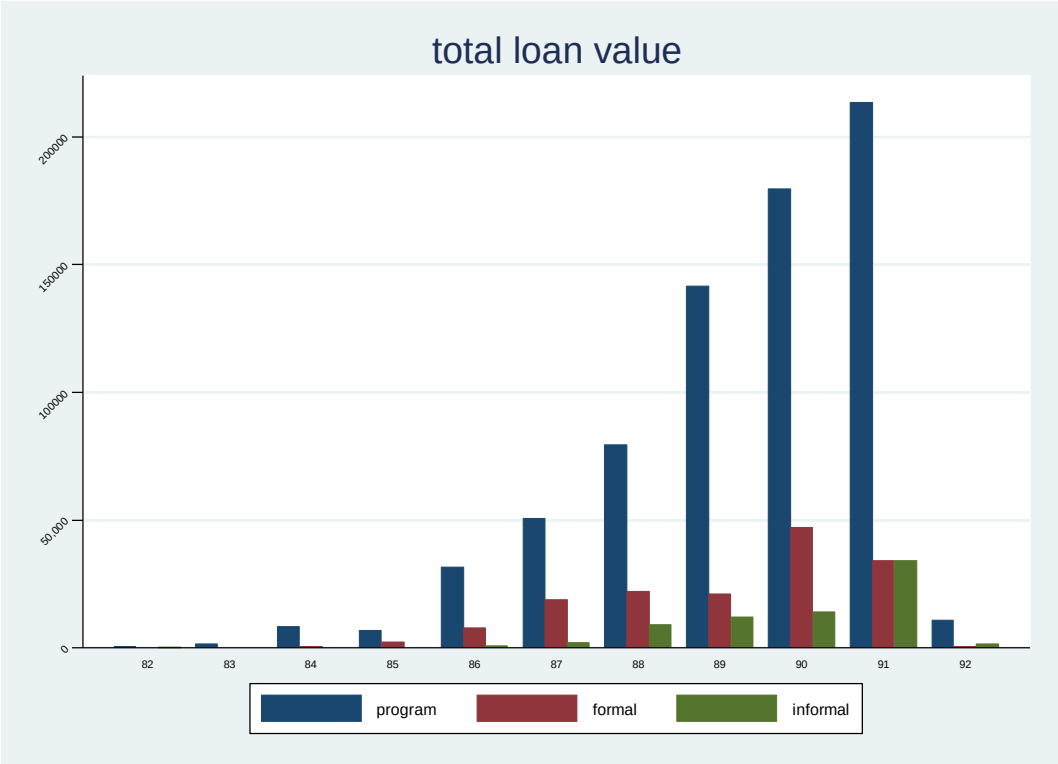

Figure 3: Total loan value by sources and year of receipt.

Mis-targeting

Furthermore, while households in treatment villages which have more than 0.5 acres of cultivable land are categorised as "target" if they are not "nontrgth == 1", all households in control villages with more than 0.5 acres are classified as "non-target". Thus in treatment villages there are 25 participant households and 33 non-target households with more than 0.5 acres, while in control villages there are no non-non-target households with more than 0.50 acres[39]. The asymmetric treatment of target and non-target households treatment and control villages with regard to eligibility has been the cause of some controversy (Morduch, 1998). Also, a number of households classified as non-target own less than 0.5 acres of cultivable land. While these households may have been classified as non-target because they have ample non-land assets, this is not reported. Thus, to our surprise, the provenance (and consistency) of the non-target dummy variable is never discussed. We now present and discuss analysis and implications of these potentially inconsistent classifications.

About 20% of MF borrowers have more than the 0.5 acres that is used by PnK to define eligibility of non-borrowers, and 2% (5/219) of the "non-target" households own less than 0.5 acres of cultivable land. Pitt (1999) argues that removing the former households makes no difference, and that it is legitimate to include them because the eligibility criteria actually used by MFIs was that "a person from a household that owns less than 0.5 acre of cultivated land, or assets with a value equivalent to less than 1.0 acre of medium-quality land, is eligible to receive a loan" (Hossain 1988:25)". The issue of potential mis-classification on the "non-target" households is not discussed.

There are consequently a number of problems with the data and analysis to address. We expected an attempt to justify these eligibility classifications in terms of the area or value of land and non-land assets. Land area may not be a good indicator of its value; for example, land of different soil texture, flood level and irrigation access categories have different values; irrigated (or irrigable) land, for example, has a higher value than un-irrigated/irrigable land. Hence, the land definition should take account of the different values of different categories of land, for example by putting either a lower cut-off in terms of area for irrigated land, or a higher one for un-irrigated land. Alternatively, since even within each irrigated/non-irrigated category land values can vary (for example by soil fertility,

---

[39] Distribution of households by cultivable area owned, Treatment, non-target and participation statuses

		< 0.5 acres		>= 0.50 acres	
Village	target	Non-participant	participant	Non-participant	participant
Control	Target	255		0	0
	Non-target	1		39	0
Treatment	Target	345	702	33	203
	Non-target	4		175	0

Cells shaded light grey are potential mis-classifications – i.e. non-target households with less than 0.5 acres, and target households with more than 0.5 acres. Particularly noteworthy are the 203 participant and 33 non-participant non-target households with more than 0.5 acres in treatment villages; note also, the no non-"non-target" households with more than 0.5 acres in control villages (dark grey). This table is replicated in "Pitt1999Table3.do", line 238.

vulnerability to flooding, and so on[40]), one could use a "value of (cultivable) land" criterion, derived, perhaps, from estimates of the value of different types of land, and land areas owned to estimate this value. We experimented with various eligibility cut-off criteria, and report some of the results below.

The key point is to treat equivalent households in an equivalent way. MP (and PnK) define a category of households as "non-target" and introduce a variable reflecting this (nontrgth in PKexp.dta. There are two crucial variables of interest – those households which are (not) eligible and do not borrow and those which are not eligible but do borrow (mis-targeted). We found it difficult to verify the construction of this crucial non-target variable which appears in PKexp.dta and which is undocumented (i.e. there is no account of how it was produced). Defining eligibility in terms of households owning less than 50 decimals of cultivable land (prior to MF for MF clients or at the time of survey) does not reproduce the eligibility variable used by PnK – nontrgth – but it does apparently reproduce the "mistargeted" variable used by Pitt (1999) - see below.

To attempt to construct (or verify) these variables we used data on land ownership (13ARr) and credit use (11ARr), or MFI membership (11ACr). Cross-checking was done with the classifications of households by the "progid" variable in the weights file (wgt123.dta) (nontrgth is the same as progid ==5; the origin of this (progid == 5) classification is also not described); when reproducing the results in Table 3 of Pitt (1999) to check our construction of the mis-target variable we also used the sample weights in this file. Since the land areas in 13ARr did not result in nontrgth (which we presumed was equivalent to non-borrower with more than 50 decimals of cultivable land), we also used cultivated land from 06ARr. This file reports land used for farming, and classifies it as near/far, irrigated and non-irrigated. These variables also do not produce the "nontrgth" variable.

Non-target

None of these trials gave rise to a variable which accurately reproduced nontrgth. It turned out, as David Roodman pointed out to us when, after many hours of experimentation, we asked him, that nontrgth is defined by the variable csg in 11CRr. This variables is defined as "HH Qualify" in the codes9192.pdf file (takes the value 1 if yes, 2 if no), and corresponds to a question which seems to be "Is anybody qualified to be group members from this household" in the questionnaire file (hhsurvey91Quest.pdf). Roodman suggested (personal communication) that this might be a "self-definition", but it is also possible that it was defined by the enumerator, or derived from the village census for stratification of the sample. How this variable was produced whether in the village census or subsequently) is not documented. Since it is not clear how the "nontrgth" (csg) variable was produced in the survey it is hard to speculate about its meaning. Since csg seems to have been specified in the sample design, it is likely that it was produced during the village census described by PnK (1996:24-5) and PnK (1998: 973)[41]. Both these locations imply that a land criterion was used. In any case, owned cultivable land does not correspond exactly with nontrgth, and we spent a lot of time attempting to understand it.

---

[40] A common classification of soil in Bangladesh focusses on soil "type" – which generally refers to elevation in relation to flooding, and soil texture, which refers roughly to water retention capacity and inherent fertility

[41] "land ownership is used as the primary eligibility criteria for these credit programs only to proxy for unreliable indicators of income, consumption or total asset wealth" (PnK, 1996:20).

We find this usage very surprising because, given the known ambiguity of eligibility criteria and suspicions of mis-targeting by MFIs even in the 1980s, we would expect self-report or (expert) enumerator report, or indeed the stratification derived from the villages census, to be liable to error, and, as our earlier effort suggests, we would expect this variable to be justified – ex-post - in terms of the (cultivated) land owned (and, or value of assets) reported in the data. To define mis-targeting in terms of cultivated land while using a survey stratification variable (csg) to define "non-target" without exploring the consistency of this asymmetric derivation with the data on ownership of cultivable land is surprising. We report experiments with data defined eligibility variables below.

To cut an already long story short, we found that PnK, Pitt (1999) and MP do not use a data consistent non-target variable; rather they used a variable which appears to have been either self, or enumerator reported, or defined in the village census (csg in 11CRr[42]), and which is not entirely consistent with the data on land ownership or total assets[43]. The variable for "mis-targeting" does seem to derive from reported land ownership. Using a data consistent indicator of non-eligibility (i.e. substituting nontrgth with a variable based on owned cultivated land, or assets[44]) does nothing to rescue the PSM results. To have used such a variable without exploring its consistency with the other data and testing the sensitivity of results to alternative plausible definitions of non-target households is hard to comprehend.

Land and land values

Defining eligibility and mis-targeting should, we have suggested, involve cut-offs in terms of area or value of cultivable land owned, or total assets (subject to the caveats mentioned above). A simple cut off in terms of area alone without taking account of its unit value (irrigation status, soil type and texture, and so on) is unlikely to mimic what MFIs were attempting to do in Bangladesh at the time of the survey, when it was common to identify irrigated land as equivalent to substantially larger areas of un-irrigated land. We cannot readily use value of cultivated land owned plus value of non-land assets, as suggested by Hossain's account of eligibility, because, although the data report land values before MF, they do not report the value of assets prior to MF. Hence we focus on using land area and value of land as eligibility and mis-targeting variables (although we do report results using an eligibility cut off in terms of value of cultivable land plus current value of non-land assets).

Land Unit values

In order to construct eligibility in terms of cultivable land area we first replicate Pitt's (1999) tables 1-3. We are able to (almost) exactly replicate tables 1 & 2, which report areas of cultivable land owned by participating households, by MFI membership and for participants with more than 50 decimals, both prior to joining the MFI and at the time of the survey. We can approximate table 3. Details of our replication are given in Appendix 2.

---

[42] We are grateful to David Roodman for pointing us in this direction and putting us out of our misery.

[43] Although the latter could be explained by the inappropriate nature of non-land assets data.

[44] An alternative would be to define the "mis-targeted" in the same way that the "non-target" were defined. Since this would have to be ex-post for the mis-targeted, it would have to involve some sort of indicator which performs better than cultivable land owned does for the non-target.

These results would appear to confirm Pitt's (1999) argument that the unit values of land of mis-targeted households are significantly lower than those of non-participating or appropriately targeted household. However, this is a mis-interpretation of the results because the mis-target dummy variable is derived from area of land which is also included in the estimation. Specifically, the model does not show that the unit price is less than for "other households with similar quantities of land" (Pitt, 1999:3). Unit values of land differ because participants and target (eligible) households have different proportions of land of different categories, with the eligible having a higher proportion of the total land as house plots (Table 5) which is generally of higher unit value. The intuition is shown in Figure 4 and Figure 5[45].

Table 5: Land and land value by Target and Participant Status

	Target		Non-target	
	Participant	Non-participant	Participant	Non-participant
	1	2	3	4
Proportion of housing land	69.04	73.1	13.49	9.76
Proportion of own cultivated land	12.86	12.72	54.94	62.92
Proportion of other cultivable land	2.04	4.66	12.26	15.97
Unit value (taka/decimal)	2201.26	1972.94	905.47	773.31

Weighted means.

Target is defined in terms of cultivated land ownership.

[45] These figures differ particularly for the non-target category which is data defined (>= 50 decimals) in the former, and using nontrgth == 1 in the latter.

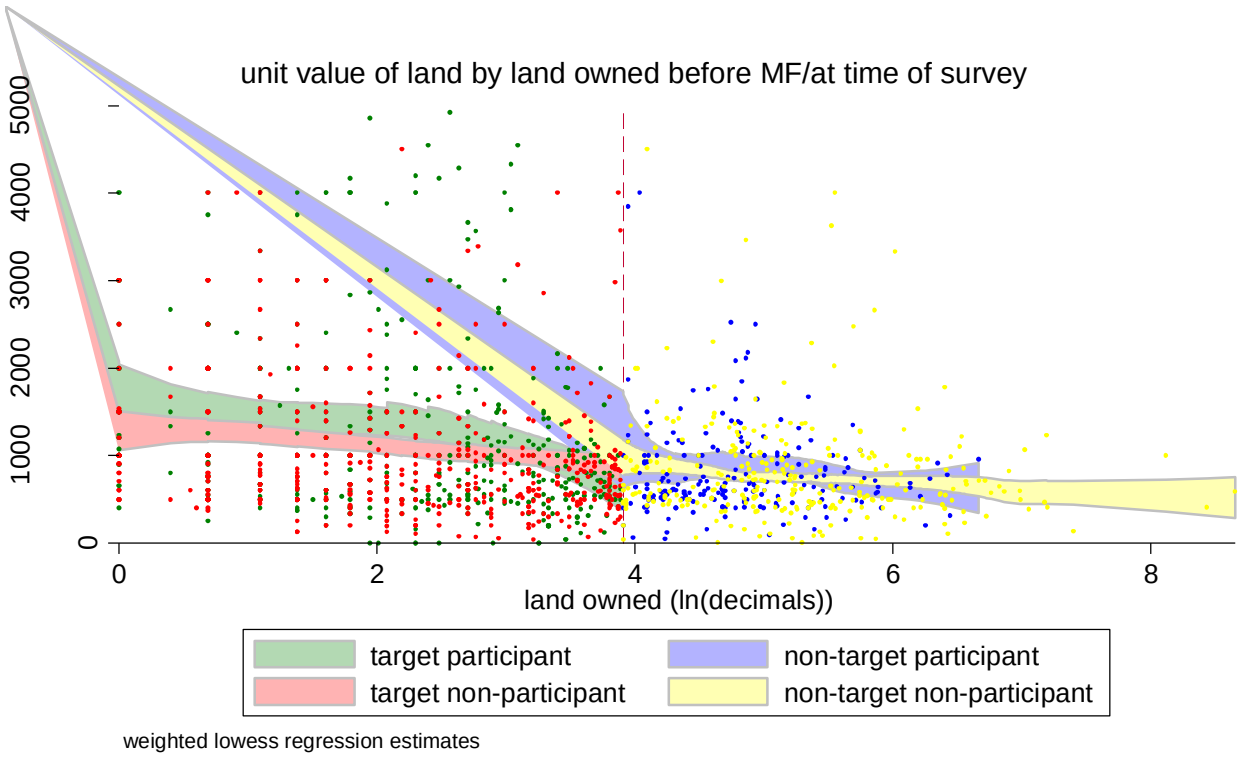

Figure 4: Locally weighted regression of land unit values on land owned, by categories of participation and eligibility (target = < 50d cultivable land owned).

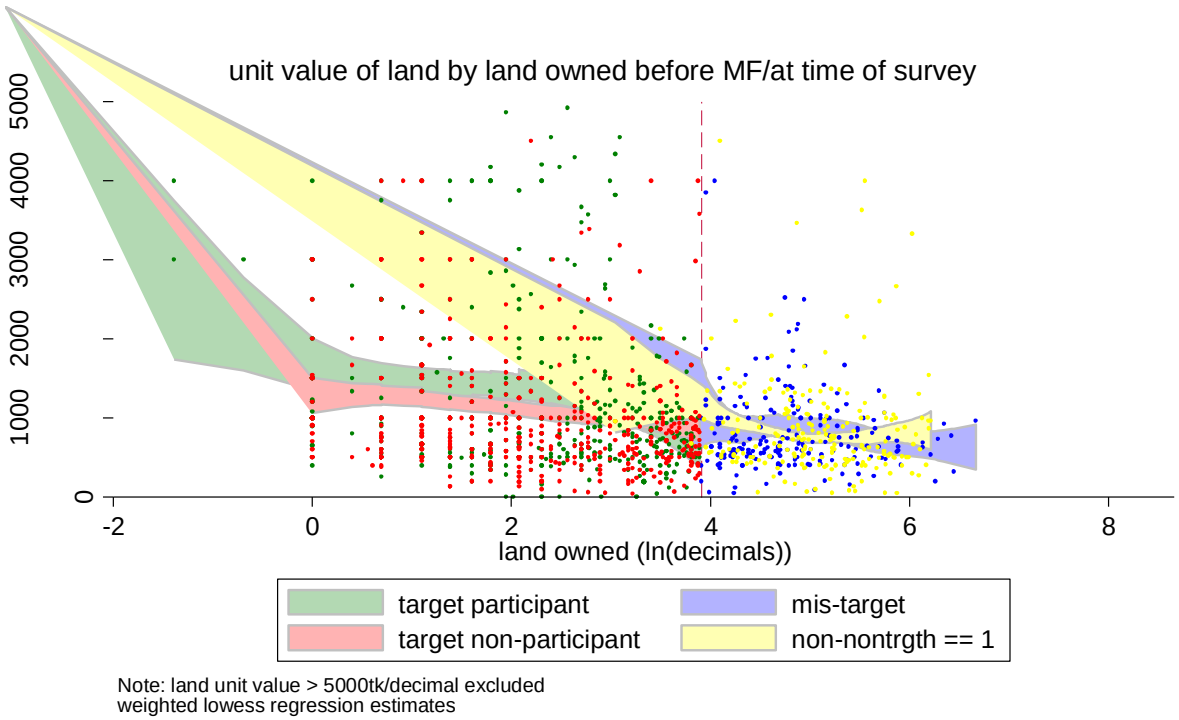

Figure 5: Locally weighted regression of lf land unit values on land owned, by eligibility and participation, mis-target & nontarget (omits extreme values).

Hence, the specification used by Pitt cannot be interpreted in terms of the value of β_3 (in equation 1 below).

$$UV = \alpha + \beta_1 A + \beta_2 A^2 + \beta_3 M + \beta_4 P + \delta T + \varepsilon \qquad (1)$$

Where UV is the unit value of land, which we computed a total land owned divided by total value of land, A is total area owned, M is a dummy for mis-targeted (total or cultivated land owned > 50 decimals), P is a dummy for participation and T is a set of dummies for thana fixed effects. To show that the unit values of land of "mis-targeted" households are less than for non-target households we need to include a variable for the non-target group. We need to estimate:

$$UV = \alpha + \beta_1 A + \beta_2 A^2 + \beta_3 M + \beta_4 P + \beta_5 NT + \delta T + \varepsilon \qquad (2)$$

Where NT is a variable taking the value 1 for households identified as "non-target" (i.e. in PKexp.dta, the variable nontrgth). We can then test the difference between β_4 and β_5. This turns out to be non-significant.

We can demonstrate more clearly whether the mis-targeted have lower unit land values than the non-participating non-target households where the latter is defined in terms of cultivable land owned being below or above the putative eligibility cut-off. Thus we can estimate:

$$UV = \alpha + \beta_1 A + \beta_2 A^2 + \beta_3 E + \beta_4 P + \beta_5 E * P + \delta T + \varepsilon \qquad (3)$$

Where E is a dummy taking the value 1 for eligible households and E*P is the interaction of eligibility and participation dummies. The results of estimating this specification are in Appendix 2 Table 10 column 6 (which otherwise replicates Pitt, 1999, Table 3), and in Table 6. While these models ((1)-(3)) are similar, the last provides an easier and more appropriate interpretation in showing that the key variable determining land unit values is "target" defined in terms of cultivable land owned, rather than participation[46]. Table 10 also shows that the coefficient on "non-target" (nontrgth) is not significantly different from that on "mis-targeted".

These results (Table 6 & Table 10) make it clear that the comparison between non-target participants and non-target non-participants, where target is defined either by having less than 50 decimals, the non-target variable defined in PnK, or some other criterion as discussed below, shows that the difference in land unit values between participants and non-participants is not statistically significant (and not large). When non-target participants (mis-targeted) are compared with the non-target non-participants the results show that on the basis of land assets there is no difference between many households within each of the target/non-target classifications. This is further demonstrated in Table 6 where we include classification by area of land, and also by value of land.

[46] The Stata command nestreg shows that adding the non-target variable to either of the specifications reported by Pitt (1999) table 3, is highly statistically significant.

Table 6: Determinants of Unit land Values by Target Status and Participation

	Target status by area of land				Target status by value of land[47]			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Land before	-0.467	-0.495	-0.659	-0.627	-5.136***	-4.458***	-4.011***	-3.469***
	(0.592)	(0.487)	(0.497)	(0.408)	(0.550)	(0.458)	(0.464)	(0.387)
Land before^2	0.000	0.000	0.000	0.000	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Non-target/ non-participant	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)
Non-target/ non-participant	58.0	84.5	-125.3	-192.4	252.6	530.4*	-129.0	-14.8
	(312.8)	(281.1)	(263.1)	(239.1)	(285.5)	(256.2)	(242.3)	(220.5)
Target/ participant	1288.6***	1083.3***	930.8***	792.4***	-1603.0***	-1337.9***	-1207.4***	-981.6***
	(233.5)	(188.5)	(197.1)	(158.2)	(218.5)	(178.0)	(184.4)	(150.8)
Non-target/ non-participant	1288.6***	1311.2***	786.9***	717.9***	-1466.0***	-1115.4***	-1132.9***	-954.4***
	(257.0)	(220.7)	(217.0)	(189.5)	(250.1)	(217.9)	(212.013)	(188.2)
_cons	918.5***	896.2***	550.8	611.7	2994.8***	2634.4***	2003.9***	1819.1***
	(198.9)	(160.5)	(354.7)	(320.7)	(184.3)	(150.6)	(347.9)	(316.8)
Num. Obs.	1404	1701	1404	1701	1404	1701	1404	1701
R2	0.052	0.051	0.368	0.370	0.077	0.072	0.379	0.377
thana fe	n	n	y	y	n	n	y	y
domain	program	all	program	all	program	all	program	all
Wald tests								
target within participants	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
target within non-participants	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
participants within target	1.000	0.228	0.423	0.646	0.530	0.249	0.684	0.870
participants within non-target	0.853	0.764	0.634	0.421	0.376	0.039	0.594	0.946

Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001.

Notes: In this table target is defined independent of participant status, and likewise with participation; thus there are no non-target.[48]

[47] Total value of land either before access to MF or at the time of the survey less than 40000Tk (50 decimals * unit value of land = 1000). The results do not depend crucially on the unit value of land.

[48] This two way classification makes this clear:

Participant	Target status		
	Non-target	Target	Total
Non-participant	295	578	873
Participant	206	622	828
Total	501	1,200	1,701

We can further support the analysis by looking at the distributions other variables of mis-targeted and non-targeted households including areas and value of cultivable land, and total assets and so on. There are for each variable considerable overlap in these distributions, suggesting that they do not provide good reasons to drop non-target households from the analysis. Even when we define non-target status using the "nontrgth" variable specified in the PnK data set (corresponding to the csg variable in 11CRr files) there is considerable overlap in the distributions of variables such as area (Figure 6), total value (Figure 7), and unit value of land (Figure 8), and total assets owned (Figure 9) of mis-and non-target households. Figure 10 shows that for non-target households with land below the 0.5 acres cut-off for eligibility non-target and mis-targeted households have similar values of assets. Of course it is possible that the total value of assets of mis-targeted households was lower than for non-target households prior to the mis-targeted borrowing from MFIs, the similarity in assets is striking.

Even if there were statistically significant differences in these variables between mis- and non-target groups there is also considerable overlap in their distributions; thus at least some of the non-target households can provide the best matches in terms of propensity scores to so-called mis-targeted households unless there is some unobserved or unobservable variables that are likely to have distinguish them .

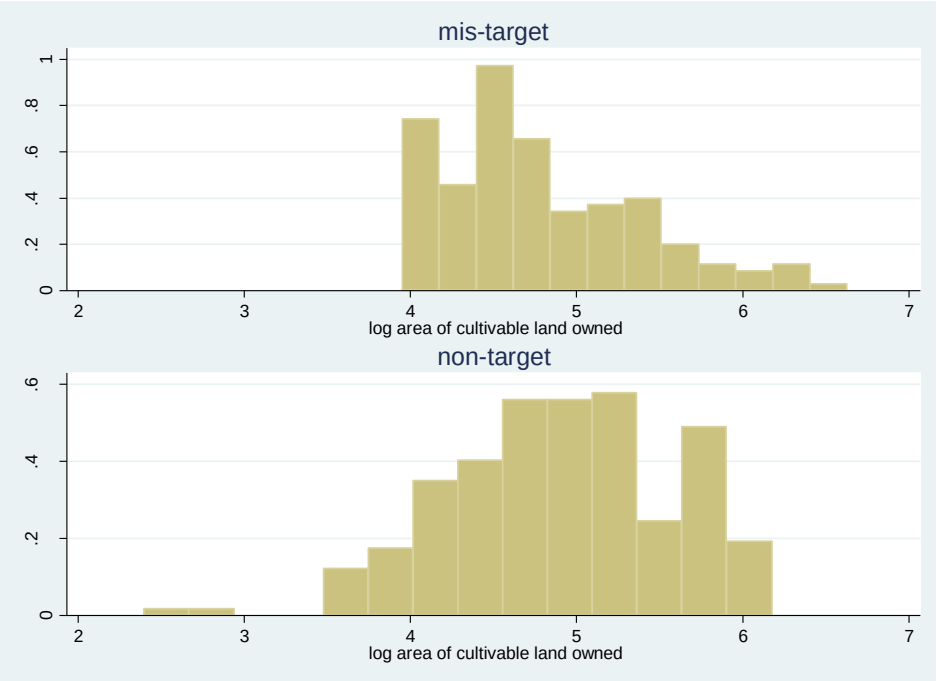

Figure 6: distribution of area of cultivated land owned: mis-target vs non-target.

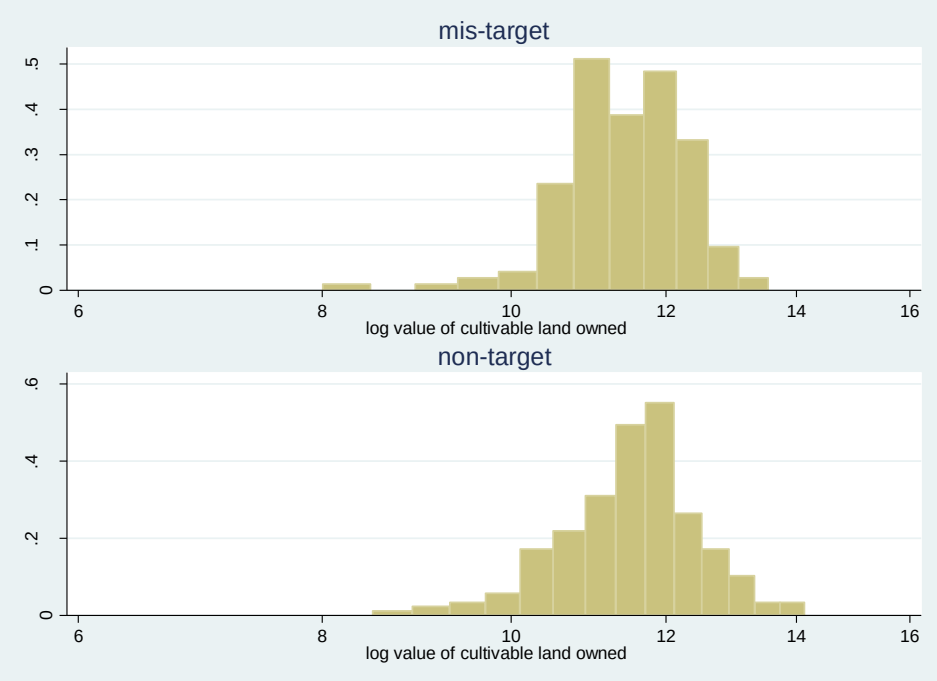

Figure 7: distributions of value of cultivable land owned – non-target vs mis-target.

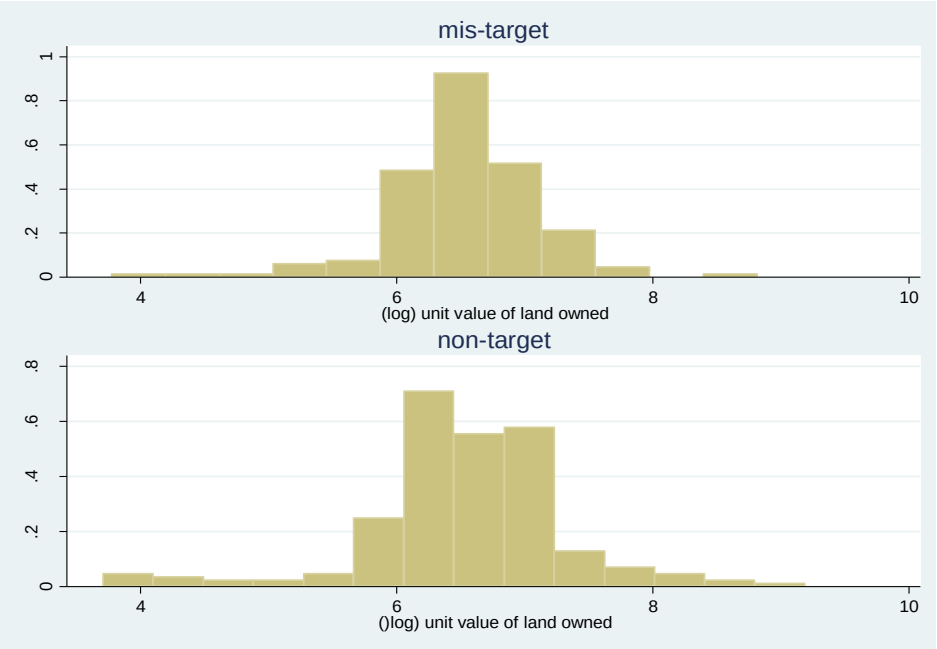

Figure 8: distributions of (log) unit values of land owned – non-target vs mis-target.

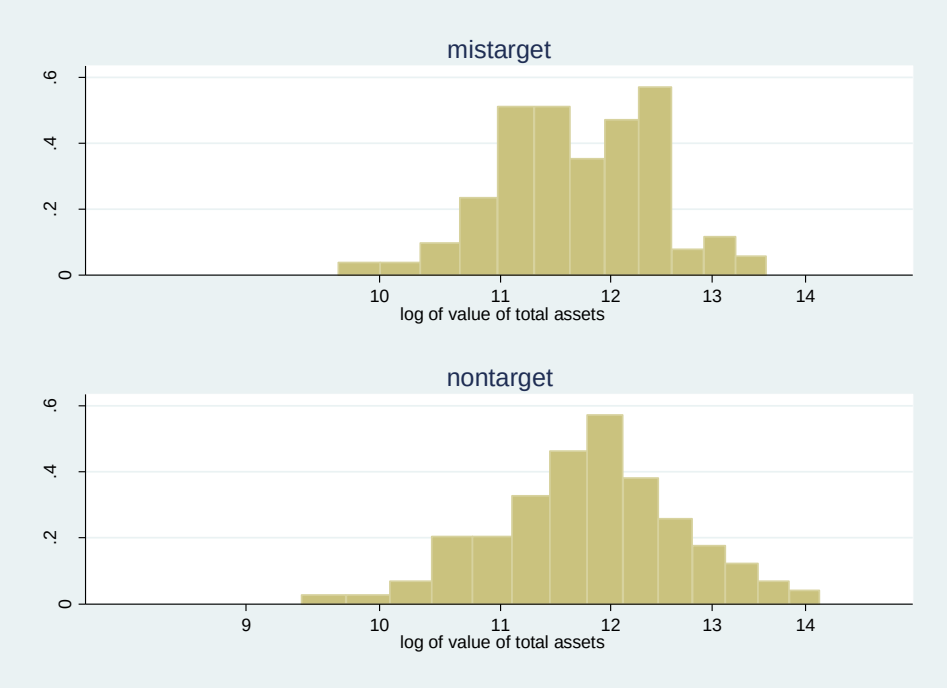

Figure 9: distributions of (log) total value of assets– non-target vs mis-target.

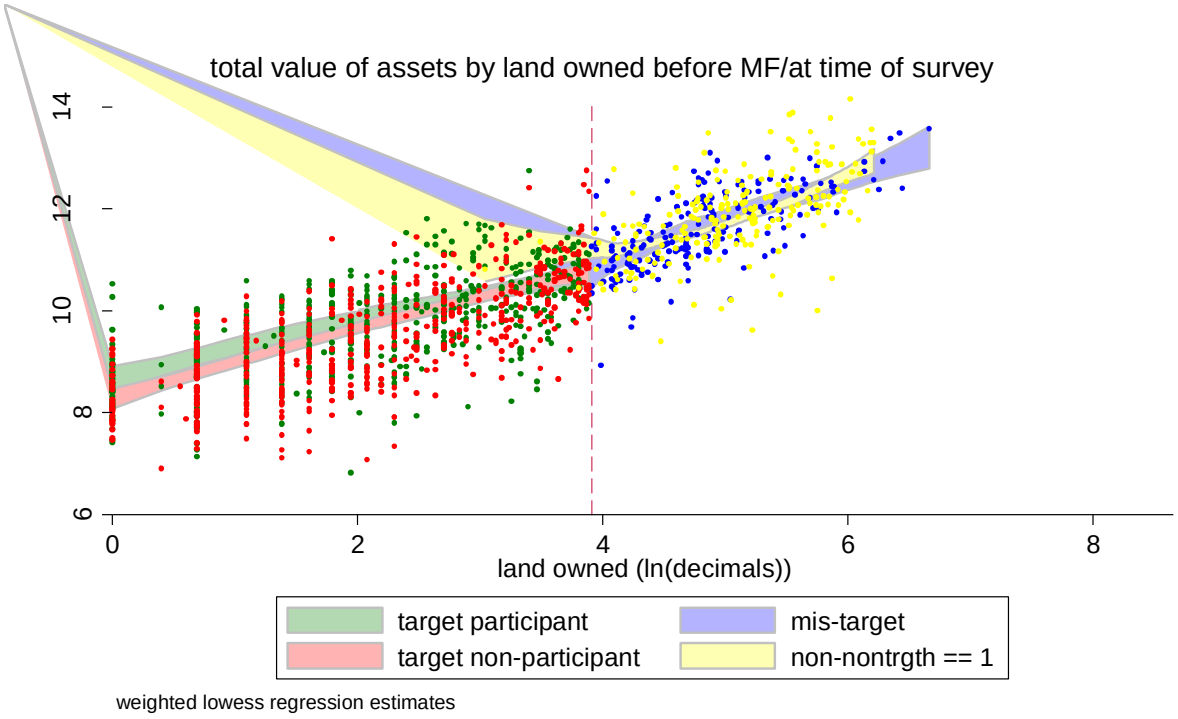

Figure 10: Total value of land and no-land assets by land owned.

This suggests that the PnK model requires some modification to the "non-target" dummy variable that would enable PSM to match between mis-target and non-target households since, by the eligibility criterion generally put forward (cultivable land owned, or total assets), there is little to distinguish them that we have so far been able to discern. Hence, households can be matched on propensity scores estimated by logit models which do not exclude the non-target households.

We do not discuss further the evidence on differences in unit values of land that can be derived from analysis of data in the data files 13ARr and 06ARr.

Replication of DPJ: dropping the non-target dummy variable

The PSM with village fixed effects or village covariates has so far been undertaken using the nontrgth variable which causes non-target households to be dropped. We now conduct PSM without the nontrgth variable. Pitt frequently emphasises the importance of this variable which seemingly derives from the sampling scheme of the survey since it does not correspond to the eligibility criteria emphasised in PnK (i.e. the distributions of owned cultivable land, and of value of total assets, are not obviously different to those of the "mis-targeted" households in the sample), or the total value or the unit values of owned cultivable land emphasised by Pitt (1999).

We have argued that there is little reason to distinguish between "mis-targeted" participants and non-target non-participants in terms of either cultivated area owned or value of land owned. If the non-target households in fact have no unobserved variables which could further confound the association of treatment with impact at least by comparison with the "mis-targeted" variables, we should not include a non-target variable such as the "nontrgth" variable in the propensity score estimation, as used by MP in his replication of DPJ. The effects of observed variables should of course be taken into account in the propensity score and matching process which would let the data decide which households to match rather than have this imposed by the researcher. We can assess the extent to which non-target households are in fact matched with mis-targeted households (or the weights attached to them if we use kernel matching). We can also assess the effects of using alternative non-target variables which may better identify households which are non-target, for example, according to the data on cultivable land area or value of assets which were rather more realistic criteria than strict criteria according to area of cultivable land owned (see discussion above). We will also conduct SA on these results.

Because the results are quite predictable we do not give much detail. We first drop the nontrgth variable as it seems it does not credibly identify a unique set of households; this allows the households that are dropped by MP to be included as potential matches.

Table 7 reports the results of a number of these experiments (Appendix 3, Table 12 provides an equivalent table for male borrowing households). First we report the effects of using MP's estimation comparing female borrowing households with non-borrowing households in treatment villages. We exclude male borrowing households and include dummies for villages 11 and 143, and alter the wchoice variable accordingly. Models 1 & 2 show the estimated impact of female borrowing when the non-target variable is included; it is positive but highly vulnerable to confounding by unobserved variables[49]. Model 3 reports the comparison of households with female MF borrowing with matched households in control villages, using Chemin's method (see above). In this case the impact is also positive, and not significant.

The remaining models in Table 7 all use village covariates rather than village fixed effects. This allows all the comparable households to be included whether in female borrowing villages, male only

[49] In this table we report only results from 10 nearest neighbour estimations of sensitivity to hidden bias.

borrowing villages, or control villages; male borrowing households are excluded. Models 4 & 5 include the non-target dummy variable and hence still drop these households. Impacts are positive but vulnerable to unobservables. In models 6 & 7 we drop the non-target dummy variables allowing these households to be matched with any household with female borrowing. Male borrowing households are dropped. The estimated impacts are negative but not significant (and hence highly vulnerable to unobservables). The remaining models report results of including non-target variables based on land owned and on value of assets owned. In the remaining cases, including a non-target variable results in a positive impact highly vulnerable to unobservables.

Matching without a non-target dummy enables non-target households to be included in the propensity score and matching process; we find that the non-target households are more likely to be matched with mis-targeted households than target households whether target is defined as owning less than 50 decimals or less having assets worth less than 100,000Tk[50]. But a considerable number of the non-target households match with target households even though land area or assets are included in the propensity score. Undoubtedly one of the problems with application of PSM to these data is the low explanatory power of the logit used to estimate the propensity score. At this point we do not engage in pursuit of better models of participation.

Clearly, the inclusion of a plausible data driven non-target dummy variable is associated with positive and significant results, whereas excluding such a variable results in estimates of impact which not statistically significant. However, all the positive and significant results are highly vulnerable to hidden bias, which, as we have argued above, is highly likely to be present in the absence of variables reflecting entrepreneurial energies and abilities. Moreover, the estimated impacts are small.

[50] Results available from authors, and in .do files for this paper.

Table 7: PSM Results: Impact and Sensitivity Analysis – female borrowers[51]

Model	Treat - ment variable	Domain[a]	Matching method	Non-target variable[b]	Fixed effects.village covariates FE	covars	N	Difference mean	se	p-value	t-diff	Critical gamma	Confidence levels 95% hi	lo
1	creditff	treat	kernel	nontrgth	y	n	2511	0.0431	0.0173	0.0126	2.4989			
2	creditff	treat	nn10	nontrgth	y	n	2511	.03884	.0183	.03411	2.1208	1.25	.11	0
3	creditff	control	kernel	nontrgth	y	n	2182	0.0169	0.0174	0.3316	0.9713			
4	creditff	all	kernel	nontrgth	n	y	3487	0.0379	0.0157	0.0161	2.4089			
5	creditff	all	nn10	nontrgth	n	y	3487	0.0463	0.0164	0.0049	2.8151	1.3	0.17	0
6	creditff	all	kernel	none	n	y	4190	-0.007	0.0154	0.6417	-0.465			
7	creditff	all	nn10	none	n	y	4190	-0.0004	0.0166	0.9790	-0.0263	1.20	0.00	1
8	creditff	all	kernel	nontar1	n	y	3524	0.0394	0.0160	0.0137	2.4669	.	.	.
9	creditff	all	nn10	nontar1	n	y	3524	0.0394	0.0168	0.0190	2.3473	1.30	0.13	0
10	creditff	all	kernel	nontar4	n	y	3697	0.0370	0.0154	0.0163	2.4040	.	.	.
11	creditff	all	nn10	nontar4	n	y	3697	0.0448	0.0163	0.0060	2.7542	1.30	0.22	0
12	creditff	target[c]	nn10	nontar1	n	y	2877	0.0330	0.0182	0.0705	1.8108	1.20	0.16	0

Notes: creditff = all female credit households; fhh1 – households with female credit in round 1.

a. treat = treatment villages only; control -= all female borrower households AND all households in control villages; all – all households without female borrower.
b. nontrgth = MP's non-target variable; nontar1 = 1 if non- borrower AND household cultivable land > 50d; nontar4 = non-borrower & total households assets > 100,000Tk.
c. All households excluding mis-targeted households.

[51] An equivalent table for households with male borrowers is given in Appendix 3, Table 12.

The results with RnM's data are equivalent to those reported above (which used MP's data set except where augmented with variables which we computed because they are not available in PKexp.dta). Female borrowing does not appear to have significant positive impacts on the household per capita consumption if we drop the nontrgth variable. When we substitute the design based nontrgth variable with ones which are more data driven the estimated impacts are positive, but mainly insignificant and are still highly vulnerable to hidden bias.

Conclusions

Pitt's reply has considerably clarified matters; the estimations in DPJ have been revealed as full of errors, some of which are shared with Chemin, others due to lack of accurate documentation in Chemin, and some our own (see also DPJb & c). Much of the difficulty we experienced in replicating Chemin related to lack of documentation of the variables and the sample in Chemin. Many of the problems for which MP criticises us are already present in Chemin, but MP does not comment on this, any more than he comments on the difference between his results for treatment villages only and Chemin's[52].

The problems of constructing an estimation data set from the PnK raw data set and information made available by the original authors have also been aired; many of the problems we faced, some of which are offered as mitigating circumstances, arose because of the lack of code performing the original data construction and estimation, inadequate description of the survey instrument and associated data variables, and qualified cooperation from the original authors. Cooperation from original authors will often be restricted because it can become demanding, but recourse to original authors would have been unnecessary had the original data been adequately documented and variable construction and estimation code made available by the original authors either at the time of publication or when it became clear there was a demand for it[53].

It may be the case that archiving of data and code to an appropriate standard was not common at the time PnK were writing, although the need for this had been frequently noted, and often advocated. The AER, for example, initiated a voluntary code in the 1980s following the critical paper by Dewald et al (1986). Interest in replicating PnK started with Morduch in 1998, and by Roodman (and Morduch) around 2008[54]. While a version of the data was available through the World Bank it was (and is, as of early January 2013) in a poor state. The interest demonstrated in replication could have prompted the original authors to make available better documentation and code to replicate their papers as by this time the practice of archiving was mainstream (e.g. Bernanke, 2004). MP has reported that the code had been lost, but one can wonder whether, given the interest expressed in (and doubts about) PnK, as well as the numerous publications arising from these data by the original authors and their co-workers, it would not have been appropriate to re-do the original work in a form that would convince would be replicators (and others), earlier.

---

[52] MP gets positive and significant impacts of female MF borrowing while Chemin gets negative and significant results for this comparison (Chemin Table 4 row 1). It could be argued that they are not comparable because MP is concerned with female borrowing while Chemin includes all borrowers, and the propensity score and sample restrictions are different. However the difference is striking and surely noteworthy.

[53] We do not agree that our replicatory intentions were not adequately communicated to Chemin; see DPJb.

[54] And Pitt was a co-author of a paper using these data as recently as 2006.

The difficulties of reasonably applying PSM to these data have been further exposed, although both Chemin's methods and the use of village level covariates instead of village-level fixed effects provide usable approaches; however, we only have some confidence in the use of village covariates. These PSM methods cannot deal with hidden bias, but SA does allow estimates of the likely degree of vulnerability of the results to hidden bias. The results of SA of both MP's method and those using control villages described here indicate high levels of vulnerability to hidden bias[55]. MP's method using treatment villages only, comes to different conclusions to Chemin for the same comparison, although MP's estimates are restricted to female borrowers. However, as noted, even MP's estimate is vulnerable to hidden bias. Since both theory and ethnographic research suggest that it is likely that there are unobserved variables confounding MF borrowing and outcomes such as per capita expenditure, we conclude that these results do not contradict DPJ's conclusions and that we should exercise extreme caution in interpreting these data as showing significant benefits to the poor from female MF borrowing.

Without trying to excuse ourselves, many of the problems in the estimation related to the difficulty in replicating Chemin's descriptives and following Chemin's propensity score model (particularly using endogenous RHS variables). There is a lesson here for replicators to take a more thorough going and critical view of the work they are replicating. We also followed Chemin's choice of control units neglecting, as did Chemin, that some were individuals who themselves did not borrow but were living in households in which another individual obtained loans from MFIs – e.g. were treated. None of this excuses, although it may partly explain, some of the errors we made, but it does point to difficulties in establishing the credentials of original papers lacking ready replicability. Replication, even though faulty, has clarified matters and suggests significantly different conclusions to the original papers. PnK has only been (purely) replicable after very considerable effort, and the results so far, whether by replication using WESML-LIML-FE (by RnM) or by PSM, do not contradict the view that reliance on them for policy analysis would require "incredible" belief (Manski, 2011). If, as MP seems to think, PSM has no bearing on interpretation of PnK, then he should respond fully to the critique of the method used in PnK by RnM[56].

The results of these replications fail to confirm the robustness or perhaps the existence of the main conclusions of PnK, and it is not clear that any of MP's critiques should alter our interpretation of the experience, namely that "You can't fix by analysis what you bungled by design" (Light et al, 1990: viii), a rather more graphical statement of the conclusion presented by DPJ in their last paragraph.

MP writes that replicators lack creativity; maybe, but perhaps the greatest creativity in PnK (and Chemin) is to have found in their data a robust positive effects of MF borrowing by the poor on the well-being of their households, especially when delivered through women, and seeing this as a

[55] Again, it can be objected that impose the most demanding criteria, e.g. that one or more missing variables account for all the observed impact, and that when one has strong prior beliefs in the correctness of the results one can accept lower standards (Haughton, 2011). However, we doubt these are valid reasons for adopting less stringent criteria than used in the medical sciences (Duvendack and Palmer-Jones, 2011b&c).

[56] Pitt and Khandker (2012) do not respond to the vulnerability of the PnK estimates to outliers as described in RnM (2011), and it is not clear that much of the critique offered does not support RnM's critique (see http://blogs.cgdev.org/open_book/ "Perennial Pitt and Khandker", 10/12/2012).

general characteristic of microfinance. The evidence does not seem to be robust even if there is such an effect, in these data. Until, either new analyses using the WESML-LIML-FE or PSM can convincingly overturn this conclusion, or another convincing replicable method developed, we think our conclusion is maintained, in spite of the egregious errors we made. Perhaps Wyatt Earp *has* met *his* High Noon?

References

Abadie, A. & Imbens, G., 2008. On the Failure of the Bootstrap for Matching Estimators. *Econometrica*, 76(6):1537-1557.

Acemoglu, Daron, Johnson, Simon & Robinson, James A., 2000. The Colonial Origins of Comparative Development: An Empirical Investigation, National Bureau of Economic Research. *Available at: http://www.nber.org/papers/w7771 [Accessed July 6, 2012].*

Acemoglu, D., Johnson, S. & Robinson, J. A., 2001. The colonial origins of comparative development: an empirical investigation. *American Economic Review*, 91:1369–401.

Acemoglu, Daron, Johnson, Simon & Robinson, James A., 2005. Institutions as the Fundamental Cause of Long-Run Growth' P. Aghion & Durlauf, S.N., eds. *Handbook of Economic Growth*, 1:385–472.

Acemoglu, D., Robinson, J.A. & Johnson, S., 2011. Hither thou shalt come, but no further: Reply to "The Colonial Origins of Comparative Development: an Empirical Comment".

Acemoglu, D., Johnson, S. & Robinson, J.A., 2012. Hither Thou Shalt Come, But No Further: Reply to "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment. *American Economic Review*, forthcoming.

Albouy, A., 2004. The Colonial Origins of Comparative Development: A Reexamination Based on Improved Settler Mortality Data.

Albouy, A., 2008. The Colonial Origins of Comparative Development: An Investigation of the Settler Mortality Data.

Albouy, D.Y., 2012. The Colonial Origins of Comparative Development: an Empirical Investigation: Comment Investigation of the Settler Mortality Data. *American Economic Review*, forthcoming.

Armendáriz, B. & Morduch, J., 2005. The Economics of Microfinance, Cambridge: MIT Press.

Bateman, M., 2010. Why Microfinance Doesn't Work? The Destructive Rise of Local Neoliberalism, London: Zed Books.

Bernanke, B.S., 2004. Editorial Statement. *American Economic Review*, 94(3):404.

Caliendo, M. & Kopeinig, S., 2008. Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22(1):31-72.

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).

Concato, J., Shah, N. & Horwitz, R.I., 2000. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *New England Journal of Medicine*, 342(25):1887–1892.

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., *et al*., 2011a. *What is the evidence of the impact of microfinance on the well-being of poor people?* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Duvendack, M. and R.W. Palmer-Jones, 2011b. Comment on: Abou-Ali, H., El-Azony, H., El-Laithy, H., Haughton, J. and Khandker, S., 2010. Evaluating the impact of Egyptian Social Fund for Development programmes. Journal of Development Effectiveness, 2 (4), 521-555. Journal of Development Effectiveness, 2011, 3(2):297-299.

Duvendack, M. and R.W. Palmer-Jones, 2011c. Much ado about something: response to Haughton's reply to Duvendack and Palmer-Jones, Journal of Development Effectiveness, 3(2):302-330.

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

Duvendack M. & Palmer-Jones, R. 2012b. What Mat[t]hieu Really Did; Rejoinder to Chemin, 2012b, online (DPJb).

Duvendack, M. & Palmer-Jones, R. 2012c. Reply to Chemin and to Pitt. *Journal of Development Studies,* 48(12) (DPJc).

Duvendack M. & Palmer-Jones, R. 2012d. Wyatt Earps' High Noon; Rejoinder to Pitt, 2012a, this document (DPJd).

Duvendack M. & Palmer-Jones, R., forthcoming. Replication of Quantitative Work in Development Studies: Experiences and Suggestions. *Progress in Development Studies*.

Fernando, J.L., 1997. Nongovernmental Organizations, Micro-Credit, and Empowerment of Women. *The ANNALS of the American Academy of Political and Social Science*, 554(1):150-177.

Gelman, A. & Hill, J., 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: Cambridge.

Glandon, P., 2010. Report on the American Economic Review Data Availability Compliance Project. *Available at: http://www.aeaweb.org/aer/2011_Data_Compliance_Report.pdf.*

Goetz, A.M. & Gupta, R.S., 1996. Who takes the credit? Gender, power, and control over loan use in rural credit programs in Bangladesh. *World Development*, 24(1):45–63.

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

Haughton, J., 2011. Reply to comment by Duvendack and Palmer-Jones. *Journal of Development Effectiveness*, 3(2):300–301.

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.*

Hoxby, C. M., 2000. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *The Quarterly Journal of Economics*, 115(4):1239–1285.

Hoxby, C M., 2005. Competition Among Public Schools: A Reply to Rothstein (2004). *National Bureau of Economic Research Working Paper Series*, No. 11216. *Available at: http://www.nber.org/papers/w11216.*

Hoxby, C.M., 2007. Does Competitions am Public Schools Benefit Students and Taxpayers? A Reply. *American Economic Review*, 97(5):2038–2055.

Hulme, D., 2000. Impact Assessment Methodologies for Microfinance: Theory, Experience and Better Practice. *World Development*, 28(1):79-98.

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.

Iversen, V. & Palmer-Jones, R., forthcoming. Draft of an Outline of Good Practice for Replication of Quantitative Impact Evaluations: Draft Report for "Consultancy to develop Replication Guidelines for 3ie", October, 2011.

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.

Karim, L., 2011. *Microfinance and Its Discontents: Women in Debt in Bangladesh.* University of Minnesota Press: Minneapolis.

Khandker, S.R., Koolwal, G.B. & Samad, H.A., 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. World Bank Publications: Washington D.C.

Leamer, E.E., 1983. Let's take the Con out of Econometrics. *American Economic Review*, 73(1):31–43.

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*.

Light, R.J., Singer, J.D. & Willett, J.B., 1990. *By Design: Planning Research in Higher Education*. Harvard University Press: Cambridge, Mass.

Manski, C.F., 2011. Policy Analysis with Incredible Certitude. *The Economic Journal*, 121:261-289.

Meyer, M.M. & Fienberg, S.E., 1992. *Assessing Evaluation Studies: The Case of Bilingual Education Strategies.* National Academy Press: Washington D.C.

McCullough, B.D. & McKitrick, R., 2009. Check the Numbers: The Case for Due Diligence in Policy Formation.

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

Peng, R.D., 2011. Reproducible Research in Computational Science. Science, 334:1226–1227.

Pitt, M.M., 1999. Reply to Jonathan Morduch's 'Does microfinance really help the poor? New evidence from flagship programs in Bangladesh'. Unpublished mimeo.

Pitt, 2011a, Response to Roodman and Morduch's "The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence. *Available at: www.pstc.brown.edu/~mp/.*

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

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

Pitt, M.M. & Khandker, S.R., 2012. Replicating Replication: Due Diligence in Roodman and Morduch's Replication of Pitt and Khandker (1998). *Available at: http://www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2012/11/21/000158349_20121121125524/Rendered/PDF/wps6273.pdf [Accessed November 30, 2012].*

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.*

Roodman, D., 2011a. Response to Pitt's Response to Roodman and Morduch's Replication of …., etc. *Available at: http://blogs.cgdev.org/open_book/2011/03/response-to-pitts-response-to-roodman-and-morduchs-replication-of-etc.php.*

Roodman, D., 2011b. A somewhat Less Provisions Analysis of Pitt and Khandker. *Available at: http://blogs.cgdev.org/open_book/2011/04/a-somewhat-less-provisional-analysis-of-pitt-khandker.php.*

Roodman, D., 2011c. Due Diligence: An Impertinent Inquiry into Microfinance, Centre for Global Development, Washington, D.C.

Roodman, D., 2012a. Perennial Pitt and Khandker. *Available at: http://blogs.cgdev.org/open_book/2012/12/perennial-pitt-khandker.php.*

Rothstein, J., 2007a. Does Competition Among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000). *American Economic Review*, 97(5):2026–2037.

Rothstein, J., 2007b. Rejoinder to Hoxby. *Available at: http://gsppi.berkeley.edu/faculty/jrothstein/hoxby/documentation-for-hoxby-comment/.*

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

Roy, A. 2010. *Poverty capital: microfinance and the making of development*. Routledge: London.

Shadish, W.R., 2012. Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology. O*nline 18/11/2012. *Available at: http://www.springerlink.com/index/10.1007/s11292-012-9166-8 [Accessed December 22, 2012].*

Sinclair, H., 2012. *Confessions of a Microfinance Heretic: How Microlending Lost its Way and Betrayed the Poor.* Berret-Koehler Publication: San Francisco.

Stewart, R. et al., 2010. What is the impact of microfinance on poor people? A systematic review of evidence from Sub-Saharan Africa. *Available at: http://www.dfid.gov.uk/R4D/PDF/Outputs/SystematicReviews/MicroFinance_FOR+WEB%5B1%5D.pdf.*

Stewart, R. et al., 2012. Do micro-credit, micro-savings and micro-leasing serve as effective financial inclusion interventions enabling poor people, and especially women, to engage in meaningful economic opportunities in low- and middle-income countries? A systematic review of the evidence. *Available at: http://www.dfid.gov.uk/r4d/pdf/outputs/systematicReviews/Microcredit2012StewartReport.pdf.*

Appendix 1: Specifics

Here we respond to specific points made by MP; we use what we learnt from these points and our responses, to conduct and report our replication of MP.

We need to emphasise that our estimation code contained many errors, some of which were simple coding errors, others due to failing to reinstate correct code after performing various (more or less well conceived) experiments, and yet others due to conceptual problems. The work of variable construction was very demanding (as Roodman found for the related task of constructing the PnK estimation data set[57]); we were unable to closely replicate much of Chemin at least partly because of imprecise reporting by Chemin of key parts of his variable construction (DPJb), so that, partly as a consequence, we undoubtedly devoted less effort to estimation than we should.

We included those with no empirical possibility of treatment [1][58]

According to MP, there are three types of exclusion from possibility of treatment – those living in villages without any MFI, those without a gender specific MFI, and those who are underage or too old. Thus there are some villages with female only MFIs, some with male only MFIs, and some with MFIs which cater to both female and male. Pitt argues it would be inappropriate to include observations from villages in which there is no empirical possibility of participation (MPa:4), and people who are excluded from MF on the basis of age should also be excluded from the analysis.

With regard to this last point, inclusion or exclusion of people under 16 (or over some higher age), will make little if any difference since they will not provide matches to the treated who are within this age range (or will have low weight in kernel matching), if age is included in the propensity score estimation. We used some specifications with age restrictions (>= 15), and found no meaningful difference.

Using observations without choice might be no problem, according to Chemin, or MP, if one could control for village level effects using village-level fixed effects. One cannot use village-level fixed effects for villages without the possibility of treatment because for these villages fixed effects would perfectly predict non-membership. Our code was compromised, as MP could ascertain because we provided it, in a number of ways. Firstly we actually used only Thana level fixed effects (in the sample there are three sample villages per Thana) rather than village-level fixed effects; secondly in some estimations of the propensity scores we failed to exclude the control villages but continued to use village-level fixed effects, thereby treating control Thana as equivalent to the excluded Thana (Thana 1).

It is unfortunate if it is only possible to include households in the same village. One could however use village level covariates if these accounted for differences in placement, or we could follow Chemin who estimates the propensity score for treatment villages only and predicts the propensity score for households in control villages, or one could pursue a variation on Chemin's approach using village level covariates (as discussed in the section "Replication using PKexp"). Any of these would

[57] See the quotation in footnote 13 above.

[58] The numbering of this section relates to the numbering of the points made in MPa.

avoid throwing away information, and would have a better chance of mitigating the selection on unobservables problem that PSM encounters.

Assigns individuals to control groups when they are members of treatment households [2]

Yes, we do mistakenly do this. However, so do Chemin and MP. Chemin does not notice the problem. MP in his replication of our replication has two problems in this regard; firstly there are some treated households (with female borrowers) who also have male borrowers. Secondly, there are male borrower households among the control group. Excluding all households with any male borrowers removes their effects from both the treated and untreated, and also removes households with both male and female borrowers (71 out of 879 in DPJ's sample; 53 out of 728 households in wave 1 of MP's sample). Our estimations using MP's data set suggest that this has no effect on the outcome, which although positive, and statistically significant, is shown below by sensitivity analysis (Rosenbaum, 2002) to be highly vulnerable to unobservables.

Inaccurate claim to include village fixed effects [3]

This is correct - we made a mistake. There are three villages per Thana in the sample. In our code we used Thana to define fixed effects for Thana 2-24, and then we dropped the exclusion of Thana 25-29 from some estimations (e.g. when we did not use "ps2" with psmatch2).

Using village fixed effects compared to Thana fixed effects makes no substantive difference.

Treat observations in Thanas 25 to 29 as equivalent to those in Thana 1 [4]

This is the same as the previous mistake; however, contrary to MP our "preferred" propensity score is estimated with a restriction to treatment villages (see footnote 59).[59]

Inclusion of endogenous variables [5]

In "High Noon" we did this following Chemin's preferred specification (Spec. 3). In other papers we used a specification closer to Chemin's first specification (see $Chemin1 in footnote 60).[60]

---

[59] gen control1 = thanaid <= 24 // all treated vs non-part in treatvill
*Chemin's table 2, row 1:all treated, part vs non-part in treatvill, Chemin's replication:
logit treatgrp2 $Chemin3 $treatvilldumm if control1 //elig_defacto_treatpp
predict ps2
estimates store Ch2
`* later on we use "ps2" in when we use the "pscore()" option in psmatch2.

[60] Pitt says we used Chemin's second specification. We are not clear where in our code he gets this impression. We specify:
global Chemin1 "hgrade male agey agehhhh no_of_adultmales relhhh_1_trland relhhh_2_trland "
global Chemin2 "hgrade male agey agehhhh no_of_adultmales relhhh_1_trland relhhh_2_trland cssv nonfarm livestockvalue hhsize wageag wagenonag agesq age4 no_of_adultfemales agriincome_net sexhhh relhhh_3_trland maxedhhh highedufat highedumat flopt flopi mliv fliv marital equipment transport other halaa injury nfexpfv nferevw agricost dairy"
global Chemin3 "male agey agehhhh no_of_adultmales maxed cssv nonfarm livestockvalue hhsize wageag wagenonag agesq age4"
*Chemin's table 2, row 1:all treated, part vs non-part in treatvill, Chemin's replication:

Use of household level covariates in the propensity score function [6]

The education variable in the data set we sent to Pitt is not invariant among household members and infants and for those less than five years of age is generally coded zero. This education variable does include the appropriate level for those currently attending school, although we use level last year as more appropriate than level this year to represent the level of education achieved.

The education of the household head, and the maximum education of any male, female or any person in the household are rightly invariant at household level.

In the estimation code we did have an unfortunate error which used only the education level of a person who had completed schooling to compute what we thought was Chemin's variable "hgrade", which we had not been able to replicate. As MC clarifies, this variable is in fact a zero/one dummy variable for having any education, which we should have been able to pick up from Chemin, Table 1. The availability of our data preparation code in the .do files we sent to MP enabled him to identify our error. We should have used the correctly calculated "education" variable to compute Chemin's "hgrade" variable, and include this "education" for individual level education in propensity score estimations.

Agricultural and non-agricultural wages in the propensity score function [7]

This was a mistake, although we thought we were following Chemin. However, we had been unable to replicate the descriptives of the wage variables given in Table 1 of Chemin, which are not defined. MC1&2, and code, clarifies that these variables are daily wage rates; our original data construction code was quite close to this, but as a result of experimentation to better approximate Chemin's descriptives, we left erroneous code in our data preparation file. We explained to Pitt in our emails of 6 & 8 July 2012 (listed below in Appendix 4) that there was a problem with our wage calculations, which remained until we received MC's code. Our inability to replicate Chemin's Table 1 derived in significant part because Chemin used operated rather than owned land to define eligibility and to restrict his sample (see DPJb & c). Chemin used "flopt" rather than "halab" in the PnK World Bank data set. We argue that this is inexplicable, and never mentioned in Chemin, who never uses the word "operated".

Release of variable construction code

MP's claim of a striking inconsistency in our behaviour in regard to releasing variable construction code depends on an interpretation of "published" as "published online". JDS had in fact only very recently started making pre-publication copies of forthcoming papers available online and we had not prepared for this.

```
logit treatgrp2 $Chemin3 $treatvilldumm if control1
predict ps2
…
`* And e.g.,
psmatch2 treatgrp2, outcome(lnconsweekpc ) labsupwomMD1 labsupmenMD1 fedec517_rpj medec517_rpj
///
        pscore(ps2) kernel k(normal) bwidth(`k')
```

Perhaps because we specify pscore(ps2) Pitt thinks we are using $Chemin2 but clearly that is not the case as the logit …$Chemin3 command immediately before "predict ps2" makes clear.

MP is wrong to claim (footnote 14) that he cannot make use of the RnM dataset for the same reasons we cannot. Our text, which he quotes, is supposed to mean that we cannot use SQL code in Stata, for obvious reasons. As must be clear to MP, we can both use the posted RnM stata data sets, and the database including the SQL data construction code (at least we can interpret most of it). We do both these things, and we compare our variable constructions with RnM's, and on the whole they agree closely. We have argued that data construction using different software of computer language is often desirable (or completely rewriting code in the same language) in order to check computations and constructions. Iversen and Palmer-Jones (forthcoming) have an example where completely rewriting Jensen and Oster's (2009) stata code (Stata often allows the same computation to be performed in several different ways) identified an error and came up with a plausible alternative variable construction which had significant effects on the results.

Inappropriate p-values [8]

This is correct, although we partly follow the suggestions for stratification matching in the help file for psmatch2. There is a simple solution, which MP does not suggest, instead using the pscore and ATTs commands in his own stratification matching. The solution is to calculate the difference between the observed outcome for the treated observation and that for the matched sample for each case in the strata and then compute the mean and difference from zero of all the cases in all strata.[61]

Wrongly define strata adding yet another misspecification [9]

Sub-classifying treatment cases across the propensity score is one way to perform stratification matching, is seemingly the most commonly performed method of stratification of interval matching, and is that adopted in the Stata "pscore" user written command. However, it is not the only way to undertake stratification matching, and a different procedure is suggested in the psmatch2 user written command. There are disadvantages to using the pscore approach, especially when the distributions of propensity scores of treated and control cases are very different, as was the case with Chemin's analysis of the PnK data. There are not many control cases with propensity score which are close to those of the treated cases (Chemin, 2008: 474; even with MP's estimation using his data set and a very restricted control group the propensity scores of the treated and controls are distributed quite differently). There seems no reason why one should not explore stratification by the outcome variable, or any other variable of interest, and this is suggested by the online help for the psmatch2 command. The help file for psmatch2 suggests using a categorical grouping variable and propensity score estimation within strata of "groupvar" (see footnote 61), as MPa also notes (footnote 17). Thus it is not self-evident that propensity scores should be estimated for the treated as a whole rather than within strata, especially perhaps if one is using a variable other than the

---

[61] E.g. within strata

```
foreach g in groupvar {
psmatch2 …. if groupvar == `g'
replace diff = _outcome - _outcome if `g' == groupvar
}
ttest diff == 0
or reg diff [weight = weight], cluster(nh)
test _cons == 0
```

propensity score for stratification. It is true our p-values were inappropriately calculated but this is easily remedied, and that the propensity score estimation included some endogenous variables, but in this we were following Chemin which we started out replicating. In other work we have used some of the variables suggested by MP (e.g. based on Chemin's specification 1). The example of the lottery winner suggested by MP to illustrate the problems of stratifying by an outcome variable to test the effect of treatment on the same outcome variable seems valid. It was inappropriate to stratify by the outcome variable that was being tested.

Significance by bootstrapping [10]

It is true that neither the code nor the results we reported used bootstrapping. We had earlier made the estimates of significance using bootstrapping but this was not used for the published results, and the statement that we used bootstrapping was not removed from the text. The results of bootstrapping are not meaningfully different from those obtained without bootstrapping, as MP knows from the results of his replication reported in his tables 2 and 3. Whether our significance, had they been conducted on the individual differences between treatment and controls, rather than the means within strata, and with clustering (see point [11], is moot. But, as noted, bootstrapping makes little substantive difference. Furthermore, Abadie and Imbens (2008) cast doubts about the validity of bootstrapping in the context of PSM which encouraged us further not to report and publish the bootstrapped results.

Inappropriate t-statistics [11]

Because he has our .do file MP can see that we calculated our significance tests using the means within strata thereby radically increasing the t-statistics. A more appropriate approach would be to compute the difference between the outcomes and matched controls for each treated case in each strata, as shown in footnote 61.

Used probit rather than logit [12]

Probit is the default procedure for psmatch2, and we did indeed fail to add the "logit" option, as MP could see from the code we supplied. Use of probit rather than logit makes no substantive difference.

Stratification estimates in DPJ Table 2 row 4 have little resemblance to the methods described in the paper [13]

This is not the case, except that we used probit rather than logit to estimate the propensity scores. The code calculating our estimates for Table 2 row 4 shows that we use the "pscore(ps2)" option, where ps2 was estimated from a logit using Chemin's specification 3. MP seems to have ben mislead by our use of the variable name "ps2" into thinking that we used Chemin's specification 2, whereas the logit command immediately prior to "predict ps2" reads "logit depvar $Chemin3 ….", and the global macro "$Chemin3" clearly uses variables equivalent to Chemin's specification 3. The commands used to calculate Table 2 rows 1 – 3 estimate the propensity score directly (i.e. do not make use of the "pscore()" option), and with the covariates specified individually rather than using the $Chemin macro; we use the same covariates as specified in $Chemin3. Also, as noted in [12] we did not specify the "logit" option so in rows 1-3 the propensity scores were estimated by probit with the same covariates. These two differences have no substantive effect on the outcomes.

MP's points out that many of the covariates in our within strata estimates are not estimable, and this is the case. MP's estimates reported in his Tables 2 & 3 also drop inestimable covariates, including the crucial "nontar" variable which defines cases as not eligible for membership of MFIs.

As Morduch (1998) first noted PnK (and MP), treat eligibility of non-members of MFIs (defined by nontrgth = 1) differently in that the former are defined as ineligible if they own more than 0.5 acres of land, regardless of its "value", while the latter are deemed eligible even if the own (often considerably) more than 0.5 acres. Pitt (1999) justifies this on the grounds that the value of land of those who are MFI members but own more than 0.5 acres is no more than that of 0.5 acres of average quality land. Since the PnK data set has a variable reporting the value of owned land this suggestion is readily controverted (see discussion above for the appropriate specification of this variable and a critique of Pitt's, 1999 justification).

Kernel estimates do not use Chemin 3 [14]

They do.

Comparison of female and male treatment effects [15]

MP points out that in making or estimations for female and male borrowing only we include members of the opposite sex among the treated. This is true and arises because there are some households in which there are both female and male borrowers (71 out of 554 and 343 households in which there are female and male borrowers). Our code failed to exclude males/females from the comparison. We explore the effects of correcting this elsewhere.

Choice based sampling [16]

The literature is unclear with regard to accommodating sampling weights in the context of matching. Leuven and Sianesi (2003) suggest that they should not be used in the propensity score estimation but they can be used in the impact estimation. Stata help for psmatch2 recommends to investigate the balancing of the independent variables in order to reach a conclusion on whether sampling weights should be used or not. Hence, we re-ran the analysis with and without sampling weights and investigated the balancing properties of the independent variables. We find that using the sampling weights after PSM made no substantive difference to estimated impacts. Chemin did not use sampling weights in his estimations as he did not have the required weights Stata file which we obtained from Roodman.

The issue of sampling weights should have been discussed but including them or not does not make any substantive differences.

We do not use village-level fixed effects [17]

This is true, and we have no explanation other than failure to read Chemin carefully enough or to think of this for ourselves. Our failures and frustrations with replication of the earlier parts of Chemin can be offered as only partial mitigation.

Do all these issues matter [18]

MP suggests that whether all this matters for estimating the impact of MF using PSM can be assessed from his propensity score matching estimates using the data set he posted in early 2011, and the logit model specifications used in PnK. Apart from repeating the claim that this is relevant to "the impact of microfinance" (MPa:14) unqualified by the possibility that it might only apply to these data, from the particular temporal and geographic domain of their production, it is not clear that the

results reported do credibly support the claim that MP makes. Apart from a number of issues we discussed in the main text, the main reason we believe these results do not provide credible support for the headline PnK claim is that MP fails to conduct SA to assess the vulnerability of the estimated impacts to “hidden bias” (Rosenbaum (2002). Furthermore, correcting for the mistakes we made as suggested by MP, and others, we often find that there is substantive difference in our overall conclusion. Fortunately, perhaps, for us, we came to the right decision albeit for the wrong (or ill-founded) reasons.

Appendix 2: Replication of Pitt, 1999, Tables 1-3

13ARr contains four variables of interest including a categorical variable (halaid) that defines the values in halab (land before access to MF) and halaa (and at time of survey); halab is missing for non-borrowing households (i.e. non-borrowing households in treatment villages and all households in control villages. There are 6 categories in halaid, with 1-5 for various categories of land in decimals (100^{th} of an acre), and category 6 is the total value of land[62]. Thus we do not have values of each category of land separately. Moreover, the proportions of land of different categories We computed various definitions of area of cultivable land, and value of land, including based on predictions from regressions of total value of land on areas/proportions of land of different categories (with and without thana/village dummies). We also varied the definition of MFI membership, focussing on having a loan at the time of the first round of the survey but including membership (there are some who are members but had not taken out any loan).

However, this line of argument is not carried through. Firstly, there are a number of problems in identifying the variables used in Pitt's analysis of land values, and his analysis is not documented by Pitt. Second, while it is possible to match quite closely the results of Pitt (1999) and so infer the variables and specifications probably used there, the analysis Pitt (1999) presents does not address a number of questions one can raise about the finding that mis-targeted borrowers do indeed have land of significantly lower value than others. Thirdly, since Pitt (1999) refers to an extended definition of eligibility (less than 0.5 acres or assets to a value less than the value of 1 acre of medium quality land, one would expect an analysis which sets the non-target households defined in this way rather than the 0.5 acre cut off specified. Specifying non-eligibility in this way will categorise some non-borrowing households with more than 0.5 acres as eligible (not eligible) who are categorized as eligible under the < 0.5 acres rule imposed throughout the PnK oeuvre. We construct an alternative eligibility criteria based on the value of land[63] and of non-land assets.[64]

Sources of land value information

There are two possible sources in the data, files 06AR[1/2/3], and 13AR[1/2/3]. The former appears in the survey section on farming and livestock, and the latter in that on assets. The .pdf file of the

---

[62] 1 "homestead land" 2 "pond/orchard/bamboo" 3 "Own cultivation" 4 "Sharecrop out" 5 "Own land other" 6 "total value of land"

[63] There is a problem with using a land value criterion because there is a clear difference in value between irrigated and non-irrigated land, and between thana; although one could construct a non-irrigated land equivalent using the relative prices of irrigated and non-irrigated land, there is no way of translating this into the eligibility criterion where it is not specified whether it is irrigated or non-irrigated land that is the criterion. Furthermore, both irrigated and non-irrigated land vary in value by other characteristics (including quality of irrigation). We assume that the land component is medium quality unirrigated land, where median quality is the coefficient of the preferred regression on land value on areas of irrigated and non-irrigated land.

[64] Unfortunately there are no variables reflecting value of non-land assets prior to getting loans from MFIs. Hence, while we construct mis-target variables using non-land assets, we also construct mis-target indexes using the value of land prior to joining the MFI. Thus, a household can be considered mis-targeted if it had a value of land more than the value of 50 decimals at the median, mean or upper quartile unit value of land prior to joining.

survey forms (hhsurvey91Quest.pdf) seems to be missing some relevant page[s] for the former[65]. The pages for section 13 seem to be present if not easy to read. The codebook (codes9192.pdf) provides some information on the values in 06AR[r] and 13AR[r].

File	variables
06AR[r]	land owned near/far, irrigated/un-irrigated/unusable; total owned; value of non-irrigated, irrigated, and total[66] land; fixed rented out; share-cropped out; fixed rented in (I/NI); share-cropped in (I/NI); Total operational (I/NI/T); about 700 households report no land in this file raising the question of how to treat these missing values. We presume they are zeros
13AR[r]	land assets code (homestead, pond/orchard/bamboo, own cultivation, sharecrop out, other, total value; areas "before/after" by acquisition (purchase, inheritance, gift, dowry, other); after; before; 1765 households report a value land at the time of the survey (halaa) in this file (i.e. no missing values).

In principle the values in these two files should agree in that total "own cultivation" in 13AR[r] should match the net cultivated land (i.e. total owned less fixed/share-rented out); however, about half of all households report no values of variables in 06ARr files, presumably because, even if they own some land, they do not have cultivated land.

This is unfortunate as it means that it is not possible to identify the value of cultivable land owned by a household. The data in 13AR[r] files contain a single row for the value of land of each household; 5 categories of land are reported (1 "homestead land" 2 "pond/orchard/bamboo" 3 "Own cultivation" 4 "Sharecrop out" 5 "Own land other" 6 "total value of land"). Presumably, although this is not explicitly stated, the value of the category "Total Value" refers to all land reported rather than only cultivated land. While it is possible that the value of non-cultivable land is very stable, it is possible that variations in the value of non-cultivable land with overall assets is similar to that of the value of cultivable land (i.e. that the poorer have non-cultivable land of lower unit values as well as cultivable land of lower unit value).

As noted above, dropping non-borrowing households with more than 0.5 acres using household owned land as one of the covariates in the propensity score estimation might bias upwards the impact estimate because only households with lower land values (and maybe lower outcome variables) can be matched to borrowers. This raises two questions; firstly whether mis-targeting by the extended criterion reported by Hossain (1988) was in fact used by MFIs at the time, and secondly, whether matching by other criteria such as the value of land, land cultivated, or value of assets might be more appropriate.

Since the PKexp.dta file does not include either cultivated land or value of land or assets we can only test these ideas using our own data constructions. In what follows we use wave 1 files only.

[65] Thus page 7 is titled "SECTION 6 FARMING AND LIVESTOCK PART A. LANDHOLDING AND TENURE (cont.) and the first row is numbered 13.

[66] It is not clear whether the variables in 06AR only refer to cultivable land.

Table 8: replication of Pitt, 1999, Table 1				
Panel A: Cultivable land Owned by participation prior to joining MFI programme				
	BRAC	BRDB	GB	Total
Mean	34.19	32.94	35.26	34.24
Median	0	0	0	0
Num > 50 decimals	46	45	66	157
|% > 50 decimals	16.14	14.61	21.15	17.35
Total	285	308	312	905
Panel B: All land Owned prior to joining MFI programme				
Mean	47.09	44.31	52.12	48.06
Median	10	10	16	11
Num > 50 decimals	59	56	88	203
|% > 50 decimals	20.70	18.18	28.20	22.43
Total	285	308	312	905

Table 9: Replication of Pitt, 1999, Table 2				
Panel A: Cultivable land owned by participation at time of survey				
	BRAC	BRDB	GB	Total
Mean	41.50	35.95	42.05	40.34
Median	0	0	4	0
Num > 50 decimals	54	51	74	179
|% > 50 decimals	18.95	16.56	23.72	19.78
Total	285	308	312	905
Panel B: All Land by participation at time of survey				
Mean	54.94	48.19	59.57	54.82
Median	10	13	21	15
Num > 50 decimals	67	66	94	227
|% > 50 decimals	23.51	21.43	30.13	25.08
Total	285	308	312	905

Value of owned land

Pitt (1999) provides an analysis of the unit values of land which we reconstruct here. Our descriptive statistics (Table 8 & Table 9) correspond closely with those of Pitt (1999) but Table 10 is not so close to Pitt, 1999, Table 3, although not far off.

Table 10: Determinants of Unit Land Values (replication of Pitt, 1999, Table 3, plus column 6)

	Taka per decimal all land					
	(1)	(2)	(3)	(4)	(5)	(6)
Land before	-2.626***	-2.412***	-2.110***	-2.041***	-1.819***	-0.627
	(-5.60)	(-4.97)	(-5.40)	(-5.04)	(-5.45)	(-1.53)
Land before ^2	0.0005***	0.0004***	0.0004***	0.0004***	0.0003***	0.0001
	(4.61)	(4.13)	(4.64)	(4.40)	(4.63)	(1.48)
Mis-target	-632.2	-894.9*	-630.7*	-712.4*	-745.4**	-948.7***
	(-1.92)	(-2.46)	(-2.30)	(-2.36)	(-2.69)	(-3.41)
Participant		329.5		107.7	132.2	-117.6
		(1.67)		(0.65)	(0.87)	(-0.74)
Non-target						-789.1***
						(-4.98)
_cons	1935.4***	1831.4***	1241.3***	1195.0***	1164.8***	1420.5***
	(20.51)	(16.21)	(3.84)	(3.61)	(3.83)	(4.63)
N	1404	1404	1404	1404	1701	1701
r2	0.027	0.029	0.357	0.358	0.360	0.369
method	OLS	OLS	thana fe	thana fe	thana fe	thana fe
domain	program	program	program	program	all	all
P test nontarg. vs mistarg.						.594
P ΔR^2 inc. participant						0.000
P ΔR^2 inc. Non-target						0.010

Authors calculations: t statistics in parentheses; * p<0.05 ** p<0.01 *** p<0.001.

Appendix 3: Further results for Male borrowers

Table 11: ATT and sensitivity analysis for Male borrowing (treatment villages only) [67]

Panel A								
kernel estimate, bw = 0.08							N households with any female borrowers	
Sample	Method	n	diff	sediff	p <=	't-value'	Un-treated	treated
MP [a]	kernel, bw 0.08	2553	0.0560	0.0204	0.0061	2.7484	27	42
Excluded male credit households[b]	"	1715	0.0434	0.0249	0.0812	1.7460	0	0
Panel B: Sensitivity Analysis								
Nearest neighbour estimates							gamma p>0.10	
Excluded male credit households[b]	n(10)	1715	0.0506	0.0263	0.0551	1.9210	1.25[c]	
"	n(1)	1715	0.0593	0.0340	0.0812	1.7463	1.20[d]	

Notes: a. corrects area owned of hh 32111; b. also includes villages 11 & 143 (see text)
c. estimated using rsens; d. estimated using rbounds

[67] Equivalent to Table 1 in this document. Same specification as MP, i.e. including nontrgth & nadultf. Results dropping households which borrow in rounds 2 and 3 for the first time make no substantive difference (results available in code).

Table 12: PSM Results: Impact and Sensitivity Analysis – Male borrowers (equivalent to Table 7)

Model	Treat - ment variable	Domain[a]	Matching method	Non-target variable[b]	Fixed effects.village covariates FE	covars	N	Difference mean	se	p-value	t-diff	Critical gamma	Confidence levels 95% hi	lo
1	creditmm	treat	kernel	nontrgth	y	n	1682	0.0484	0.0239	0.0431	2.0256	.	.	
2	creditmm	treat	nn10	nontrgth	y	n	1682	0.0302	0.0261	0.2476	1.157	1.15	0.1719	0
3	creditmm	control	kernel	nontrgth	y	n	1601	0.1087	0.0298	0.0003	3.6471	.	.	
4	creditmm	all	kernel	nontrgth	n	y	2958	0.0508	0.0195	0.0093	2.6059	.	.	
5	creditmm	all	nn10	nontrgth	n	y	2958	0.0518	0.0202	0.0105	2.5638	1.25	0.1702	0
6	creditmm	all	kernel	none	n	y	3609	0.0111	0.0187	0.5528	0.5938	.	.	
7	creditmm	all	nn10	none	n	y	3609	0.0168	0.0195	0.3898	0.8604	1.3	0	1
8	creditmm	all	kernel	nontar1	n	y	2943	0.0544	0.0197	0.0059	2.7607	.	.	.
9	creditmm	all	nn10	nontar1	n	y	2943	0.0466	0.0205	0.0232	2.2746	1.25	0.1404	0
10	creditmm	all	kernel	nontar4	n	y	3116	0.0513	0.0188	0.0065	2.7263	.	.	.
11	creditmm	all	nn10	nontar4	n	y	3116	0.0565	0.0195	0.0039	2.8933	1.3	0.2006	0
12	creditmm	target[c]	nn10	nontar1	n	y	2518	0.033	0.0222	0.143	1.4674	1.1	0.1052	0

Notes: credit = all male credit households; mhh1 – households with male credit in round 1

a. treat = treatment villages only; control = all male borrower households AND all households in control villages; all = all households without female borrowing
b. Nontrgth = MP's non-target variable; nontar1 = 1 if non- borrower AND household cultivable land > 50d; nontar4 = non-borrower & total households assets > 100,000Tk.
c. Excludes mis-targeted households.

Appendix 4: Emails about wage rates

-----Original Message-----
From: Maren Duvendack
Sent: 08 July 2012 09:59
To: 'Mark'
Cc: Richard Palmer Jones (DEV)
Subject: RE: wage data

Dear Mark

We still prefer not to share the full data preparation code until High Noon is in print in its final form - what we write in the paper is clearly intended to be acted on only after it appears in print. We think we have made everything available to you that you need to re-construct the analysis in High Noon and the sub-groups paper from the estimation data set, and have responded to your other queries as promptly as is reasonable, consistent with our work schedules.

Let me explain the wage code I sent earlier in more detail with the hope this will clarify the matter:

The "wage" variables in High Noon were constructed as:

gen wageag = wepv + wescv

and

gen wagenonag = wena1v + wena2v

these were later renamed

gen sumnonagri = wagenonag
gen sumagri = wageag

and sumnonagri and sumagri appear in the logit used for psmatch2

This is not how we do it now, but it was how we did it in High Noon; we told you in the email on May 24th that we had cleaned up the code since the paper was submitted to JDS, and it is almost inevitable that there will be errors - hopefully minor - in such a large amount of code. Our previous message explained what happened, and the code included in that message makes it fairly obvious how this error arose.

It is clear that these two variables do not correspond to any meaningful variable. But, recall that in High Noon we were trying to replicate Chemin, 2008, and Chemin, 2008, does not make entirely clear what his variables represented, and we could not (and we have more or less given up trying to) replicate all his descriptives (column (1) of his Table 1). We don't know what his two wage variables mean. We did not have access to the relevant code, and by the time this issue arose we were no

longer in communication with him - he stopped responding to emails and we decided not to burden him further.

What we do it now is:

```
gen wageag = wepdpm * wepv + wescdpm * wescv                    // wage * days
gen wagenonag =  wena1d * wena1v + wena2d * wena2v          // assuming wena*v is pay per day!
```

We then sum over waves 1-3, and over household members to give household wage earnings.

This (the second pair of "gen wage..." statements) is what it was prior to experimentation with the other code (the first pair of "gen wage ... " statements), as shown by the code from our file that we copied to you. The proper code (the second pair) will give slightly different results to those published in High Noon. As far as we can see the results with the "proper" code are not substantively different from those included in High Noon whether we use total household wage income, or wage income per capita (dividing by hhsize).

As far as we know you have asked to comment on High Noon prior to its appearance in print, after it has been reviewed and accepted. This is unusual, if only because you are not the author of the paper we were trying to replicate (e.g Chemin, 2008). However, our paper clearly relates to your and Shahid Khandker's analysis and uses the data collected by the World Bank with you as a PI, and speaks about results which contrast with those of PnK (but are, we hope duly, qualified). Hence, our cooperation up to and beyond the data and code availability protocol of the current AER data availability guidelines prior to appearance in print. A working paper version of our paper has been available for some time (Jan 2011) should you have wished to comment on that.

We hope this explains the wage variables that appear in High Noon, and how they arose, and acknowledges that they are not what we would do now. And we hope this assures you of our appropriate cooperation. We look forward to your comments.

Best
Maren

-----Original Message-----
From: Mark [mailto:mark_pitt@brown.edu]
Sent: 07 July 2012 21:03
To: Maren Duvendack
Cc: Richard Palmer Jones (DEV)
Subject: Re: wage data

Hi Maren:

I do not understand your coding below. I really do not want to bother you with this. If you just send me the code used to prepared the dataset used in estimation, which you say in the paper is

available, I will find the answers that I need by myself as long as the code is complete. Also, as you know, I need to prepare whatever comments that I may write very soon in order to meet the terms of the agreement made with the Editor of the JDS. Thanks.

Best regards,

Mark

--- Mark M. Pitt
--- Research Professor of Population Studies
--- Emeritus Professor of Economics

On 7/6/2012 7:33 AM, Maren Duvendack wrote:
> Dear Mark
>
> In High Noon we used
>
> ....
> gen wageag = wepv + wescv // is this wages + wages? wepdpm * wepv + wescdpm * wescv // wage * days per month in agriculture
> * calculate agricultural wage
> egen sumagri = rowtotal(wepv wescv)
> label var sumagri "Agricultural wage earnings"
> ...
> and
> ...
> gen wagenonag = wena1v + wena2v // wena1d * wena1v + wena2d * wena2v // assuming wena*v is pay per day!
> * calculate non-agricultural wage
> egen sumnonagri = rowtotal(wena1v wena2v)
> label var sumnonagri "Non-agricultural wage"
> ...
>
> and then collapsed (sum) over individuals within households, and then over waves within households.
>
> Later we drop sumnonagri and sumagri and
>
> gen sumnonagri = wagenonag
> gen sumagri = wageag
>
> This is the source of the wage variables in High Noon.
>

> We now do this differently; as you will notice the code for wageag and wagenonag were originally different. Specifically we now remove the inserted expression [wepv + wescv, and wena1v + wena2v] and reinstate the originals [wepdpm * wepv + wescdpm * wescv, wena1d * wena1v + wena2d * wena2v] , and note that this reflects wage earnings, not wage rates, which of course "wepv + wescv " and "wena1v + wena2v" never did. We think it was only later that we were convinced that wepv and wescv, wena1/2v were wage rates (per day) rather than wage earnings (for the reference period).
>
> We are not entirely clear how this substitution came about, but suspect it was because we could not match Chemin's descriptives for agricultural and non-agricultural wages, and were experimenting with various more or less plausible alternatives, and this particular formulation was left un-corrected in the code until we conducted the revisions which we have already advised you of.
>
> Our revised specification (using wage earnings) does not seem to affect the PSM results substantively.
>
> Best
> Maren
>
> -----Original Message-----
> From: Mark [mailto:mark_pitt@brown.edu]
> Sent: 05 July 2012 17:16
> To: Maren Duvendack
> Subject: wage data
>
> Hi Maren:
>
> You and Palmer-Jones uses the variables "Non-agricultural wage" and "Agricultural wage" in your Table 1. Can you tell me how you calculated these variables?
> Thanks.
>
> Best regards,
>
> Mark
>
> --
> ---
> --- Mark M. Pitt
> --- Research Professor of Population Studies
> --- Emeritus Professor of Economics