Journal of Development Studies
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/fjds20

High Noon for Microfinance Impact Evaluations: Re-investigating the Evidence from Bangladesh

Maren Duvendack [a] & Richard Palmer-Jones [a]
[a] School of International Development, University of East Anglia, Norwich, UK

Available online: 27 Apr 2012

To cite this article: Maren Duvendack & Richard Palmer-Jones (2012): High Noon for Microfinance Impact Evaluations: Re-investigating the Evidence from Bangladesh, Journal of Development Studies, DOI:10.1080/00220388.2011.646989

To link to this article: http://dx.doi.org/10.1080/00220388.2011.646989

Journal of Development Studies,
2012, 1–17, i*First* article

Routledge
Taylor & Francis Group

High Noon for Microfinance Impact Evaluations: Re-investigating the Evidence from Bangladesh

MAREN DUVENDACK & RICHARD PALMER-JONES
School of International Development, University of East Anglia, Norwich, UK

Final version received July 2011

ABSTRACT *Recently, microfinance has come under increasing criticism raising questions of the validity of iconic studies which have justified it, such as Pitt and Khandker. Chemin applied propensity score matching to the Pitt and Khandker data, finding different impacts, but does not disaggregate by gender of borrower. We first replicate Chemin and extend his analysis in two ways. We test the robustness of propensity score matching results to selection on unobservables using sensitivity analysis, and we investigate propensity score matching estimates of impacts by gender of borrowers. The mainly insignificant impacts of microfinance differ greatly by gender of borrower, but are all vulnerable to selection on unobservables. We are therefore not convinced that the relationships between microfinance and outcomes are causal with these data.*

I. Introduction

Replication and reproduction[1] are important features of practise in the natural sciences and are desirable also in the social sciences including economics (Kane, 1984; King, 1995; Hamermesh, 2007). Economics papers applying complex statistical methods can, apparently quite frequently, have errors of both variable construction (data manipulation) and statistical estimation (Dewald et al., 1986; McCullough et al., 2006; McCullough et al., 2008). The application of different methods can lead to different conclusions with practical relevance (Dewald et al., 1986; McCullough et al., 2006; McCullough et al., 2008). However, there are few rewards for replication in the social sciences; relatively few are conducted and fewer published (Anderson et al., 2008). Doubts can be cast on the creativity of those who replicate, and also on their motivation, which might include casting doubt on the integrity or ability of the original authors. Yet the returns to finding problems in papers could be high for society; policies which are legitimated in large part by iconic studies which are subsequently shown to not to lead robustly to the conclusions for which they are known, could lead to different research or policy conclusions with high social benefits. In this article we present evidence that undermines an, if not the, iconic study which legitimated for much of the past two decades the belief that microfinance (MF) is beneficent for the poor, especially when targeted on women.

Correspondence Address: Maren Duvendack, University of East Anglia, School of International Development, Norwich, NR4 7TJ, UK. Email: m.duvendack@uea.ac.uk
An Online Appendix is available for this article which can be accessed via the online version of the journal available at http://dx.doi.org/10.1080/00220388.2011.646989

ISSN 0022-0388 Print/1743-9140 Online/12/000001-17 © 2012 Taylor & Francis
http://dx.doi.org/10.1080/00220388.2011.646989

The concept of microcredit was first introduced in Bangladesh by Nobel Peace Prize winner Muhammad Yunus. Professor Yunus started Grameen Bank more than 30 years ago aiming to reduce poverty by providing small loans to the countries' rural poor (Yunus, 1999). It is argued that microfinance enables the poor to access credit, providing them access to remunerative activities and relieving them of onerous debts, and has especially beneficial effects when targeted on women (Khandker, 1998, 2000). These arguments were for much of the last decade and a half most authoritatively supported by Pitt and Khandker (1998 – henceforth PnK). However, despite the huge expansion and popularity of microfinance, it has recently been argued that there is little convincing evidence that microfinance programmes have positive impacts (Armendáriz de Aghion and Morduch, 2005, 2010; Goldberg, 2005; Roy, 2010; Bateman, 2010; Stewart et al., 2010; Duvendack et al., 2011) in part because the PnK study came under intense scrutiny particularly from Morduch (1998), and Roodman and Morduch (2009 – henceforth RnM). However, these criticisms have seemingly been refuted (Pitt, 1999, 2011a, 2011b).

While not engaging in debate with PnK or Morduch, Chemin (2008 – henceforth Chemin), applied propensity score matching (PSM) to his reconstruction of the data and finds 'positive, but lower than previously thought' (Chemin, 2008: 463) impacts compared to PnK; Chemin does not explain whether the differences he finds are due to differences in data construction or analytical method, and does not differentiate between female and male borrowers. While RnM had considerable difficulty in reconstructing the variables and analysis reported in PnK and Khandker (2005), evidently expending much time and effort,[2] when corrected their analysis replicates PnK successfully (Pitt, 2011a, 2011b). Hence Chemin is the only outstanding reproduction of PnK that casts some doubt on their highly influential findings.

In this article we set out to replicate Chemin's results. Neither PnK nor Chemin provide a complete set of data and code that would allow reproduction of their published results. We independently and successfully replicate the data constructions of RnM, also expending much effort; we recalculate and extend Chemin's PSM findings, subject our results to sensitivity analysis, and differentiate between male and female borrowing. We cannot replicate Chemin's empirical results, and find different and mainly insignificant effects of MF on the outcome variables; the sensitivity analysis suggests that those of our results which are significant are highly vulnerable to unobservables. We fail to confirm PnK's original findings of beneficent outcomes caused by MF using PSM with these data. Consequently our study fails to lend support to the idea that causal relationships can be established between MF and wellbeing with these data. Replication is important in order to avoid errors in data construction and to triangulate complex but fragile analyses.

II. Microfinance Evaluations

A number of putatively rigorous studies suggest social and economic benefits from microfinance (Hulme and Mosley, 1996; Pitt and Khandker, 1998; Khandker, 1998, 2005; Coleman, 1999; Rutherford, 2001; Morduch and Haley, 2002). However, Dichter and Harper (2007), Roy (2010), Bateman (2010), and Bateman and Chang (2009) argue that microfinance is neither always beneficial nor rigorously demonstrated.

Many of the early microfinance impact evaluations failed to address the problem of selection bias (Sebstad and Chen, 1996; Gaile and Foster, 1996). Underlying these selection processes[3] are observable characteristics such as age, education, work experience and so on, and unobservable attributes such as entrepreneurial skills, organisational abilities, willingness to take risks, and so forth, although Armendáriz de Aghion and Morduch (2010: 272) argue that there is a high correlation between entrepreneurial skills, age and microfinance participation. Coleman (1999) lists further unobservable characteristics such as access to social networks and business skills that tend to increase the likelihood of individuals participating in microfinance.

Estimates of impacts of MF may be confounded by selection on unobservables, or 'hidden bias' (Rosenbaum, 2002), due to unobserved variables that influence treatment allocation as well as potential outcomes (Becker and Caliendo, 2007). If research methodologies do not condition

on these unobservables by, for example, randomisation, or use of convincing instrumental variables (IV) estimations, these lacunae undermine impact estimates (Heckman, 1979). A few studies have addressed this problem (for example Hulme and Mosley, 1996; Pitt and Khandker, 1998), but are not uncontested.[4] Furthermore, among other limitations, average impacts can disguise considerable heterogeneity; thus, an average impact that is indistinguishable from zero can arise because some significantly benefit while others receive negative outcomes (Heckman et al., 1999). MF impact evaluations may be vulnerable to criticisms if they do not adequately investigate impacts by subgroups.[5]

The authoritative PnK study on three microfinance programmes in Bangladesh attempts to account for participant selection and programme placement[6] biases (Pitt and Khandker, 1998; Coleman, 1999) using an IV approach, and Khandker (2005 – henceforth Khandker) adds data on the same households surveyed in 1998/1999 to construct a panel, putatively overcoming the problems for evaluation posed by participant selection.

A number of studies have attempted to replicate the findings of the original PnK study, and Khandker. Morduch contested PnK but has seemingly been refuted by Pitt (1999 – henceforth Pitt).[7] RnM replicated PnK, Morduch, Pitt, and Khandker, producing variables which in some cases differ significantly from their equivalents in original papers (see Online Appendix 1). Nevertheless, when using different estimating software and correctly specifying the PnK model, RnM support the empirical findings of PnK, but dispute causality. Doubts about PnK arise in part because of the quasi-experimental design and doubts that the econometric methodology establishes causality (Roodman and Morduch, 2009). Nevertheless RnM conclude that

> nothing [in their work] contradicts ... [the] idea .... that it [MF] is effective in reducing poverty generally, that this is especially so when women do the borrowing, and that the extremely poor benefit most [and] ... helps families smooth their expenditures, lessening the pinch of hunger and need in lean times. (Roodman and Morduch, 2009: 39–40)

Two systematic reviews (SR) (Stewart et al., 2010; Duvendack et al., 2011), of the impacts of MF, including two recent randomised control trials (RCTs), are more sceptical; using criteria generally applied in the medical literature, these two SRs pointed to the weak research designs of the most valid evaluations of MF, as well as sometimes harmful effects of MF borrowing (Stewart et al., 2010; Duvendack et al., 2011). The latter in particular argue that it is moot whether the appropriate response to the PnK study should have been to pursue more robust evidence of such impacts, or to research alternative means to the beneficent ends purportedly attached to MF.

Moreover, the original RnM paper has been contested by Pitt (2011a, 2011b); Roodman has rejected Pitt's interpretation.[8] Since we are using a different method of analysis (PSM rather than limited-information maximum likelihood ((LIML)/IV) we do not engage further in this debate other than to note that doubts about IV methods are common (Leamer, 2010; Deaton, 2010), and PSM is a possible alternative (DiPrete and Gangl, 2004).

Another approach to replication is to apply different estimation techniques to the data, motivated by doubts about the estimation strategy (rather than, or as well as the data constructions). Chemin applies PSM to his construction of the variables, and does not engage in critique of PnK. PSM (Rosenbaum and Rubin, 1983, 1984) has become a popular technique in development economics in recent years. It attempts to approximate the research design of a RCT by matching participants to non-participants drawn from a suitable population on the basis of a set of common covariates by a predicted probability of programme participation, or the 'propensity score' (Ravallion, 2001; Caliendo and Kopeinig, 2005, 2008). Units which are clearly not similar to the treated group are dropped from the calculation of impacts; the treatment effect is then estimated by comparing the mean outcomes of the participants and their matches (Ravallion, 2001). This method can account for selection bias due to observable characteristics, its drawback, however, is that bias due to selection on unobservables remains (Smith and Todd, 2005). Sensitivity analysis of PSM results can identify the vulnerability of the estimated impact to

unobservables (Rosenbaum, 2002) and is good practice in PSM studies (Ichino et al., 2006; Nannicini, 2007).

This article attempts to replicate Chemin's study which remains the only currently credible evaluation of microfinance using these data that fails to confirm the PnK results;[9] it extends Chemin's analysis to test the claim by PnK that borrowing from microfinance institutions (MFIs) by women is more beneficial, and uses sensitivity analysis to test the extent to which the PSM results are vulnerable to unobservables. The paper proceeds as follows: we outline the particularities in the PnK research design, briefly discuss the challenges of replication and (re-)construction of appropriate variables with the PnK data, apply PSM to (our reconstruction of) the data, and investigate effects of the gender of the borrower on microfinance impact; we apply sensitivity analysis to the matching results to draw conclusions as to the robustness and limitations of PSM in this context.

We find differences in the descriptive statistics from those reported in Chemin (but only minor differences from RnM – see Table A1 in the Online Appendix); our PSM results not surprisingly also differ from Chemin's (see Table A3 in the Online Appendix for an overview of headline findings). We find negative as well as positive, both often statistically insignificant, average microcredit impacts, and we cannot show that women have an obvious advantage over men as borrowers. Sensitivity analysis suggests that even the statistically significant impacts are highly vulnerable to unobservables, implying that it would be unwise to conclude that any association between MF borrowing and impact is causal.[10]

III. The Impact of Microfinance in Bangladesh

In this section we make the case that the quasi-experimental research design used by PnK means that their identification strategy is open to doubt. PnK use data from a World Bank funded study which conducted a survey in three waves in 1991–1992[11] on three leading microfinance group-lending programmes in Bangladesh, the Grameen Bank (GB), Bangladesh Rural Advancement Committee (BRAC) and Bangladesh Rural Development Board (BRDB) (Pitt and Khandker, 1998: 959). PnK sampled households from villages with at least one microfinance programme (treatment villages) and from villages without these (or any other) MFIs (control or non-programme villages). These institutions apply eligibility criteria (nominally owning/cultivating less than 0.5 acres at the time of recruitment into the MFI programme[12]); this eligibility criterion is PnK's identification strategy assuming that land operated/cultivated was exogenous.

The survey was conducted in 87 randomly selected villages from 29 thanas,[13] yielding a sample of 1798 households of which 1538 were target households (eligible in treatment or control villages) and 260 were non-target (not eligible) households (Pitt and Khandker, 1998: 974). According to PnK (1998: 974), out of those 1538 households, 905 effectively participated in microfinance (59%). Three survey waves (R1–3) were timed to account for seasonal variations (Pitt, 2000: 28–29). The study measures the impact of microfinance participation by gender of borrower on labour supply, school enrolment, expenditure per capita and its variation between rounds, and women's non-land asset ownership. PnK find that microcredit has significant positive impacts on many of these indicators and find larger positive impacts when women are involved in borrowing.

PnK adopt an estimation strategy for assessing the impact of microfinance participation involving comparisons of 'treated' and 'non-treated' households in 'treated' villages and 'non-treated' households in 'non-treated' (control) villages, using the eligibility criterion described above. PnK define all participants as eligible (that is, they are 'de facto' eligible). They sample treatment and control villages containing both non-target/landed and target/landless households. PnK's (ideal) identification strategy is represented in Figure 1.

PnK suggest assessing impact at the discontinuity between participant (eligible) and non-participant (not eligible) households in treatment and control villages; that is, comparing the discontinuity at the boundary between group B and A in control villages, and between group D to C in treatment villages (Figure 1). The difference between these two sets of comparisons is estimated by applying village-level fixed-effects to account for unobserved differences between

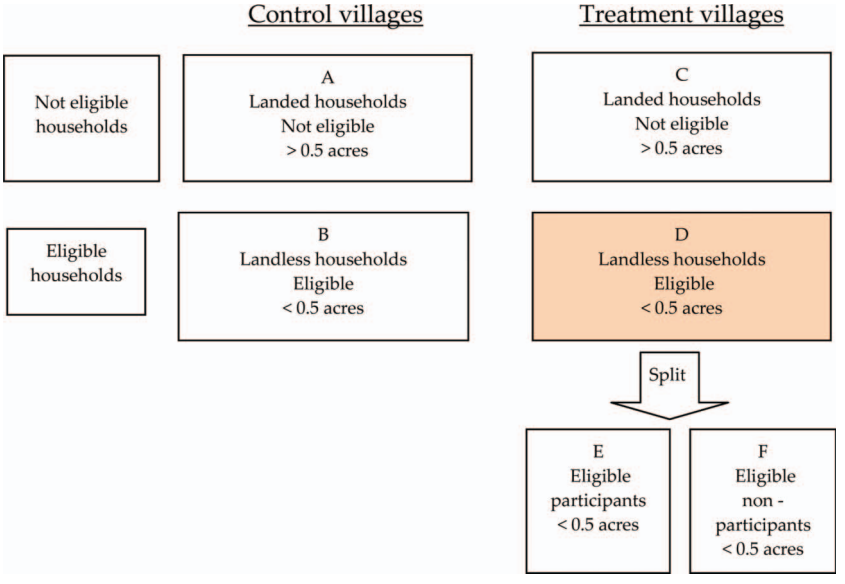

Figure 1. Intended identification strategy.
Source: Authors' illustration based on Morduch and Chemin. *Notes*: This diagram ignores that the eligibility criterion was not strictly (literally) enforced.

treatment and control villages. However, the eligibility criterion was not strictly enforced (Morduch, 1998).[14] Nevertheless, Pitt (1999, 2011a, 2011b) maintains that this does not undermine the results which are also obtained when units which borrow from MFIs and are cultivating more 0.5 acres are dropped from the estimation sample.

Chemin (2008: 465) bypasses the PnK debate arguing that the issue of eligibility is avoided by PSM. In addition to solving the eligibility issue, Chemin further claims that

> matching takes into account non-random programme placement by comparing treated individuals with the 'same' non-treated individuals in control villages. These 'same' non-treated individuals in control villages would have participated in microfinance had they had access to microfinance. (Chemin, 2008: 465)

However, it is not clear whether PSM can achieve all this, since it cannot condition on unobservables, or whether, indeed, it is an appropriate technique for solving the particular problems in the PnK data set. We use Chemin's study to demonstrate some of the challenges of replication as well as some limitations of PSM.

IV. Replicating Chemin

Our focus is on replication of Chemin but to do this we also triangulate our variable construction with PnK and RnM. A complete set of our Stata code is available from the authors to run with the data that can be downloaded from the World Bank together with additional data we can supply, and instructions on how to organise the data.

The first step in a replication from the same raw data is to recreate the variables used in the analysis; this is not a trivial exercise when using multi-topic surveys. Most of the PnK data, including questionnaires and variable codes are (at the time of writing this article) available on the World Bank website[15] or have been obtained[16] by Roodman[17] and are publicly available. Nevertheless replication remains a challenge, particularly because the survey forms and variable descriptions are problematic. We have compared our data constructions with RnM's data,

observation by observation for the key variables finding negligible differences in most cases; Table A1 in the Online Appendix reports the descriptive statistics provided by PnK, RnM, as well as Chemin, and our replications.[18] The discrepancies we have with the descriptive statistics published by RnM are minor; we have other differences with the RnM data set but these reflect differences in interpretation of some variables rather than differences or errors in variable construction (these discrepancies are detailed in Table A2 of the Online Appendix). Chemin did not provide either complete code to construct his data set or the data set he used,[19] so that differences in the descriptive statistics between our and Chemin's variables cannot be explained. Details of his variable definitions (or code embodying these definitions) are not available.[20] We invested a large amount of time and effort in understanding his incomplete code but we could neither re-construct his variables nor replicate his findings, and used our own definitions. We assume that our data constructions which triangulate with those of RnM are defensible, and we prefer them because they can be verified from both our and RnM's code. However, as a result of these differences in variables we cannot attribute differences between our and Chemin's PSM results to either data constructions or estimation methods.

Table 1 reproduces Chemin's original logit results and our estimates using the same specifications. Chemin uses microfinance participation (not eligibility) as a dependent variable which assumes the value of 1 if the individual participates in microcredit and 0 if the individual does not. Chemin explains that specification 1 is used by PnK; specification 2 contains the same variables as specification 1 as well as additional control variables which Chemin argues might be of use for predicting microfinance participation. Specification 3 is Chemin's preferred model used in his PSM analysis.

Table 1 shows that our logit coefficients for sex, age and age of household head in specification 1 are reasonably similar to Chemin's. Many of the remaining logit coefficients, however, differ. A similar pattern can be found in specification 2. According to Chemin the additional control variables were all insignificant (Chemin, 2008: 471); however, we found that 10 of these were in fact significant (see Table 1 and its notes). The pseudo R-squared in our replication for logit specification 1 is higher than reported by Chemin and lower for specifications 2 and 3; this is presumably due to differences in the data sets and variable constructions used by Chemin and ourselves. The number of observations differs as well for reasons we cannot explain.

Figure 2 shows the distribution of propensity scores produced using Chemin's specification 3; this figure is very similar to Figures 2 and 3 in Chemin (2008: 474–475). They indicate that there is limited overlap between participants and non-participants in treatment and control villages. The common support region is rather narrow and hence few good control group cases are suitable matches.[21] Generally speaking, the lack of overlap implies that the common support assumption is not fully satisfied, and consequently the question must be asked whether PSM is suitable with these data.

We turn now to the impact estimates, following Chemin's preferred matching algorithms (Chemin, 2008: 475). Compared to non-participants in treatment villages, Chemin's PSM results indicate that microcredit has a significant negative impact on participants' log of per capita expenditure (Table 2, row 1). Participants spend 3.5 per cent to 4.6 per cent less per capita than non-participants. This is surprising and contrary to the expectation that microcredit participation has positive impacts. It appears that participants in treatment villages are not necessarily better off than matched non-participants in treatment villages. The replication results presented in Table 2, row 3 also provide negative but smaller impact estimates across both matching algorithms. There are discrepancies in the kernel matching results, Table 2, row 3 provides insignificant negative estimates while Chemin's kernel estimates are all significant and negative (Table 2, row 1).

Turning now to comparison with matched samples from control villages; Chemin finds that microcredit has a significantly positive impact across most matching algorithms (Table 2, row 2). The replication results presented in Table 2, row 4 differ substantially from Chemin's. We find

Table 1. Replication of Chemin's logit predicting the probability of microfinance participation

Independent variables	Spec. 1		Spec. 2		Spec. 3	
	Chemin	Authors	Chemin	Authors	Chemin	Authors
Highest grade completed	0.041	−0.058**	0.024	−0.094***		
	0.03	*0.046*	*0.036*	*0.005*		
Sex (male = 1)	−0.886***	−0.590***	−1.515***	−0.805***	−1.136***	−0.773***
	0.123	*0.000*	*0.182*	*0.000*	*0.128*	*0.000*
Age (years)	0.051***	0.050***	1.224***	0.519***	1.065***	0.559***
	0.004	*0.000*	*0.269*	*0.000*	*0.159*	*0.000*
Age household head (years)	−0.046***	−0.027***	−0.035***	−0.004	−0.014**	−0.003
	0.006	*0.000*	*0.009*	*0.332*	*0.006*	*0.454*
Number adult male in household	1.951	−0.296***	2.854*	0.101	0.832***	0.011
	1.268	*0.000*	*1.562*	*0.157*	*0.308*	*0.873*
Landholdings HH head parents	0.137	−0.079	0.094	−0.070		
	0.14	*0.307*	*0.147*	*0.436*		
Landholdings HH head brothers	0.019	−0.097***	−0.023	−0.062		
	0.065	*0.007*	*0.068*	*0.132*		
Education					0.336***	0.000
					0.113	*0.196*
Savings			0.0002***	0.000***	0.0002***	0.354***
			0.0004	*0.002*	*0.00003*	*0.000*
Have non-farm enterprise (yes = 1)			0.763***	0.319***	0.630***	−0.000
			0.173	*0.000*	*0.111*	*0.129*
Livestock value			0.0000397	−0.000	0.00005***	−0.122***
			0.00003	*0.520*	*0.00002*	*0.000*
Household size			−0.117***	−0.075***	−0.147***	−0.001
			0.041	*0.006*	*0.028*	*0.124*
Non-agricultural wage (in Taka)			−0.002	−0.001	−0.006*	−0.000
			0.004	*0.257*	*0.003*	*0.992*
Agricultural wage (in Taka)			0.013**	−0.000	0.010**	−0.009***
			0.007	*0.612*	*0.005*	*0.000*
Age squared			−0.033***	−0.008***	−0.028***	0.000***
			0.01	*0.000*	*0.006*	*0.000*

(*continued*)

Table 1. (*Continued*)

Independent variables	Spec. 1		Spec. 2		Spec. 3	
	Chemin	Authors	Chemin	Authors	Chemin	Authors
Age power of 4			−1.73E-6*	0.000***	−1.16E-6***	−0.773***
			0.000	*0.000*	*0.0001*	*0.000*
Number adult female in household				−0.201**		
				0.013		
Agricultural income (in Taka)				0.000*		
				0.069		
Household land (in decimals)				−0.003**		
				0.040		
Marital status (married = 1)				0.422***		
				0.002		
Other assets (in Taka)				−0.000***		
				0.000		
Village dummies	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	4215	9397	4205	9397	5037	9397
Pseudo R-squared	0.150	0.163	0.356	0.333	0.331	0.315

Source: Chemin (Table 1: 471) and authors' calculations.
Notes: p-values in italics. * significant at 10 per cent, ** significant at 5 per cent, *** significant at 1 per cent. PnK data across R1–3 downloaded from the World Bank website are used. According to Chemin, specification 1 replicates PnK, specification 2 includes other control variables such as landed assets, equipment assets, transport assets, injuries, change of residence in the last two years, assets, expenses of the non-farming enterprise, agricultural costs, irrigated land, father still alive, marital status, agricultural income, mother's education, irrigated household land, mother still alive, household land, highest grade completed by household head, sex of household head, number of adult females in household, sisters of household head owning land, father's education, revenue of non-farming enterprises, dairy products sales which are all insignificant. Chemin argues that all control variables in specification 2 were insignificant. However, our replication results differ and 10 out of those 23 control variables are significant, namely: number of adult females in household, household land, marital status, equipment assets, other assets, sex of household head, agricultural income, landed assets, agricultural costs, and father's education. The remaining variables such as revenue of non-farming enterprises, expenses of the non-farming enterprise, irrigated land, mother's education, transport assets, injuries, father still alive, irrigated household land, mother still alive, highest grade completed by household head, sisters of household head owning land, dairy products sales were also included and were all insignificant. The variable 'change of residence in the last two years' could not be replicated.

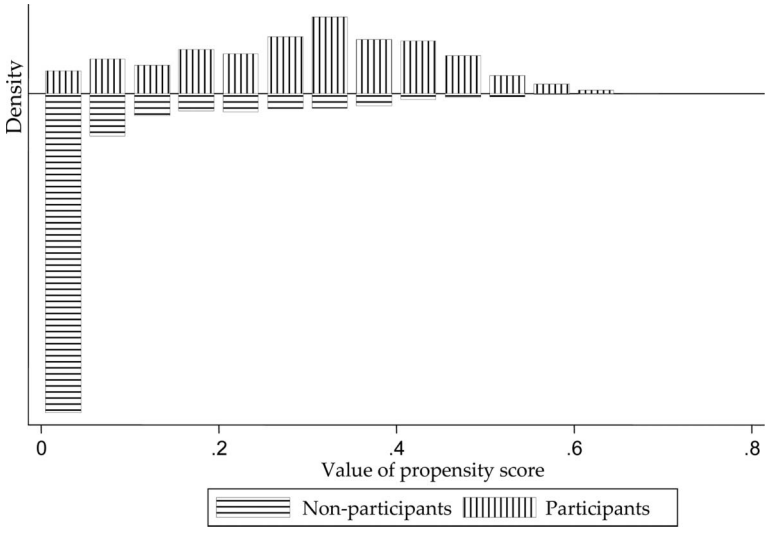

Figure 2. Distribution of propensity scores for participants and non-participants in treatment and control villages.
Source: Authors' calculations.

Table 2. Impact estimates and their replication for log of per capita expenditure (Taka)

		Stratification number of strata			Kernel matching bandwidth		
Row	Control group	20	10	5	0.05	0.02	0.01
Chemin's reported results[a,b]							
1	Non-participants in treatment villages	−0.035*	−0.044*	−0.044*	−0.039*	−0.044*	−0.046*
2	Individuals in control villages	0.028	0.028***	0.028*	0.028***	0.028***	0.028***
Authors' replication of Chemin[c]							
3	Non-participants in treatment villages	−0.001***	−0.003***	−0.010***	−0.007	−0.011	−0.012
4	Individuals in control villages	−0.004***	0.008***	−0.021***	−0.061***	−0.064***	−0.065***

Notes: * significant at 10 per cent, ** significant at 5 per cent, *** significant at 1 per cent. PnK data across R1–3 downloaded from the World Bank website are used. Chemin's specification 3 is used. The results in this table refer to the differences in the mean values between matched samples. *t*-tests before and after matching were employed for all results presented in this table to investigate the differences in the mean values for each covariate X across matched samples; as before, the test provided conclusive results. All results are bootstrapped.
[a]*Source*: Chemin (2008: Table 2, 476).
[b]Chemin's original impact estimates for log of per capita expenditure obtained by matching participants with non-participants in treatment villages and participants with individuals in control villages using Chemin's specification 3.
[c]Replication of Chemin's original impact estimates for log of per capita expenditure obtained by matching participants with non-participants in treatment villages and participants with individuals in control villages using Chemin's specification 3. Authors' calculations.

that participants are worse off than individuals in control villages spending 0.4 per cent to 6.5 per cent less than individuals in control villages – significantly so except for 10 strata (significant and positive) matching.

Table 3 presents the impact estimates for all six outcome variables comparing microfinance participants in treatment villages with non-participants in treatment and control villages combined. Table 3 lists Chemin's original results which suggest that microcredit has a significant and positive impact on male labour supply and girls' school enrolment. All other impact estimates are not significant apart from boys' school enrolment when a kernel bandwidth of 0.05 is used. Those estimates are lower than the ones presented by PnK and this is an important finding. However, Chemin does not assess the impact by gender and fails to re-examine PnK's main claim, namely that microfinance impact is more positive when women are involved in borrowing. No explanation is given for his neglect of this important issue.

Our replication (Table 3) cannot entirely confirm Chemin's results; it suggests that microcredit has a significant and positive impact on the log of women's non-landed assets and a significant and negative impact on the variation of the log of per capita expenditure (for kernel bandwidth 0.05 and 0.02). The effects on both school enrolment variables are largely significant and positive with all remaining results being insignificant.

To summarise the main results so far: Chemin finds no significant impact on the variation of the log of per capita expenditure but the majority of our estimates are significant at 10 per cent. He finds no impact on the log of women's non-landed assets while our replication supports the view that microfinance has a significantly positive impact on the log of women's non-landed assets. Chemin's finding that there is a positive and significant impact on male labour supply

Table 3. Impact estimates for all six outcome variables matching participants to non-participants across treatment and control villages (kernel matching, various bandwidths)

		Chemin			Authors' replication		
Row	Outcome variables	0.05	0.02	0.01	0.05	0.02	0.01
1	Variation of log per capita expenditure (Taka)	−0.008	−0.008	−0.008	−0.012*	−0.012*	−0.012
2	Log per capita expenditure (Taka)	n/a	n/a	n/a	−0.008	−0.011	−0.011
3	Log women non-landed assets (Taka)	0.037	0.037	0.038	0.513***	0.498***	0.489***
4	Female labour supply, aged 16–59 years, hours per month	9.503	9.507	9.521	9.448	8.879	9.104
5	Male labour supply, aged 16–59 years, hours per month	17.001***	16.996***	16.974***	21.256	22.305	20.512
6	Girl school enrolment, aged 5–17 years	0.051***	0.051***	0.052***	0.051**	0.057**	0.060**
7	Boy school enrolment, aged 5–17 years	0.035*	0.035	0.036	0.038	0.044*	0.047*

Notes: * significant at 10 per cent, ** significant at 5 per cent, *** significant at 1 per cent. PnK data across R1–3 downloaded from the World Bank website are used. Chemin's specification 3 is used. The results in this table refer to the differences in the mean values between matched samples. Results are bootstrapped.
Source: Chemin (Table 3: 477) and authors' calculations.

cannot be supported by our analysis. Furthermore, we find positive and significant impacts on boys' school enrolment while he does not. How do these results change when segregated by gender?

Table 4 shows that the effects of female and male borrowing on the log of women's non-landed assets is positive and significant, as in Table 3 but that is where the similarities end. Female borrowing has positive and significant effects on female labour supply and no significant effect on male labour supply, which suggests that a higher value is attached to women's market time and thus home time is substituted with market time. If indeed women spend more time on self-employment activities, one might expect that their daughters' time would be used for household activities to replace their mothers' leading to withdrawal of girls from school (Pitt and Khandker, 1998). However, female borrowing has significantly positive impacts on girls' school enrolment. Male borrowing significantly increases male labour supply while significantly decreasing women's labour supply, indicating that the value of men's market time increases. Also, male borrowing is more likely to have a significant negative impact on the log of per capita expenditure than female borrowing. Overall, it appears that the gender of the borrower makes a difference on the outcomes investigated here.

V. Sensitivity Analysis

Although some significant effects are found using PSM this does not answer the question whether these results are robust to unobservables. Rosenbaum (2002) developed sensitivity analysis to explore the robustness of matching estimates to selection on unobservables (Rosenbaum, 2002). Ichino et al. (2006: 19) argue that 'sensitivity analysis should always accompany the presentation of matching estimates'.

Rosenbaum (2002) invites us to imagine a number Γ (gamma) (≥ 1) which captures the degree of association, of an unobserved characteristic with the outcome and with selection into treatment, required for it (the unobserved characteristic) to explain the observed impact. Γ is the ratio of the odds[22] that the treated have this unobserved characteristic to the odds that the controls have it; a low odds ratio (near to one) indicates that it is not unlikely that such an unobserved variable exists. Cornfield et al. (1959) provide the example of the effect of smoking on lung cancer. In this case, which is now surely without doubt, data from the late 1950s gave a $\Gamma > 5$ as the level of association of an unobserved characteristic with outcome and selection into smoking required to confound the observed impact of smoking on death; such a variable is, it is

Table 4. Impact estimates for all six outcome variables segregated by gender (kernel matching)

	Bandwidth 0.05	
Outcome variables	Female	Male
Variation of log per capita expenditure (Taka)	−0.019**	−0.026***
Log per capita expenditure (Taka)	−0.018	−0.066***
Log women non-landed assets (Taka)	1.314***	0.805***
Female labour supply, aged 16–59 years, hours per month	23.440**	−25.592***
Male labour supply, aged 16–59 years, hours per month	40.860	63.909***
Girl school enrolment, aged 5–17 years	0.069**	0.103***
Boy school enrolment, aged 5–17 years	0.056	0.068***

Notes: * significant at 10 per cent, ** significant at 5 per cent, *** significant at 1 per cent. PnK data across R1–3 downloaded from the World Bank website are used. Chemin's specification 3 is used. The results in this table refer to the differences in the mean values between matched samples. *t*-tests before and after matching were employed for all results presented in this table to investigate the differences in the mean values for each covariate X across matched samples; as before, the test provided conclusive results. Results are bootstrapped.
Source: Authors' calculations.

suggested, highly unlikely not to have been observed, because a variable with such a strong association with both smoking and death from lung cancer would have been hard to overlook.

This approach can be implemented for a continuous outcome variable using the **rbounds** procedure in Stata (DiPrete and Gangl, 2004);[23] **rbounds** uses the results from the nearest neighbour matching estimates to calculate the maximum and minimum p-values from a Wilcoxon sign rank test between treatment and matched pairs, and the Hodges-Lehman point estimates of the impact of the treatment on the outcome variable and their confidence intervals (for a given level of confidence – for example 95%), for a unobserved confounding variables with different values of Γ. A value of Γ that produces a Hodges-Lehman confidence interval that encompasses zero is one that would make the estimated impact not statistically significant at the relevant level of confidence. If the lowest Γ at which this confidence interval encompasses zero is relatively small (say <2[24]) then one may assert that the likelihood of such a characteristic being unobserved is relatively high and therefore that the estimated impact is rather sensitive to the existence of unobservables (DiPrete and Gangl, 2004).

Table 3 shows that the kernel matching estimate with a bandwidth of 0.01 for log of women's non-landed assets is 0.489 and is statistically significant at 1 per cent. This suggests that the log of women's non-landed assets is significantly higher for participating households than for control households. The sensitivity analysis using **rbounds** is reported in Table 5[25] showing that when $\Gamma = 1.15$ the statistical significance of the Wilcoxon signed-rank test upper bound estimate at this level of bias is $p < 0.1727$ and the confidence interval of the estimated impact encompasses zero (-0.091 to 1.407) at this gamma (strictly the estimates become marginally insignificant at the 95% confidence level when $\Gamma = 1.10$). It is not unlikely that a variable with this level of odds of association with both outcome and being treated (being a borrower) relative to controls should be unobserved. This implies that for this outcome variable selection on unobservables could well explain the observed association between taking microfinance and the log of women's non-landed assets. Consequently, we conclude that the statistically significant association between taking microfinance and the log of women's non-landed assets may well be due to unobservables. We have argued elsewhere that in social science studies a critical $\Gamma < 2$ (at $p < 0.05$) surely suggests vulnerability to unobservables if a similar figure is accepted by natural sciences where one might expect unobservables to be less likely (Duvendack and Palmer-Jones, 2011a); see also Guo and Fraser (2010: 318) who write that 'because [gamma=] 1.43 is a small value, we can conclude that the study is very sensitive to hidden bias'.

Table 5. Sensitivity analysis for log of women's non-landed assets for microfinance participants across R1-3

	Significance levels (Wilcoxon signed-rank test)		Hodges-Lehmann point estimates		95% Confidence intervals	
Gamma (Γ)	Lower	Upper	Lower	Upper	Lower	Upper
1	0.0031	0.0031	0.469	0.469	0.048	0.920
1.05	0.0004	0.0174	0.326	0.607	2.7e-07	1.089
1.1	0.0000	0.0650	0.189	0.730	−2.7e-07	1.251
1.15	2.7e-06	0.1727	0.077	0.882	−0.091	1.407
1.2	1.7e-07	0.3453	−2.7e-07	1.031	−0.220	1.560
1.25	8.6e-09	0.5497	−2.7e-07	1.166	−0.325	1.710
1.3	3.8e-10	0.7348	−0.026	1.307	−0.418	1.865

Notes: See note 25. The table shows magnitude of selection on unobservables, range of significance levels, Hodges-Lehmann point estimates and confidence intervals.
Source: Authors' calculations.

Sensitivity analysis was conducted on all the outcome variables we presented in Table 3 and Table 4 (results available from the authors). The evidence provided by those tests suggests that the impact estimates presented in Table 3 and Table 4 appear to be sensitive to selection on unobservables.[26]

VI. Conclusion

The replication of PnK and associated studies poses a challenge due to the weak research design, complex statistical analysis, poor documentation of the data and the absence of code. RnM corroborate the PnK findings, but dispute confidence in causality; Morduch, Chemin and this study argue that PnK overstate the impacts of microcredit. PnK estimated positive and significant impacts for six outcome variables, all of which were stronger when women were microcredit borrowers (Pitt and Khandker, 1998: 987– 988). The largely successful replication by RnM of PnK, and doubts cast on Morduch, left Chemin as the remaining study coming to significantly different conclusions to PnK (see Table A3 in the Online Appendix for an overview of headline findings of main PnK related studies).

Using PSM, Chemin found lower impact estimates than PnK for all outcome variables except male labour supply (Chemin, 2008: 478). Doubts about Chemin arise because of problems in replicating his data constructions, and his failure to conduct sensitivity analysis, or to examine borrowing by women and men separately. Our study finds few convincing effects of microcredit when comparing participants with non-participants. However, the gender of the borrower matters; female borrowing has significant negative effects on the variation of the log of per capita expenditure and significant positive impacts on the log of women's non-landed assets, female labour supply and girls' school enrolment while male borrowing has significant negative effects on expenditure and female labour supply and significant positive effects on the log of women's non-landed assets, male labour supply and girls' and boys' school enrolment. However, sensitivity analysis shows that all these estimates of impact are highly vulnerable to unobservables, in part, perhaps, because of the poor quality of the matches. Thus, we do not corroborate PnK or Chemin; however, it is not clear how to rank the results of PSM and LIML when their results differ; DiPrete and Gangl (2004: 303, emphasis in original) state that

> these two approaches [PSM and IV] cannot be given an *in principle* ranking in terms of their information content.[27]

Nevertheless, in not providing corroboration our study raises concerns about placing heavy reliance on the results of one data set using sophisticated analytical methods to compensate for weak research design.

The analysis in this article also raises doubts about the capabilities of PSM to provide robust estimates of impact, at least with these data. A critique of sophisticated analytical techniques is not new; in a landmark paper Leamer (1983: 38) complained about 'the whimsical character of econometric inference' (see also Leamer, 2010). Despite his pessimistic view of the usefulness of econometric methods, there has been a trend towards ever more sophisticated techniques which, however, do not necessarily provide convincing solutions to the challenges of impact evaluation. A similar conclusion would seem to apply to PSM.

Policymakers would have been well advised to have placed less reliance on PnK, and to have facilitated replication of that study with a more robust research design and better quality data production, sooner rather than later, to avoid missing opportunities for better researched policy. Funders should also encourage replication of analyses using the data sets from which they – the funders – draw policy conclusions. It would be more 'scientific', and save resources in the long run, if data processing and analysis were made more transparent and accountable through the availability of raw data and data processing and analysis code.

Acknowledgements

We wish to thank David Roodman for supporting the data set re-construction and Matthieu Chemin for sharing some Stata do-files. Many thanks also to the World Bank for providing additional data files that were not available online but essential for our analysis. Thanks very much to Ben D'Exelle and to two anonymous reviewers for very helpful comments.

Notes

1. While used in various ways, and sometimes interchangeably, we generally use reproduction to mean using different methods applied to the same raw data, and replication to mean using the same (or very similar) methods (McCullough et al., 2006). Differences in replications, or reproductions, may also occur because of differences in variable construction due to (correction of) errors in calculation, or differences in operationalisation of the underlying concepts.
2. Roodman says: 'A couple of years ago I spent a good deal of time scrutinizing what was then the leading academic study [Pitt and Khandker, 1998] of the impacts of microcredit' (http://www.house.gov/apps/list/hearing/financialsvcs_dem/roodman_testimony_4.28.10.pdf). Pitt did provide a data set including constructed variables to RnM but he cautions that this might not be the same as the one used by PnK (see Roodman and Morduch, 2009: 12, footnote 14).
3. Selection may be by self, peer, or MFI (Armendáriz de Aghion and Morduch, 2010).
4. Hulme and Mosley (1996) were contested by Morduch (1999) and PnK by Morduch and RnM, and these views have been reiterated in readily accessible reviews (Goldberg, 2005; Odell, 2010).
5. We examine only subgroups by gender of borrower in this article.
6. The locations of programmes are also chosen in a non-random way and therefore differ from other places that could be used as controls.
7. The complexity of the PnK and Pitt method, using unique and unrecoverable computer code, seemingly meant this debate remained unresolved in the grey literature until RnM replicated PnK.
8. For details: http://blogs.cgdev.org/open_book/2011/04/a-somewhat-less-provisional-analysis-of-pitt-khandker.php.
9. Imai et al. (2010) use PSM as a check on their IV estimation, finding that PSM confirms their results. Another approach to testing the robustness of PnK analysis would be to perform robustness tests of the LIML estimates involving relatively small perturbations of the underlying data as suggested by Vinod (2009).
10. Other outcome variables have been addressed in other papers using the PnK dataset (Pitt et al., 1999; Pitt, 2000; McKernan, 2002; Menon, 2006; Pitt and Khandker, 2002; Pitt et al., 2003; Pitt et al., 2006), but are not covered in this article.
11. In areas not affected by the cyclone of April 1991.
12. There is some confusion about whether the eligibility criterion is cultivated (operated) or owned land, and whether this includes homestead land. As noted below, the eligibility criterion was not strictly enforced.
13. A thana (literally police station, also known as upazila) is a unit of administration in Bangladesh; in 1985 there were 495 upazilas (Bangladesh Bureau of Statistics, 1985) and 507 upazilas in 2001 (Bangladesh Bureau of Statistics, 2004).
14. Thus there are de jure (cultivating less than 0.5 acres), and de facto (participating) eligibility categories; all de jure are de facto eligible, but not vice versa.
15. http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:21470820~pagePK:64214825~piPK:64214943~theSitePK:469382,00.html.
16. Such as data on consumer price indices, sampling weights and landholding details.
17. The RnM data and code are available at: http://www.cgdev.org/content/publications/detail/1422302. Their variable construction is mainly in SQL, with statistical analysis is in Stata; our data manipulation and analysis is all in Stata. This difference in data management software facilitates triangulation because we were not tempted (able) to borrow code from RnM or take their results as correct.
18. There are small discrepancies in the number of cases used to compute these statistics. PnK exclude households owning/cultivating more than 5 acres. However, with this exclusion neither we nor RnM end up with exactly the same number of cases. Our difference with RnM derives from differences in the construction of the land variable explained in Table A2 in the Online Appendix.
19. RnM do not replicate Chemin or other studies by Pitt and or Khandker which used 'the' PnK data (Khandker, 1996, 2000; Pitt et al., 1999; Pitt, 2000; McKernan, 2002; Pitt and Khandker, 2002; Pitt et al., 2003; Menon, 2006; Pitt et al., 2006). From what we can understand none of Pitt's co-authors has a copy of the data set used in papers of which they were sole or co-authors (personal communications with McKernan and Millimet).
20. We have large differences with Chemin for schooling of individual aged five or above, number of adult male in household, non-agricultural wage and agricultural wage (see Table A1 in the Online Appendix). Many of these variables (apart from schooling of individual aged five or above) are not produced by RnM since they are not required for replicating PnK.

21. A general rule seems to be that there should be at least two control cases per treatment case (King et al., 2011: 8).
22. Odds, which are widely used in assessing probabilistic outcomes, are derived from probabilities ($0 \leq \pi_i \leq 1$) by the following formula: $\pi_i/(1-\pi_i)$.
23. **mhbounds** can be used for a binary outcome variable (Becker and Caliendo, 2007).
24. See Duvendack and Palmer-Jones (2011a, 2011b) for a discussion of levels of Γ that can be considered to demonstrate robustness to unobservables.
25. Results of the **rbounds** tests may differ from the test of the ATT because they are based on different tests (Wilcoxon signed-rank test and Hodges-Lehmann point estimate confidence intervals versus a *t*-test). In this case we use Hodges-Lehmann point estimates (see Rosenbaum, 2002). These are median shifts between treatment groups. Therefore, they are likely to be smaller than the mean shifts reported in Table 3 which provide the average treatment effects.
26. With one exception all the estimates in Table 4 and similar estimates for the other outcome variables have critical values of gamma below 2, with the majority close to 1. The relevant Stata do-files can be made available upon request.
27. As noted above Imai et al. (2010) claim that their PSM corroborates their IV impact estimates. DiPrete and Gangl (2004: 276–278) provide a discussion of the limitations of IV and ways to assess vulnerability to endogeneity bias in both PSM and IV estimates. Their PSM and IV estimations show similar treatment effects.

References

Anderson, R.G., Greene, W.H. et al. (2008) The role of data code archives in the future of economic research. *Journal of Economic Methodology*, 15(1), pp. 99–119.

Armendáriz de Aghion, B. and Morduch, J. (2005) *The Economics of Microfinance* (Cambridge, MA: MIT Press).

Armendáriz de Aghion, B. and Morduch, J. (2010) *The Economics of Microfinance*, 2nd ed. (Cambridge, MA: MIT Press).

Bangladesh Bureau of Statistics (1985) *Statistical Yearbook of Bangladesh 1984–85* (Dhaka: People's Republic of Bangladesh).

Bangladesh Bureau of Statistics (2004) *Statistical Yearbook of Bangladesh 2004* (Dhaka: People's Republic of Bangladesh).

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

Bateman, M. and Chang, H.-J. (2009) The microfinance illusion. Accessed at http://www.econ.cam.ac.uk/faculty/chang/pubs/Microfinance.pdf.

Becker, S.O. and Caliendo, M. (2007) Sensitivity analysis for average treatment effects. *The STATA Journal*, 7(1), pp. 71–83.

Caliendo, M. and Kopeinig, S. (2005) Some practical guidance for the implementation of propensity score matching. Forschungsinstitut zur Zukunft der Arbeit (IZA) Discussion Paper No. 1588, May.

Caliendo, M. and Kopeinig, S. (2008) Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), pp. 31–72.

Chemin, M. (2008) The benefits and costs of microfinance: evidence from Bangladesh. *Journal of Development Studies*, 44(4), pp. 463–484.

Coleman, B.E. (1999) The impact of group lending in northeast Thailand. *Journal of Development Economics*, 60(1), pp. 105–141.

Cornfield, J., Haenszel, W., Hammond, E. and Lilienfeld, A. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22, pp. 173–203.

Deaton, A. (2010) Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2), pp. 424–456.

Dewald, W.G., Thursby, J.G. and Anderson, R.G. (1986) Replication in empirical economics: The Journal of Money, Credit and Banking Project. *American Economic Review*, 76(4), pp. 587–603.

Dichter, T. and Harper, M. (eds) (2007) *What's Wrong with Microfinance?* (Warwick: Practical Action Publishing).

DiPrete, T.A. and Gangl, M. (2004) Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34(1), pp. 271–310.

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

Duvendack, M. and Palmer-Jones, R. (2011b) Reply: much ado about something: response to Haughton's reply to Duvendack and Palmer-Jones. *Journal of Development Effectiveness*, 3(2), pp. 302–308.

Duvendack, M., Palmer-Jones, R., Copestake, J.G., Hooper, L., Loke, Y. and Rao, N. (2011) *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).

Gaile, G.L. and Foster, J. (1996) Review of methodological approaches to the study of the impact of microenterprise credit programs. Report submitted to USAID Assessing the Impact of Microenterprise Services (AIMS), June.

Goldberg, N. (2005) Measuring the impact of microfinance: taking stock of what we know. Grameen Foundation USA Publication Series, December.

Guo, S.Y. and Fraser, M.W. (2010) *Propensity Score Analysis: Statistical Methods and Applications* (Los Angeles, CA: Sage Publications).

Hamermesh, D.S. (2007) Viewpoint: replication in economics. *Canadian Journal of Economics*, 40(3), pp. 715–733.

Heckman, J.J. (1979) Sample selection bias as a specification error. *Econometrica*, 47(1), pp. 153–161.

Heckman, J.J., LaLonde, R. and Smith, J. (1999) The economics and econometrics of active labor market programs, in: O. Ashenfelter and D. Card (eds) *Handbook of Labor Economics, Volume 3A* (Amsterdam: Elsevier).

Hulme, D. and Mosley, P. (1996) *Finance against Poverty* (London: Routledge).

Ichino, A., Mealli, F. and 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.

Imai, K.S., Arun, T. and Annim, S.K. (2010) Microfinance and household poverty reduction: new evidence from India. *World Development*, 38(12), pp. 1760–1774.

Kane, E.J. (1984) Why journal editors should encourage the replication of applied econometric research. *Quarterly Journal of Business and Economics*, 23(1), pp. 3–8.

Khandker, S.R. (1996) Role of targeted credit in rural non-farm growth. *Bangladesh Development Studies*, 24(3–4).

Khandker, S.R. (1998) *Fighting Poverty with Microcredit: Experience in Bangladesh* (New York: Oxford University Press), pp. 181–193.

Khandker, S.R. (2000) Savings, informal borrowing and microfinance. *Bangladesh Development Studies*, 26(2–3), pp. 49–78.

Khandker, S.R. (2005) Microfinance and poverty: evidence using panel data from Bangladesh. *The World Bank Economic Review*, 19(2), pp. 263–286.

King, G. (1995) Replication, replication. *Political Science and Politics*, 28(3), pp. 443–499.

King, G., Nielsen, R., Coberley, C., Pope, J.E. and Wells, A. (2011) Comparative effectiveness of matching methods for causal inference. Accessed at http://gking.harvard.edu/publications/comparative-effectiveness-matching-methods-causal-inference.

Leamer, E.E. (1983) Let's take the con out of econometrics. *The American Economic Review*, 73(1), pp. 31–43.

Leamer, E.E. (2010) Tantalus on the road to Asymptopia. *Journal of Economic Perspectives*, 24(2), pp. 31–46.

McCullough, B., McGeary, K.A. and Harrison, T.D. (2006) Lessons from the JMCB archive. *Journal of Money, Credit, and Banking*, 38(4), pp. 1093–1107.

McCullough, B., McGeary, K.A. and Harrison, T.D. (2008) Do economics journal archives promote replicable research. *Canadian Journal of Economics*, 41(4), pp. 1406–1420.

McKernan, S.-M. (2002) The impact of microcredit programs on self-employment profits: do noncredit program aspects matter? *Review of Economics and Statistics*, 84(1), pp. 93–115.

Menon, N. (2006) Non-linearities in returns to participation in Grameen Bank programs. *Journal of Development Studies*, 42(8), pp. 1379–1400.

Morduch, J. (1998) Does microfinance really help the poor? New evidence from flagship programs in Bangladesh. Unpublished mimeo.

Morduch, J. (1999) The microfinance promise. *Journal of Economic Literature*, 37(4), pp. 1569–1614.

Morduch, J. and Haley, B. (2002) Analysis of the effects of microfinance on poverty reduction. NYU Wagner Working Paper No. 1014, June.

Nannicini, T. (2007) Simulation-based sensitivity analysis for matching estimators. *The STATA Journal*, 7(3), pp. 334–350.

Odell, K. (2010) Measuring the impact of microfinance: taking another look. Grameen Foundation USA Publication Series, May.

Pitt, M., Khandker, S.R., Chowdhury, O.H. and Millimet, D.L. (2003) Credit programmes for the poor and the health status of children in rural Bangladesh. *International Economic Review*, 44(1), pp. 87–118.

Pitt, M., Khandker, S.R. and Cartwright, J. (2006) Empowering women with micro-finance: evidence from Bangladesh. *Economic Development and Cultural Change*, 54(4), pp. 791–831.

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, M.M. (2000) The effect of nonagricultural self-employment credit on contractual relations and employment in agriculture: the case of microcredit programs in Bangladesh. *Bangladesh Development Studies*, 26(2–3), pp. 15–48.

Pitt, M.M. (2011a) Overidentification tests and causality: a second response to Roodman and Morduch. Accessed at: http://www.pstc.brown.edu/~mp/papers/Overidentification.pdf.

Pitt, M.M. (2011b) Response to Roodman and Morduch's 'The impact of microcredit on the poor in Bangladesh: revisiting the evidence'. Accessed at http://www.pstc.brown.edu/~mp/papers/Pitt_response_to_RM.pdf.

Pitt, M.M. and 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), pp. 958–996.

Pitt, M.M. and Khandker, S.R. (2002) Credit programmes for the poor and seasonality in rural Bangladesh. *Journal of Development Studies*, 39(2), pp. 1–24.

Pitt, M.M., Khandker, S.R., McKernan, S.-M. and Latif, M.A. (1999) Credit programs for the poor and reproductive behavior of low-income countries: are the reported causal relationships the result of heterogeneity bias? *Demography*, 36(1), pp. 1–21.

Ravallion, M. (2001) The mystery of the vanishing benefits: an introduction to impact evaluation. *The World Bank Economic Review*, 15(1), pp. 115–140.

Roodman, D. and Morduch, J. (2009) The impact of microcredit on the poor in Bangladesh: revisiting the evidence. Center for Global Development, Working Paper No. 174, June.

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

Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), pp. 41–55.

Rosenbaum, P.R. and Rubin, D.B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), pp. 516–524.

Roy, A. (2010) *Poverty Capital: Microfinance and the Making of Development* (London: Routledge).

Rutherford, S. (2001) *The Poor and Their Money* (New Delhi: Oxford University Press).

Sebstad, J. and Chen, G. (1996) Overview of studies on the impact of microenterprise credit. Report submitted to USAID Assessing the Impact of Microenterprise Services (AIMS), June.

Smith, J. and Todd, P. (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1–2), pp. 305–353.

Stewart, R., van Rooyen, C., Majoro, M. and de Wet, T. (2010) What is the impact of microfinance on poor people? A systematic review of evidence from sub-Saharan Africa London. Social Science Research Unit, Institute of Education, University of London.

Vinod, H.D. (2009) Stress testing of econometric results using archived code for replication. *Journal of Economic and Social Measurement*, 34(2–3), pp. 205–217.

Yunus, M. (1999) *Banker to the Poor: Micro-Lending and the Battle Against World Poverty* (New York: Public Affairs).