Smoke and Mirrors:
Evidence from Microfinance Impact Evaluations in India and Bangladesh

Maren Duvendack

Thesis submitted for the degree of Doctor of Philosophy

University of East Anglia

School of International Development

Submitted September 2010

Abstract

This thesis is framed by the current methodological debate raging in the impact evaluation arena between so-called 'randomistas' (Banerjee et al, 2009; Duflo and Kremer, 2005) who support the use of randomised control trials (RCTs), and their critics (e.g. Deaton, 2009; Imbens, 2009; Pritchett, 2009) who emphasise the continuing value of observational data analysed with advanced econometric methods. It examines the value of observational studies using the evidence of two prominent microfinance impact evaluations conducted in India and Bangladesh.

The thesis first re-examines the microfinance impact evaluation of SEWA Bank conducted by the United States Agency for International Development (USAID) in India. Existing panel data and newly collected cross-section data are subjected to propensity score matching (PSM) and panel data techniques to eliminate selection bias. Sensitivity analysis of the matching results is used to explore their robustness. Furthermore, various sub-group comparisons between borrowers, savers and controls are conducted to examine the impact of savings versus credit. This analysis is supported by direct observation and suggests that selection processes driven by unobservables are at play which cannot be fully controlled by the econometric techniques employed.

Secondly, the evidence of what is regarded as the most authoritative microfinance impact evaluation, conducted by Pitt and Khandker (1998) on paradigmatic microfinance interventions in Bangladesh, is revisited. A number of studies have attempted to replicate the findings of the original study with mixed results. After carefully reconstructing the data set I find a number of inconsistencies, and draw attention to variables overlooked in the data including borrowings from non-microcredit sources. The application of PSM and differences-in-differences (DID) results in the conclusion that these data and methods do not support the main claims of the original study, namely that microcredit has significant poverty reduction outcomes, and is more beneficial when targeted on women than on men.

The application of advanced econometric techniques to observational data was unable in these cases to provide convincing evidence of impact.

Table of contents

List of tables ... 7

List of figures ... 9

List of boxes ... 9

List of appendices ... 9

List of abbreviations ... 10

Acknowledgements ... 12

1. Introduction ... 13

1.1. Definitions and rationale for impact evaluations ... 14

1.2. The impact evaluation challenge ... 15

1.3. Why microfinance? ... 18

1.4. Main hypothesis and thesis outline ... 20

2. Microfinance in India and Bangladesh – history, theory and practice ... 25

2.1. Introduction ... 25

2.2. Historical overview of rural credit markets in colonial and post-colonial India from the early 20th century ... 25

2.2.1. The Integrated Rural Development Programme ... 31

2.2.2. Liberalisation ... 32

2.3. Microfinance in India and Bangladesh ... 34

2.3.1. Group lending ... 40

2.3.2. Theoretical models ... 42

2.4. Microfinance impact evaluations ... 45

2.4.1. Measuring microfinance impact ... 48

2.4.2. Measurement challenges ... 53

2.5. Conclusion ... 58

3. Impact evaluation methods – a dead end? ... 59

3.1. Introduction ... 59

3.2. The evaluation problem 61
3.3. Causal inference and counterfactual modelling 62
3.3.1. The potential outcomes framework 63
3.3.2. The Stable-Unit-Treatment-Value-Assumption 65
3.4. Parameters of interest & selection bias 65
3.4.1. Econometric specification of potential outcomes framework 68
3.5. Experimental approaches 71
3.5.1. Comparison of randomised experiments and analysis of observational data…… 71
3.5.2. Randomisation in development 72
3.5.3. Critique of randomisation in development 77
3.6. Non-experimental approaches 78
3.6.1. Propensity score matching 80
3.6.1.1. Theory 80
3.6.1.2. PSM in practice 82
3.6.1.2.1. Model 83
3.6.1.2.2. Algorithm 84
3.6.1.2.3. PSM in microfinance 85
3.6.1.2.4. Common support 86
3.6.1.2.5. Quality of matches 86
3.6.1.2.6. Sensitivity analysis 87
3.6.1.3. PSM compared to randomised experiments: LaLonde, Dehejia & Wahba and Smith & Todd 88
3.6.2. Instrumental variables 90
3.6.3. Regression discontinuity designs 92
3.6.3.1. Applications of regression discontinuity designs 93
3.6.4. Selection models 95

3.6.5. Differences-in-differences 97

3.6.5.1. Differences-in-differences in microfinance 98

3.7. Conclusion 102

4. Smoke and mirrors: evidence of microfinance impact from an evaluation of SEWA Bank in India 105

4.1. Introduction 105

4.2. Social capital and microfinance 107

4.3. The SEWA Bank context 113

4.4. Impact of savings 115

4.5. Research design and description of data 117

4.5.1. Drop-outs 128

4.6. Sampling procedure 128

4.7. Estimation strategy 130

4.8. Results 135

4.8.1. Income 140

4.8.2. School enrolment 141

4.8.3. Sensitivity analysis 142

4.8.4. Panel analysis 146

4.8.5. Summary 148

4.8.6. Sub-group comparisons 149

4.9. Conclusion 153

5. High noon for microcredit impact evaluations: re-investigating the evidence from Bangladesh 157

5.1. Introduction 157

5.2. The debate 158

5.3. Replication and other challenges 161

5.3.1. Identification strategy - eligibility criterion 164

5.3.2. Multiple sources of borrowing ... 173

5.3.3. Chemin replication ... 176

5.3.4. Limitations ... 186

5.3.5. Sensitivity analysis ... 187

5.4. PSM – ship of fools? ... 190

5.4.1. Determinants of microfinance participation ... 196

5.4.2. Treatment group results ... 202

5.4.2.1. Sensitivity analysis on treatment group comparisons ... 209

5.4.3. cmp with new model specifications ... 209

5.4.4. Comparison of results ... 211

5.5. Panel data ... 214

5.6. Conclusion ... 219

6. Conclusion ... 222

6.1. Lessons learnt and recommendations ... 230

Bibliography ... 234

Appendix ... 258

List of tables

Table 1: Share of rural credit in India by source, 1951 – 1991, in % 27

Table 2: Characteristics of India's financial sector.. 33

Table 3: Ranking of microfinance borrowers by country 37

Table 4: Delivery of microfinance in India by economic activity and delivery model, figures are in % 40

Table 5: Hypotheses and impact variables of USAID study... 119

Table 6: Descriptive statistics of female research respondents....................................... 124

Table 7: Units or levels of evaluation and their advantages and disadvantages 126

Table 8: Logit regression of probability of microfinance participation, without sampling weights.. 134

Table 9: PSM impact estimates of selected household level results – microfinance participants versus controls; without sampling weights 139

Table 10: Sensitivity analysis for household income per annum per capita in Rupees for microfinance participants for Round 1 145

Table 11: PSM and DID results - impact of microfinance participation; without sampling weights 147

Table 12: Selected household level PSM results – sub-group comparisons; without sampling weights 151

Table 13: Degree of mistargeting, total land owned (< 0.5 acres), prior to joining the microcredit programme 170

Table 14: Degree of mistargeting, cultivable land owned (< 0.5 acres), prior to joining the microcredit programme.. 170

Table 15: Degree of multiple programme memberships, household level data, n = 1,798 174

Table 16: Multiple sources of borrowing among microfinance participants and non-participants, person level data 174

Table 17: Chemin's logit specifications predicting the probability of microfinance participation.. 179

Table 18: Replication of Chemin's logit specifications predicting the probability of microfinance participation... 180

Table 19: Chemin's impact estimates and their replication for log of per capita expenditure (Taka) 184

Table 20: Chemin's impact estimates for all 6 outcome variables matching participants to non-participants across treatment and control villages. 185

Table 21: Replication of Chemin's impact estimates for all 6 outcome variables plus log per capita expenditure matching participants to non-participants across treatment and control villages. 186

Table 22: Sensitivity analysis for boys' school enrolment for microfinance participants across R1-3 189

Table 23: Eligibility by village category in % 192

Table 24: Available treatment options in treatment villages 194

Table 25: Descriptive statistics of individuals belonging to any of the four treatment groups across treatment and control villages and across eligibility criteria 197

Table 26: Logistic regression model predicting the probability of microfinance participation using eligible individuals 201

Table 27: Simple matching estimates across gender using 1-nearest neighbour matching with replacement and kernel matching bandwidth 0.05 for all four comparison groups 204

Table 28: Simple matching estimates split by gender using 1-nearest neighbour matching with replacement and kernel matching bandwidth 0.05 for all four comparison groups 207

Table 29: 2SLS estimates of RnM replication and with new model specifications 211

Table 30: Impact of microcredit participation, comparison of random effects model with PSM & DID model 217

List of figures

Figure 1: Growth of India's rural bank branches between 1969 and 2006 30
Figure 2: The objectives of an impact assessment 49
Figure 3: The conventional impact chain model.. 50
Figure 4: Illustration of DID approach, microfinance context 99
Figure 5: Distribution of propensity scores for participants and non-participants 138
Figure 6: Identification strategy corresponding to equations (30a) to (30d)................. 167
Figure 7: Land unit values by total land value and targeting... 172
Figure 8: Distribution of propensity scores for participants and non-participants in treatment and control villages.. 182
Figure 9: Availability of treatment options in PnK study 191

List of boxes

Box 1: Vertical social capital at SEWA Bank 111

List of appendices

Appendix 1: Descriptive statistics for all outcome variables across round 1 - 3........... 258
Appendix 2: Cross-section data results: Detailed household, enterprise and individual level hypotheses - microfinance participants versus controls – without sampling weights... 260
Appendix 3: Sub-group comparisons: Detailed household, enterprise and individual level hypotheses, without sampling weights.. 263
Appendix 4: Weighted means and standard deviations, PnK and RnM 269
Appendix 5: Simplified summary overview of headline results for all PnK related studies dealing with R1-3.. 270

List of abbreviations

2SLS	2-stage least-squares
AER	The American Economic Review
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
ATE	Average treatment effect
ATNT	Average treatment effect on non-participants
ATT	Average treatment effect on the treated
BASIX	Livelihood promotion institution, name of holding company
BRAC	Bangladesh Rural Advancement Committee
BRDB	Bangladesh Rural Development Board
CGAP	Consultative Group to Assist the Poor
CIA	Conditional independence assumption
CPI	Consumer price index
DFI	Development financial institution
DID	Differences-in-differences
DIME	Development Impact Evaluation initiative
FY	Fiscal year
GB	Grameen Bank
IPO	Initial public offering
IRDP	Integrated Rural Development Programme
IV	Instrumental variable
JLG	Joint liability group
LATE	Local average treatment effect
LIML	Limited-information maximum likelihood

MFI	Microfinance institution
MF-NGOs	Microfinance-NGOs
MIS	Management information system
MTE	Marginal treatment effect
NABARD	National Bank for Agriculture and Rural Development
NBFC	Non-banking financial company
NGO	Non-governmental organisation
NSW	National supported work programme
OLS	Ordinary least squares
PROGRESA	Programa de Educación, Salud y Alimentación (Education, health and nutrition programme)
PSM	Propensity score matching
RBI	Reserve Bank of India
RCT	Randomised control trial
RD	Regression discontinuity
RDD	Regression discontinuity design
ROSCA	Rotating savings and credit associations
RRB	Regional rural bank
SEWA	Self-Employed Women's Association
SDI	Subsidy Dependency Index
SHARE	Society for Helping, Awakening Rural Poor through Education
SHG	Self-help group
SOCAT	Social Capital Assessment Tool
SUTVA	Stable-Unit-Treatment-Value-Assumption
USAID	United States Agency for International Development

Acknowledgements

I am deeply indebted to both my supervisors Dr Richard Palmer-Jones and Dr Arjan Verschoor for guiding me through this PhD. Their constant support, encouragement and constructive feedback have been invaluable. I have been very fortunate to work with Richard in particular who shaped and challenged many of my views. Also thanks to Dr Nitya Rao who was part of the supervisory team during the first year of my PhD.

Dr Tara Nair deserves a special thanks not only for patiently listening to my research plans over numerous meals at the Institute of Rural Management Anand but also for facilitating the contact to SEWA Bank and for giving me the opportunity to publish one of my thesis chapters as a book chapter. Thanks also to Professor Keshab Das for guiding me through the affiliation process with the Gujarat Institute of Development Research in Ahmedabad, for sharing his valuable hands-on survey experience in India and for always having time to listen to my fieldwork adventures over a cup of chai.

I am grateful to Ms Jayshree Vyas, Managing Director of SEWA Bank, who has been granting me permission to work with SEWA Bank. Thanks to Ms Hiral Shah for making the logistic arrangements.

Last but not least, I cannot thank my Indian survey team enough for all their hard work. My thanks to Professor J.M. Shah for his field supervision, translation and cross-checking of questionnaires. A special thanks also to my enumerators Mr Hashmukh M. Shah, Mr Parin H. Shah, Mr Krunal G. Shah, Mr Gagendra R. Shah and Mr Mehul G. Paghdad who endured heat, torrential rains and waded through mud to collect data for their 'boss'.

Thanks to Kathryn Forrest for providing a non-economist view on the thesis and to my Mum, Uwe and Shail for endless support, patience and believing in me when the going was tough.

This PhD has been funded by the University of East Anglia.

1. Introduction

Recent years have seen a drive towards encouraging better impact evaluations since donors and policy makers are under increasing pressure to justify public spending on social and economic development interventions. There is generally acknowledged to be a lack of evidence on whether the resources spent on particular programmes actually reach their designated goals in a cost-effective manner (Pritchett, 2002). In other words, there is no clear evidence on what kind of programmes work or do not work, for whom, and under what circumstances.

The evaluation of social and economic programmes using experimental and observational methods has a long tradition. Interest in this area of work intensified in the early 1970s and the evaluation of education and labour market programmes became popular (Imbens and Wooldridge, 2008). The main concern of impact evaluations is to understand how participation in a programme affects individuals. Evaluators are trying to understand how outcomes differ when an individual participates in a programme versus had this person not participated (Caliendo, 2006; Caliendo and Hujer, 2005). Since only one of these outcomes can exist (one either participates or does not), this involves constructing a counterfactual. Constructing a counterfactual that allows observing the potential outcomes of programme participants had they not participated is the main challenge of every evaluation study (Blundell and Costa Dias, 2008; Heckman and Vytlacil, 2007a). The process of constructing a counterfactual commonly is associated with selection bias; selection bias occurs when individuals in a programme select themselves, or are selected by some criteria in such a way that they are different to the general population with whom they are to be compared. Participants may self-select (or be selected) into a programme based on observable and/or unobservable characteristics; e.g. observable characteristics can be employment status, age, sex, educational attainment, and so on, while unobservable characteristics can be motivation, entrepreneurial ability, business skills, etc. (Armendáriz de Aghion and Morduch, 2005). Selection bias is thought to be a particular problem in the context of microfinance (discussed in more depth later in this chapter).

The occurrence of selection bias can lead to errors in the measurement of impact of participation which, it is argued, can be dealt with by a wide range of experimental and observational methods[1]. However, many of these methods have drawbacks of one sort or another, and many fail to control for selection bias due to unobservable characteristics, thus potentially adversely affecting the accuracy of impact evaluation results. These shortcomings have been recognised by numerous government authorities, non-governmental organisations (NGOs), and academics and thus there has been a recent drive towards encouraging better impact evaluations, e.g. organisations such as 3ie (http://www.3ieimpact.org/) or the World Bank's Development Impact Evaluation (DIME) (http://go.worldbank.org/1F1W42VYV0) initiative encourage more rigorous approaches. This thesis aims to contribute to the current methodological debate which is concerned with the appropriate use of experimental versus non-experimental techniques and their ability to control for selection bias due to unobservable characteristics.

1.1. Definitions and rationale for impact evaluations

Impact evaluations attempt to quantify and value the effects that can be attributed to policies or projects. White (2009) argues that two main definitions of impact evaluation are commonly distinguished. The first definition is concerned with the long-term effects of a development intervention on a range of pre-defined outcome indicators. In other words, any evaluation that is concerned with impacts, outcomes and long-term effects is by definition an impact evaluation (White, 2009). The second definition refers to the issue of attribution. The challenge of every impact evaluation is to attribute the observed impact of an intervention to a particular change in the lives of participating individuals or households brought about specifically by the intervention (Hulme, 2000). According to White (2009):

---

[1] Experimental data are produced when units of observation – usually individuals – are randomly allocated by the experimenter to treatment and to control groups (untreated, or placebo treatment). Observational data are produced when some attempt is made to find a comparable group, but without random allocation by the treater. Given the pervasive presence of placebo effects (Goldacre, 2008; Imbens and Wooldridge, 2008), a further level of complication arises with the issue of blinding, i.e. does the treated individual and/or treater know who is receiving the treatment and who is placebo treated. Observational data are not single or double-blinded, while experimental data may be, although this is generally very uncommon if not impossible in social experiments (compared say to pharmaceutical treatments) (Scriven, 2008) – more on this in section 1.2. and chapter 3.

"impact is defined as the difference in the indicator of interest (Y) with the intervention (Y_1) and without the intervention (Y_0). That is, impact = $Y_1 - Y_0$. An impact evaluation is a study which tackles the issue of attribution by identifying the counterfactual value of Y (Y_0) in a rigorous manner" (p. 4).

White (2009) further argues that every methodological discussion should begin by stating which definition of impact is used. This thesis works with the second definition and re-investigates two prominent impact evaluations that claim to demonstrate that an intervention has led to specific changes in selected outcome indicators.

A wealth of impact evaluations of social and economic policies and programmes already exists but unfortunately rigorous evaluations are rare. Pritchett (2002) argues that it is not surprising that there are so few rigorous impact studies. He claims that programmes usually have few incentives to be assessed seriously (chapter 3 develops this point further). Therefore, many impact studies lack credibility because of weak methodologies, data inconsistencies and little understanding of realities in the field (Adams and von Pischke, 1992). Few impact evaluations follow a clear evaluation strategy and lack the correct application of the appropriate methods which can lead to distorted impact assessment results; thus positive or negative results are attributed to a particular programme when in reality those impacts may be due to entirely different reasons (Armendáriz de Aghion and Morduch, 2005).

Overall, it is argued in this thesis that impact evaluations can make a valuable contribution to providing more information about the effectiveness and the efficiency of a programme and can thus help to improve programme practices and policies. This in turn can support policy makers in making better decisions when allocating funds to programmes. However, many impact studies do not provide reliable estimates and hence researchers have spent a considerable amount of time and resources on improving impact evaluation strategies with the objective to provide robust techniques as well as better and more reliable impact estimates.

1.2. The impact evaluation challenge

As mentioned earlier, the task of impact evaluations is to assess how the lives of participants would have turned out had they not participated in the programme; since they do participate it cannot be known what would have happened to them – this is the challenge of measuring the so-called counterfactual (White, 2009; Blundell and Costa Dias, 2008; Heckman and Vytlacil, 2007a). Often, this process is seen as requiring the finding of an adequate control group which would allow a comparison of programme participants with non-participants. This, however, is not straightforward because programme participants commonly differ from non-participants in many ways and not just in terms of programme participation status (Armendáriz de Aghion and Morduch, 2005). A simple comparison between participants and non-participants, i.e. analysing the mean differences of their outcomes after treatment, will give rise to selection bias and will not provide any convincing impact estimates (Caliendo, 2006; Caliendo and Hujer, 2005). Selection bias usually comes in the form of self-selection (individuals commonly self-select into development interventions) or non-random programme placement (specific criteria that drive the placement of programmes, e.g. infrastructure, access to markets, etc.) (White, 2009). Therefore, finding ways to eliminate selection bias should be one of the main preoccupations of those engaged in impact evaluations.

The last two decades have seen advances in the improvement of putatively rigorous econometric techniques designed to account for selection bias. Given those developments randomised control trials (RCTs) slowly took centre stage in particular in the area of development economics (the recent drive towards RCTs is discussed in chapter 3). This thesis examines the effectiveness of those techniques in particular in terms of controlling for selection on unobservables. As mentioned earlier, since the early 1970s researchers increasingly engaged in this area of research. Econometricians such as Heckman (1974, 1976, 1978 and 1979) proposed non-experimental methods to deal with issues such as selection bias, while statisticians such as Fisher (1935), Neyman (1923), and Cox (1958) promoted the application of experimental methods (Imbens and Wooldridge, 2008). There are well known problems with both types of methods which are discussed in detail in chapter 3.

As briefly mentioned in footnote 1, in experimental designs, data are derived from units of observation who are assigned randomly to treatment and control groups, hopefully without any bias in allocations; since, in principle, other things apart from treatment are equal between the treatment and control groups, it is reasonable to attribute differences between these groups after the treatment has been applied to the treatment (Chapter 3 discusses the difficulties with experimental designs in more depth). On the other hand, non-experimental or observational designs are based on data that occur naturally, generally from surveys or censuses, direct observation or administrative data. In these observational situations subjects (individuals, households) generally choose what they do (or are chosen to do what happens to them), in particular, their 'treatment' status. Observational data are what happens in everyday life, and are generally characterised by non-random assignment; in everyday life people who participate are generally different to those who do not. There are numerous threats to both internal and external validity[2] that arise as a result (Shadish, Cook and Campbell, 2002). For example, in assessing the impact of higher education on earnings we observe the earnings of people with and without higher degrees; however, people are not assigned to higher degree courses at random, the more academically able people, those from more advantaged backgrounds and those who attach higher value to education are much more likely to take higher degrees (and until recently, in most countries, they are more likely to be males rather than females). Hence the difference in earnings between those with and without higher degrees will reflect both the effects of different backgrounds and (non-educational) characteristics and those of higher education. Thus, in observational data the effects of education are confounded with those of other characteristics, i.e. confounding variables are both related to the outcome that is being measured and the exposure, and hence can be misleading and should either be excluded or at least minimised because they otherwise produce distorted impact estimates. In other words, failure to take confounding into account produces misleading estimates of impact. There are a number of econometric methods for overcoming, mitigating, or at least documenting the existence and consequences of these biases, which are discussed in depth in this thesis. However, I argue that these econometric techniques have limitations and are often poorly executed or simply misunderstood. A critique of econometric techniques is not new; in a landmark paper Leamer (1983) criticises the key assumptions many econometric methods are built on and complains about "the whimsical character of econometric inference" (p. 38). Despite his pessimistic view on the usefulness of econometric methods, there has been

[2] Internal validity refers to the rigour with which one can assert that the outcomes between the treatment and control groups are different; external validity refers to whether the findings of this comparison are relevant to the broader population from which they are drawn (Shadish, Cook and Campbell, 2002) – chapter 3 discusses this further.

a trend towards ever more sophisticated techniques which, however, did not necessarily provide the solution to the impact evaluation challenge.

As discussed above, obtaining bias-free impact estimates for social experiments is a challenging task, mainly because of the limitations of the evaluation strategies available. There is currently a wider methodological debate raging in the area of development economics dominated by the so-called 'randomistas' (Banerjee et al, 2009; Duflo and Kremer, 2005) who support the use of RCTs and their critics (e.g. Deaton, 2009; Imbens, 2009; Pritchett, 2009). RCTs have been challenged on the grounds of technical fallacies and critics call for a closer look at the value of observational studies which are not flawless either. To illustrate the points made above, I use microfinance as an example and provide the rationale for doing so in the next section.

1.3. Why microfinance?

The concept of microcredit was first introduced in Bangladesh by Nobel Peace Prize winner Muhammad Yunus. Professor Yunus started Grameen Bank (GB) more than 30 years ago with the aim to reduce poverty by providing small loans to the countries' rural poor (Yunus, 1999). Microcredit has evolved over the years and does not only provide credit to the poor but now spans a myriad of other services such as savings, insurances, remittances and non-financial services such as financial literacy training and skills development programmes, and is hence now referred to as microfinance (Armendáriz de Aghion and Morduch, 2005). A key feature of microfinance has been the targeting of women on the grounds that, compared to men, they perform better as microfinance institutions (MFIs) clients and that their participation has more desirable development outcomes (Pitt and Khandker, 1998 – henceforth PnK).

MFIs have become important in the fight against poverty, growing worldwide in number, number of clients, as well as in terms of donor funding (see http://www.mixmarket.org). The sector continues to develop and innovate (Collins et al, 2009; Karlan and Morduch, 2009). A recent trend has been a drive towards commercialisation, i.e. MFIs are scaling up their operations and listing their shares on the stock market – a move which is not entirely unproblematic.

However, despite the apparent success and popularity of microfinance, there is no clear evidence yet that microfinance programmes have positive impacts (Armendáriz de Aghion and Morduch, 2005). There have been four major reviews of microfinance impact (Sebstad and Chen, 1996; Gaile and Foster, 1996; Goldberg, 2005; Odell, 2010) indicating that anecdotes and other inspiring stories (Todd, 1996) show that microfinance can make a real difference in the lives of those served but that rigorous quantitative evidence on the nature, magnitude and balance of microfinance impact is still scarce and inconclusive (Armendáriz de Aghion and Morduch, 2005 and 2010). Overall, there is no well known study that robustly shows any strong impacts (Armendáriz de Aghion and Morduch, 2005, p. 199-230).

A number of studies suggest social and economic benefits of microfinance such as examinations by Hulme and Mosley (1996), Coleman (1999), PnK, Khandker (1998 and 2005), Rutherford (2001) and Morduch and Haley (2002). However, on the other end are investigations by Roodman and Morduch (2009 – henceforth RnM), Bateman and Chang (2009) and Dichter and Harper (2007) that are more cautious and indicate that microfinance is not always beneficial. The debate over microfinance impact intensified recently with the publication of the first two RCTs in the sector (Banerjee et al, 2009; Karlan and Zinman, 2009) raising doubts about the causal link between microfinance participation and poverty alleviation.

Many of the early microfinance impact evaluations fail to address the problem of selection bias as indicated by Sebstad and Chen (1996) and Gaile and Foster (1996). Both studies reviewed a total of 43 microfinance impact evaluations. While many of those 43 studies stressed the importance of finding solutions to the bias problem, most did not attempt to remedy it (Sebstad and Chen, 1996, p. 21). The few studies that addressed the issue of biases more thoroughly were conducted by Hulme and Mosley (1996), PnK and Coleman (1999). Those studies, however, have not been uncontested and the selection bias problem persists until today. In particular the study by PnK caused controversy which is explored in chapter 5. Because of the growth of the microfinance industry and the attention the sector has received from policy makers, donors and private investors in recent years, existing microfinance impact evaluations need to be re-investigated and the robustness of their claims which suggest that microfinance successfully alleviates poverty and empowers women needs to be scrutinised more carefully. Hence, this thesis re-visits the evidence of two prominent microfinance evaluations conducted in India and Bangladesh and specifically focuses on discussing the technical challenges of conducting rigorous impact evaluations within the context microfinance.

Apart from the technical challenges that impact evaluations have to grapple with, they are further hampered by the conflicting agendas of the various players involved. These agendas influence the design, execution and the results of an impact evaluation. Based on observations made in the field, I argue that institutions either directly (imposing restrictions in the field) or indirectly (miscommunication/misunderstandings/provision of flawed or inconsistent data) try to control access, collection of data and essentially the results as well. Those issues shape research outcomes and need to be understood before embarking on any impact evaluation. However, it is beyond the scope of this thesis to discuss those issues in depth since the focus is on the technical aspects of conducting rigorous impact evaluations. As mentioned in section 1.1., many impact evaluations suffer from weak methodologies and inadequate data (Adams and von Pischke, 1992), thus adversely affecting the reliability of impact estimates. This can lead to misconceptions of the actual effects of a programme, thereby diverting attention from the search for perhaps more pro-poor interventions. Therefore, it is of interest to the development community to engage with those evaluation techniques and to understand their limitations so that more reliable evidence of impact can be provided which hopefully leads to better outcomes for the poor.

1.4. Main hypothesis and thesis outline

Having outlined the rationale for impact evaluations, particularly their importance within microfinance, this thesis focuses on the technical challenges that commonly occur during the conduct of quantitative impact evaluations. It is hypothesized that the use of advanced econometric techniques does not generally overcome the problem of selection on unobservables, thus significantly confounding results.

This thesis demonstrates that selection or screening processes driven by the unobservables are likely to be present determining who becomes a participant in microfinance and who remains a non-participant. These processes need to be understood, i.e. qualitative tools in the form of ethnography can be helpful in combination with advanced econometric techniques.

Hence, the following research questions are posed and addressed in the course of this study:

1. What is the main challenge of rigorous impact evaluations? How is impact measured, in particular in the context of microfinance?
2. What is the role of selection on unobservables in rigorous impact evaluations? How does their existence affect impact evaluation results? How can they be eliminated?
3. How does selection on unobservables affect the estimated impact of microfinance (e.g. in relation to control group outcomes)?
4. How do the various evaluation strategies, i.e. econometric methods commonly used in impact evaluation studies, compare in their ability to identify and control for selection on unobservables?
5. What is the impact of microfinance on the economic and social well-being of participating individuals and households?

The thesis is outlined as follows:

Chapter 2 introduces the history, theory and practice of microfinance in India and Bangladesh to provide the contextual background of this thesis. This chapter provides a historical overview of rural credit markets with a focus on their imperfections which partly explain the emergence of microfinance. This is followed by a discussion of the concept of group lending and its role in overcoming information asymmetries. The theoretical discourse implies that microfinance has some impact on reducing information asymmetries as well as transactions costs. Finally, existing microfinance impact evaluations are reviewed and the context-specific challenges of measuring microfinance impact are outlined. This is linked to the recent methodological debate on experimental versus observational studies. The evaluation problem is not unique to the context of microfinance but issues such as fungibility, drop-outs and the particular characteristics of the unobservables make microfinance a challenging area for impact evaluation, and need to be kept in mind when assessing the impact of microfinance.

Chapter 3 begins by setting out the evaluation problem and introduces the main parameters of interest such as the average treatment effect (ATE) and the average treatment effect on the treated (ATT). This chapter also conceptualises the issue of selection bias and focuses on the particular problem of eliminating selection on unobservables. Key topics in evaluation including the debate of experimental versus non-experimental/observational designs are also discussed. Since RCTs recently received a lot of attention and claim to control for selection bias, I present arguments that question the validity and use of RCTs in the specific context of developing countries. I conclude that RCTs are not the silver bullet for solving the problems of evaluation because of technical and ethical reasons, and hence observational designs should not be neglected as a result. Furthermore, there are well-known drawbacks across evaluation methods, in particular with regard to their ability to control for selection on unobservables. Many studies fail to acknowledge that the methods employed do not compensate for the weaknesses of the underlying data. Ultimately, the reliability of the impact estimates depends on the quality of the data.

Chapter 4 re-examines the microfinance impact evaluation of SEWA Bank conducted by the United States Agency for International Development (USAID) in India in 1998 and 2000. The USAID panel and a new cross-section data set are analysed using propensity score matching (PSM) and panel data techniques to address selection bias. Sensitivity analysis of the matching results is used to explore the robustness of the original USAID impact estimates. Various sub-group comparisons between borrowers, savers and controls are also conducted to shed some light on the impact of savings versus credit and to suggest the presence of unobservables which are used by SEWA Bank to select borrowers from among the larger group of savers. I conclude that while the results presented by USAID cannot be contradicted, doubts remain about the quality of the impact estimates obtained through advanced econometric techniques particularly since the panel does not have a 'true' baseline; the base year involved participants who had already been members of SEWA for some years and consequently cannot be shown to be indistinguishable from the control group. In addition, concerns are raised with regard to the robustness of the USAID control group sampling procedure, which is not described sufficiently; direct observation of SEWA Bank fieldworkers strongly suggests selection; and finally, sensitivity analysis of the PSM analysis, which is rather novel in the context of microfinance, suggests that the application of PSM and differences-in-differences (DID) to these observational data is highly vulnerable to selection on unobservables. Consequently, USAID may well have overestimated the impact of SEWA Bank's microfinance services[3].

Chapter 5 re-examines the evidence of what is commonly seen as the most authoritative microfinance impact evaluation (RnM) which was conducted by PnK on three microfinance programmes in Bangladesh. A number of studies attempted to replicate the findings of the original PnK study. For example, Morduch (1998) contradicted PnK but was refuted by Pitt (1999); RnM replicate the findings of PnK but find no evidence for either impact. Chemin (2008) applies PSM and does find significant impacts, but they are smaller than those of PnK and do not distinguish the gender of the borrower. I carefully reconstruct the data and find a number of inconsistencies in the data sets used as well as draw attention to variables overlooked in the data including borrowings from non-microcredit sources. I thus come to the conclusion that these data and methods cannot robustly resolve issues such as the claim that microcredit has significant poverty reduction outcomes, or is more beneficial when targeted on women than on men. An important lesson will be that it is very useful to conduct replications of iconic studies, including re-analysing the existing data and repeating those studies in new contexts using different survey instruments and methods such as PSM, DID and cmp (cmp estimates multi-equation recursive mixed process models and was developed by Roodman (2009) – chapter 5 explains cmp in more detail). Replications have become increasingly important in establishing the credibility and robustness of evaluations[4].

Chapter 6 synthesises the findings of the analyses presented in this thesis and outlines the lessons that may be learnt from these findings. Microfinance is believed to be pro-poor and pro-women but the evidence presented here does not support this belief. The empirical evidence shows that despite the use of advanced econometric techniques, the unobservables that drive the selection process determining microfinance participation are still at play. The analysis in chapters 4 and 5 illustrates that the impact evaluation

[3] This chapter is forthcoming in T. Nair (ed.) *Development Promise of Indian Microfinance* (provisional title), New Delhi: Routledge and is available in the UEA working paper series.

[4] I wish to acknowledge that without the persistence and support of Richard Palmer-Jones I would not have been able to construct the data set used here. Also, David Roodman was exemplary in making his data set available and clarifying various points where we had differences in the reconstruction of some variables.

strategies currently available do not adequately account for those unobservables. The objective of this thesis is to contribute to the wider methodological debate on the use of quantitative impact evaluation techniques with a focus on observational data in the context of microfinance. I argue that, despite the application of advanced econometric techniques, accounting for selection bias and the unobservables remains a challenge for impact evaluations of microfinance in particular, and for impact evaluations more generally. I suggest ways forward to improving our understanding of the unobservables, i.e. rich data sets, appropriate research designs and ethnographic insights are important. In addition, tools such as sensitivity analysis can provide additional information as to the vulnerability of impact estimates to unobservables.

2. Microfinance in India and Bangladesh – history, theory and practice

2.1. Introduction

Microfinance claims to provide an answer to the many problems that commonly plague rural credit markets such as information asymmetries, high transactions costs, the provision of incentives and the enforcement of loan contracts (Basu, 2006). The objective of this chapter is to investigate some of these claims; it begins by introducing the imperfections of rural credit markets by providing a chronological overview of the experiences of India and Bangladesh, and explains the emergence of microfinance in this context. Furthermore, microfinance is widely seen as a response to the failure of rural credit markets. But what is the evidence? This chapter moves away from the particular context of microfinance in India and Bangladesh and provides a general introduction to the theoretical foundations of microfinance, e.g. discusses the concept of group lending, which is seen as an innovative way to substitute physical collateral, and its role in overcoming information asymmetries. Finally, it presents the most prominent microfinance impact evaluations and outlines the context-specific challenges of measuring the impact of microfinance.

2.2. Historical overview of rural credit markets in colonial and post-colonial India from the early 20th century

Throughout the last century, the government of colonial and post-colonial India has made various attempts at improving credit access for the poor. The government recognised that there is a link between access to finance and poverty reduction, thus various policy initiatives were started aimed at financially including the rural poor. Many of these initiatives focused on allocating funds to farmers and rural entrepreneurs with the hope that providing subsidised credit would help to alleviate poverty. The main objective of those credit initiatives was

> "to overcome the monopoly power of private moneylenders, the lack of collateral of small farmers, and the absence of a proper market in loanable funds" (Ellis, 1992, p. 171).

However, many of these government-led lending programs, or social banking programs as they were frequently called, were not always having the desired effects

(Basu, 2006; Hoff and Stiglitz, 1990). Some of these initiatives, i.e. the introduction of credit cooperatives and the launch of the Integrated Rural Development Programme (IRDP) in India, are explored in more depth in this section with the aim to illustrate the imperfections of rural credit markets.

To begin with, in the early 20th century, moneylenders were the dominating force in the domain of credit supply. The colonial administration at the time was aware of the exploitative relationship between creditors and debtors. In many cases, creditors were not only suppliers of credit, they were also buyers of crops, labour employers and landlords. Most debtors repeatedly borrowed money in order to be able to repay debts that they had accumulated earlier, thus entering a vicious circle of indebtedness (Shah, Rao and Shankar, 2007). The colonial administration tried to put a stop to the exploitative relationship between moneylenders and the poor and encouraged the establishment of cooperative credit societies by passing the Cooperative Credit Societies Act in 1904 (Misra, 2010; Shah, Rao and Shankar, 2007). These cooperatives were a major force in Europe in the late 19th and early 20th century (Misra, 2010; Armendáriz de Aghion and Morduch, 2005 and 2010), whereas success was limited in India due to strong socio-economic divisions (e.g. caste system) among the rural poor which undermined any notion of cooperation. The cooperatives in India were mostly managed by rich landowners and moneylenders and did not help changing the exploitative relationship between creditors and debtors (Shah, Rao and Shankar, 2007). Despite the establishment of cooperative credit societies in 1904, traders, landlords and moneylenders still provided 76.8% of rural credit in 1951, while the share of cooperatives and commercial banks was merely at 5.7% (see Table 1). The meagre record of the cooperative movement until the 1950s did not stop the government from further expanding this scheme. By 1961, cooperatives and commercial banks had a 10.3% share in the provision of rural credit which rose to 24.4% in 1971 and then on to 58.6% in 1981 (see Table 1) (Shah, Rao and Shankar, 2007).

Table 1: Share of rural credit in India by source, 1951 – 1991, in %

Credit Agency	**1951**	**1961**	**1971**	**1981**	**1991**
Cooperatives and commercial banks	5.7	10.3	24.4	58.6	58.8
Government and other formal sources	3.1	5.5	7.3	4.6	7.5
Total institutional agencies	**8.8**	**15.8**	**31.7**	**63.2**	**66.3**
Professional and agriculturalist moneylenders	68.6	62.0	36.1	16.1	17.5
Traders		7.2	8.4	3.1	2.2
Landlords		7.6	8.6	4.0	4.0
Relatives and friends	14.4	6.4	13.1	11.2	4.6
Other sources	8.2	0.8	2.1	2.4	2.3
Total non-institutional agencies	**91.2**	**84.0**	**68.3**	**36.8**	**30.6**
Source not specified	0.0	0.2	0.0	0.0	3.1
Total	**100**	**100**	**100**	**100**	**100**

Source: Shah, Rao and Shankar (2007, p. 1355) based on the All-India Rural Credit Survey for 1951 and the All-India Debt and Investment Survey for the years 1951 to 1991.
Note: Traders and landlords are included in 'other sources' for 1951.

As with the cooperative credit societies, the role of commercial banks in India's rural credit markets was rather limited but this changed with the nationalisation of banks in 1969 (Nair, 2006).

Commercial banks had very few branches in the rural areas despite a directive from the Reserve Bank of India (RBI) from 1954 which stated that every bank has

> "to open at least one branch in unbanked rural and semi-rural areas for every branch opened in previously banked areas" (Shah, Rao and Shankar, 2007, p. 1353).

However, the census of 1961 has shown that

> "50% of India's towns and almost none of its villages had bank branches" (Shah, Rao and Shankar, 2007, p. 1354).

Shah, Rao and Shankar (2007) conclude that the rural elite continued to dominate cooperative credit societies and commercial banks were focused on providing credit to large urban borrowers mainly because the poor were by no means the preferred customer group of these formal banking institutions (Basu, 2006). The main reasons banks financially exclude the poor are linked to high risks and high costs. There is a tremendous amount of uncertainty with regard to the repayment capacity of the poor. Credit information is inadequate or unavailable, income and expenditure patterns are irregular and the majority of poor borrowers do not have any collateral to offer. Moreover, exposure to systemic risks such as crop failure or declining commodity prices is fairly high, thus lending to the poor is a high risk venture with high probabilities of default (Basu, 2006).

Speaking in terms of economic theory, banks are often reluctant to offer rural banking services because of existing information asymmetries. Typically, information asymmetries distinguish between adverse selection and moral hazard. In the case of adverse selection, the lender has little or no information on the quality of its borrowers, i.e. are they risky or are they safe (Chowdhury, 2010). Akerlof (1970) referred to this as the 'lemons problem'; he used the automobiles market as an example to illustrate that individuals face uncertainties about the quality of products. Akerlof's (ibid) 'lemons problem' can easily be applied to the context of microfinance. Ideally, lenders should charge higher interest rates to riskier borrowers in order to compensate for the increased probability of default, whereas safer borrowers should be charged less. However, the lender cannot clearly identify which of its borrowers is riskier than others, thus higher average interest rates are passed on to all of them. Safer borrowers may not be interested to borrow at those higher interest rates, even though they might have profitable projects to invest in, while riskier borrowers are willing to borrow at those high rates. Stiglitz and Weiss (1981) call this the problem of credit rationing. Further to this, Chowdhury (2010) argues that

> "raising the rate of interest may not bring efficiency. Hence, the lender's lack of information on the type of the borrowers (who can be good or bad) leads to a situation where the lender may not [*Author's note: be*] able to find an interest rate that appeals to all creditworthy customers and allows the bank to break even" (p. 68).

Moral hazard distinguishes between ex ante and ex post moral hazard. Ex ante moral hazard occurs when lenders are uncertain about the effort the borrower is intending to exert on making the investment project a success. The borrower might just walk away with the loan and use it for purposes other than what agreed upon (Armendáriz de

Aghion and Morduch, 2005 and 2010). This behaviour can affect the returns of the investment project and generates inefficiencies (Chowdhury, 2010). Ex post moral hazard emerges after the investment has been made and the returns of the investment project have been realised. Borrowers might not want to inform the lender of the magnitude of the project's returns (positive or negative) and simply take the money and leave without repaying the loan, claiming exogenous shocks. The lender is often unable to verify the borrowers' project returns and claims, and hence cannot enforce repayment. Therefore, ex post moral hazard is also called the enforcement problem (Chowdhury, 2010).

In addition, transactions costs of servicing the rural poor are high. Poor borrowers require small loan sizes but frequent loan transactions. As briefly mentioned earlier, providing adequate repayment incentives and designing and enforcing loan contracts in a weak regulatory environment with inadequate financial institutions are also seen as major drawbacks (Basu, 2006). In a nutshell, these are seen as the main issues that led to the failure of rural credit markets and to the continued exclusion of the poor from those markets. Microfinance emerged as a response to these issues with the objective to resolve them by employing a range of financial innovations. In fact, the principle of group lending lies at the centre of those innovations. Group lending is mainly being used as a tool to keep default rates low, replace collateral and take

> "advantage of local information, peer support, and, if needed, peer pressure" (Armendáriz de Aghion and Morduch, 2005, p. 13)

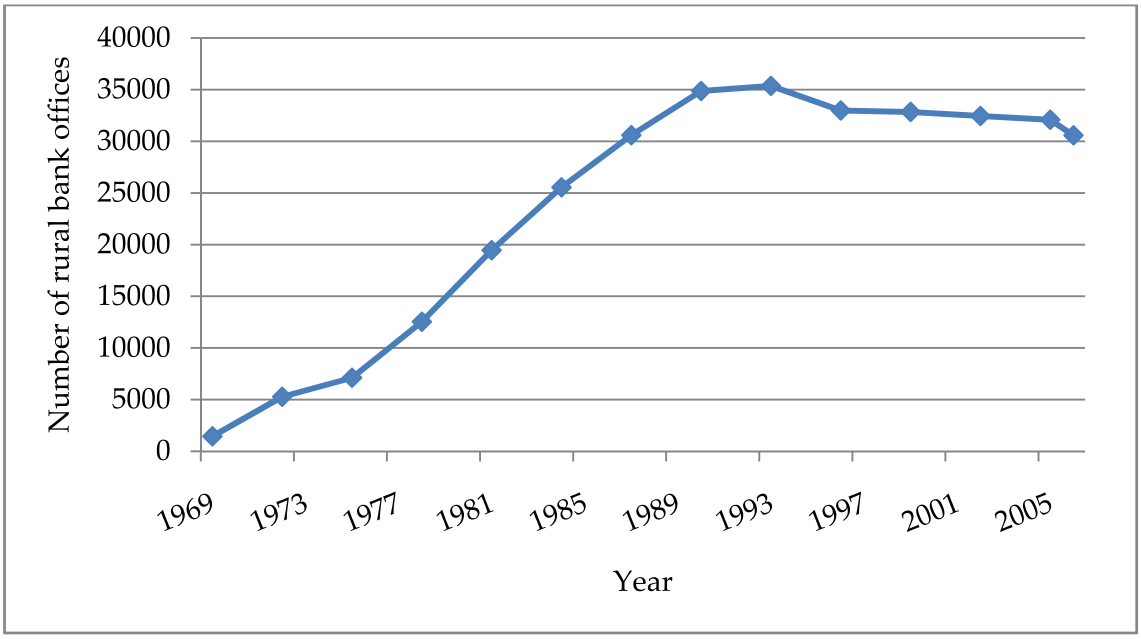
but this topic is further discussed in a subsequent section of this chapter.

Before the emergence of microfinance, however, the nationalisation of banks was initially seen as a way to address these inequalities. Nair (2006) argues that the

> "nationalisation of banks would help improve the flow of formal institutional credit to rural households to relieve them of the burden of usurious credit from informal agencies" (p. 1 - 2).

The Banking Companies (Acquisition and Transfer of Undertakings) Act was passed in 1969 and allowed the state to nationalise commercial banks which it duly did in the same year with 14 of India's largest banks (Shah, Rao and Shankar, 2007; Fisher and Sriram, 2002). Due to the nationalisation, the RBI started to play a more active role in defining banking policies and promoted the expansion of bank branches to unbanked areas. The RBI thus decreed that for every branch opened in a banked area, at least three new branches will have to be opened in unbanked rural and semi-rural areas. Furthermore, in 1976, regional rural banks (RRBs) were set up with the aim to provide credit to India's rural poor. As illustrated by Figure 1, the number of rural bank branches increased tremendously between 1969 and 1993, thanks to RBI's policy, which Shah, Rao and Shankar (2007) call 'social coercion', in which the RBI forced banks to expand into unbanked areas. The RBI also recommended imposing interest rate ceilings under this policy (Ramachandran and Swaminathan, 2005).

Figure 1: Growth of India's rural bank branches between 1969 and 2006



Source: Adapted from Shah, Rao and Shankar (2007, p. 1354).

A final pillar of the policy that Shah, Rao and Shankar (2007) call 'social coercion' was RBI's advocacy of priority sector lending, which meant that a certain percentage of bank lending had to go to priority sectors, which included agriculture, small businesses and industries and weaker sections which had not yet received much attention from banks (Nair, 2006). According to RBI guidelines, 40% of all bank lending had to go to priority sectors. This target was missed by 1% in 1999, whereas 2003 saw a 3% rise to 42% (Sriram, 2005).

2.2.1. The Integrated Rural Development Programme

Thanks to the nationalisation of banks in 1969, the share of rural credit provision by professional moneylenders, traders and landlords fell from 53.1% in 1971 to 23.7% in 1991, as illustrated by Table 1. Moreover, Burgess and Pande (2003 and 2005) argue that RBI's policy of rural bank expansion or 'social coercion' in Shah, Rao and Shankar's (2007) terminology had positive impacts in terms of reducing rural poverty and increasing non-agricultural output.

However, government-run poverty alleviation programmes were not always having the desired effects; the IRDP is a classic example of one of those failed government-led initiatives. The IRDP was initiated in 1978-79 with the objective to provide subsidised credit to India's rural poor, allowing them to create income-bearing assets (Ramachandran and Swaminathan, 2005). At its peak in 1987, the programme covered roughly 4 million households (Shah, Rao and Shankar, 2007, p. 1356). The IRDP, however, had disastrous consequences. It was rife with corruption and abuse by government officials (Dreze, 1990; Shah, Rao and Shankar, 2007). Nair (2006) argues that the IRDP is "a text book case study of a failed directed credit – subsidy programme" (p. 2). Under the IRDP, loans were distributed to virtually any household without first collecting adequate information on borrowers and their cash flow situation. Also, there was no assessment of what the money would be used for and whether it would in fact help to generate income-bearing assets (Shah, Rao and Shankar, 2007). The focus was on increasing credit supply to the rural poor in a short period of time without paying any attention to lending quality (Nair, 2006; Shah, Rao and Shankar, 2007). As a result of this irresponsible lending behaviour, many of the poor defaulted (Shah, Rao and Shankar, 2007). Moreover, the IRDP badly affected the banks active in rural lending. They were saddled with a huge number of non-performing loans, suffered from low profitability and were incapable of covering their own costs; hence the need for frequent capital injections by apex organisations (Shah, Rao and Shankar, 2007). The first-ever formal loan waiver was granted in 1989 and did not particularly help these banks to improve their situation. Moreover, it completely ruined the little bit that was left of credit discipline among the poor (Fisher and Sriram, 2002). Bank profitability remained low due to poor loan recovery rates, low interest rates and high transaction costs (Shah, Rao and Shankar, 2007). Despite the poor record of the IRDP, it was a first serious attempt at financially including the rural poor (Nair, 2006; Ramachandran and Swaminathan, 2005).

2.2.2. Liberalisation

The rural credit sector was in dire need of reforms. Things changed in 1991 when Manmohan Singh became India's Finance Minister under the government of Narasimha Rao which brought the liberalisation of India's economy on its way (Shah, Rao and Shankar, 2007). In this context, the so-called 'Narasimham Committee' (M. Narasimham was a former governor of the RBI) was set up by the RBI with the aim to build a competitive financial sector (Nair, 2006; Shah, Rao and Shankar, 2007). The era of liberalisation saw a reversal of the RBI's policy of 'social coercion' (Shah, Rao and Shankar, 2007). It was recommended to put a stop to the irresponsible lending behaviour which was encouraged under the IRDP programme, deregulate interest rates and reverse the policy of branch expansion (Nair, 2006; Shah, Rao and Shankar, 2007). In addition, banks were required to adhere to market principles and focus on financial performance which, however, negatively affected their presence in the rural credit market (Chakrabarti, 2004). In fact, as a result of those reforms, bank branch offices in rural India saw mergers and closures. After 1993, the number of rural bank offices slowly and steadily declined (as indicated by Figure 1), thus effectively limiting once again the access to credit for the poor. Furthermore, the number of RRBs also declined, from a peak of 196 in 1990 to 104 in 2006 (Shah, Rao and Shankar, 2007).

In general, rural credit markets consist of various players, such as lenders, borrowers and savers, which are brought together by financial intermediaries, i.e. the various institutions involved in rural credit programmes (Ellis, 1992). Credit provision may be through formal, semi-formal and informal channels. Table 2 shows the set-up of India's financial sector to further illustrate this point.

Table 2: Characteristics of India's financial sector

Formal	Semi-formal/microfinance	Informal
Apex development banks	MFIs	Moneylenders
Development financial institutions (DFIs)	SHG (Self-Help Group) – Bank linkage programme	Trade creditors
Commercial banks		Local shopkeepers
RRBs		Relatives, neighbours, friends
Cooperative banks		
Insurance companies		
Mutual funds		
Post office network		

Source: Adapted from Basu (2006, p. 6).

Formal sector providers are institutional and commonly regulated, they typically channel credit to small borrowers primarily through cooperatives and banks (Bouman, 1989). Due to their formalised and rather bureaucratic processes, formal sector providers are not particularly popular with rural households (Sinha, 2000). Sinha (2000) further argues that consumption loans, which are hugely popular among poorer households, are usually not provided by the formal sector. Moreover, formal sector finance tends to discriminate against women who, however, frequently require relatively small sums of money to keep the household running (Sinha, 2006). In addition, credit is only granted when collateral is provided thereby effectively excluding the poor from accessing credit (Sinha, 2000). In contrast to the formal sector, the informal sector is characterised by flexibility and accessibility. However, defining the informal sector clearly is difficult due to its heterogeneity (Bouman, 1989). A range of professional and non-professional financial intermediaries offer a variety of products, e.g. local moneylender, pawnshops, landlords and/or shopkeepers (Bouman, 1989). There is no need to go through a lengthy bureaucratic process, no collateral is required and women have easy access to credit. Loan sizes vary and loan use is not restricted (Sinha, 2000). However, the convenience of informal sector finance comes at a price considering interest rates are higher than the ones offered by formal sector institutions, which are mainly subsidised by the government. Hence, the semi-formal sector, namely microfinance, has become increasingly important because it tries to combine the advantages of informal sector lending with a promise of fair prices and practices (Sinha, 2000).

India's and Bangladesh's rural credit markets have undergone very similar developments due to their shared colonial past – India gained independence in 1947 and the Indo-Pakistani war in 1971 led to the establishment of Bangladesh. After independence, Bangladesh embraced a two-tier cooperative system that specifically targeted the rural poor with less than 0.5 acres of land. This cooperative system is based on the Comilla model which organised farmers into cooperative societies with the main objective to distribute agricultural inputs including subsidised credit. These cooperative societies were then federated into central societies at the thana (an administrative unit governing numerous villages) level (Khandker, 1998). In addition, the government of Bangladesh had also adopted a policy of 'social coercion' and forced nationalised banks and agricultural development banks to expand their networks into rural areas. However, inefficiencies in the formal financial sector led to its dismal performance in supporting economic growth in the country (Khandker, 1998). As in the case of India, many of the government's rural credit initiatives failed and hence the semi-formal and informal financial sector expanded heavily from the early 1970s onwards, i.e. microfinance providers such as Bangladesh Rural Advancement Committee (BRAC) and GB were established in 1972 and 1983 respectively. As of 1998, microfinance programmes in Bangladesh supported more landless households than formal financial institutions, e.g. 80% of the loans disbursed to the landless poor came from microfinance institutions (Khandker, 1998).

The next section now takes a closer look at the role of microfinance in rural credit markets with a particular focus on microfinance in India and Bangladesh.

2.3. Microfinance in India and Bangladesh

The emergence of microfinance can partially be attributed to the failure of rural credit markets and ill-directed government programmes and policies (Chowdhury, 2010). Microfinance has changed banking practices as well as contracts and encouraged the emergence of new financial institutions specifically targeting the poor. In particular, microfinance claims to resolve the issues of information asymmetries, high transactions costs, lack of incentives and enforcement of credit contracts with an innovative approach centred on the formation of joint liability groups (JLGs) (Armendáriz de Aghion and Morduch, 2005 and 2010). Furthermore, microfinance claims to have changed power relationships among the rural poor by specifically targeting and thus empowering women (Hashemi, Schuler and Riley, 1996).

The beginnings of microfinance are usually associated with the start of GB in Bangladesh. Since independence, the country suffered from natural disasters, i.e. cyclones, floods and political unrest. In particular the famine in 1974 was disastrous and provided the backdrop for the emergence of microfinance in Bangladesh. Professor Yunus, who was an economist at Chittagong University at the time, was deeply shocked by the poverty he saw. Hence, he designed a research project in 1976 examining the possibility of providing financial services to the rural poor with the overall objective to alleviate poverty (Khandker, 1998; Hossain, 1988). His project appeared to be successful in helping the poor climb out of poverty and with funds from Bangladesh's central bank as well as other nationalised commercial banks, he then founded GB in 1983 (Yunus, 1999). GB typically organises its borrowers into groups consisting of 5 members each and follows a 'sequential lending' model, i.e. initially a loan is given to two members only and upon successful repayment or at least partial repayment another two group members obtain a loan, and so on (Hossain, 1988). Furthermore, the concept of joint liability (discussed in more detail later in this chapter) plays a crucial role in GB's approach. If a group member falls behind in its ability to repay, then other group members will have to step in. The concept of joint liability replaces the need for physical collateral (Khandker, 1998; Hossain, 1988). Furthermore, GB offers 'dynamic incentives' to its borrowers, e.g. the loan size progressively increases in the case of timely repayments (Chowdhury, 2010). These features are some of the main innovations that still shape the microfinance landscape until today and which have been adopted by MFIs worldwide. With regard to the case of India, microfinance started at a similar time with the establishment of SEWA Bank, a cooperative bank that was founded by Ela Bhatt in 1974. SEWA Bank grew out of a trade union movement with similar objectives as GB, namely to provide financial services to the poor, but with a focus on self-employed women (Bhatt, 2006) – chapter 4 explores the beginnings of SEWA Bank in more depth.

Microfinance has grown rapidly since those early beginnings in the 1970s and received a tremendous amount of publicity since then. For example, 2005 was declared as 'The International Year of Microcredit' (http://www.yearofmicrocredit.org) by the United Nations (UN). In 2006, Professor Yunus and GB won the Nobel Peace Prize for their efforts to alleviate poverty through microfinance. By the end of 2006, more than 10,000 MFIs existed worldwide, serving an estimated 100 million microfinance borrowers (Dieckmann, 2007). Out of those 100 million borrowers, 25 million or 25% were served by MFIs in Bangladesh and roughly 11 million or 11% were serviced by India's microfinance sector. Microfinance in India is still in its nascent stages, in particular relative to the countries' population size and in comparison to microfinance giants such as Bangladesh. India reaches only 3% of its poor, whereas Bangladesh has a penetration rate of 35% (see Table 3 for details). Furthermore, microfinance expanded heavily between 2004 and 2006 thanks to foreign capital investments. Investors tripled their investments in microfinance within these years to USD 4 billion; this figure increased to USD 5.4 billion in 2007 (Karlan and Morduch, 2009).

Table 3: Ranking of microfinance borrowers by country

Country	Total population (m)	Poor people (m)	Borrowers (total, '000)	Penetration rate for microfinance borrowers/poor (%)
Bangladesh	142	70.7	24,757	35%
India	1,090	311.7	10,886	3%
Indonesia	221	59.9	6,421	11%
Vietnam	83	24.0	6,116	25%
Mexico	103	18.1	2,615	14%
Peru	28	14.9	2,036	14%
Philippines	83	30.6	1,919	6%
Colombia	46	29.2	1,449	5%
Sri Lanka	20	4.9	1,422	29%
Ethiopia	71	31.5	1,420	5%
Nigeria	132	45.0	1,392	3%
Morocco	30	5.7	1,046	18%
Pakistan	156	50.9	926	2%
Brazil	186	40.0	915	2%
Nepal	27	8.4	707	8%
Kenya	34	17.8	692	4%
Ecuador	13	6.1	632	10%

Source: Dieckmann (2007, p. 20).

As mentioned earlier, microfinance rapidly expanded in Bangladesh from the late 1970s onwards, while microfinance in India started to seriously take off with the launch of the SHG – Bank linkage programme in 1992 by the National Bank for Agriculture and Rural Development (NABARD). NABARD was established in 1982

> "with a mandate for facilitating credit flow for agriculture, rural industries and all other allied economic activities in rural areas" (Shah, Rao and Shankar, 2007, p. 1355).

NABARD functions as an apex development bank. The aim of the NABARD initiated programme was to provide the rural poor with access to formal credit in a cost-effective and sustainable manner. The SHG – Bank linkage programme works by establishing a linkage between a formal financial institution and an informal SHG (Nair, 2005). According to Pathak (2003) and Nair (2005), the SHG – Bank linkage programme was possibly the largest microfinance intervention in the world in terms of outreach. As of 2002/03, more than 2,000 non-governmental organisations (NGOs) and 444 banks with 17,000 branches were associated with this programme (Pathak, 2003). The SHG programme expanded rapidly: a mere 255 groups were linked to banks in 1992-93 but this number increased to 1,618,476 in 2004-05. The cumulative amount of loans disbursed over the same time period increased from 30 lakh[5] Rupees to 7000 crore[6] Rupees (Nair, 2006).

The potential growth prospects of the microfinance industry and the ensuing money-making opportunities have attracted investments from the public as well as the private sector. Recent years have seen a trend towards commercialisation in the microfinance industry with MFIs preparing for initial public offerings (IPOs), i.e. listing their shares on the stock market to attract investments from private investors (Cull, Demirguc-Kunt and Morduch, 2009). The most prominent examples are Banco Compartamos of Mexico which went public in 2007, Equity Bank in Kenya which listed its shares on the Nairobi Stock Exchange in 2006 and SKS Microfinance which is the first MFI in India to file for an IPO in 2010.

The reason for the drive towards commercialisation is the desire of MFIs to scale up operations and to satisfy their capital needs (Nair, 2005). Nair (2005) argues that

> "though the industry has grown in outreach and disbursements, it is still starved of adequate resources to service the estimated gap between demand and supply of funds" (p. 1697).

Commercialisation is described as the transformation from being a subsidised, donor dependent operation to becoming a regulated financial intermediary (Nair, 2005). The recent trends in microfinance have sparked controversy and divided the sector into those supporting the recent developments and into those opposing them. Professor Yunus is clearly one of the opponents and he responds to the IPO of Banco

[5] A unit used in the Indian numbering system, 1 lakh is equal to 100,000.
[6] A unit used in the Indian numbering system, 1 crore is equal to 100 lakh or 10 million.

Compartamos and the high interest rates they are charging (Compartamos charged 94% per annum at the time of the IPO) as follows:

> "I am shocked by the news about the Compartamos IPO. Microcredit should be about helping the poor to get out of poverty by protecting them from the moneylenders, not creating new ones. A true microcredit organization must keep its interest rate as close to the cost-of-funds as possible. Compartamos' business model, and the message it is projecting in the global capital markets, is not consistent with microcredit. There is no justification for interest rates in the range of 100 percent. My own experience has convinced me that microcredit interest rates can be comfortably under the cost of funds plus ten percent, or plus fifteen percent at the most" (Yunus, 2007, p. 1).

The microfinance sector is clearly on a growth trajectory with MFIs transforming themselves from donor dependent NGOs into mainstream financial intermediaries which compete with formal banking sector institutions (Nair, 2005). The boundaries of microfinance are in the process of being redefined; more products and services are made available with the objective to better address the diverse financing needs of the poor. There is now a move away from group lending to individual lending schemes; which are considered to be more flexible (Hermes and Lensink, 2007). However, the literature so far has primarily focused on group lending schemes which still dominate the area of microfinance, e.g. the PnK study examined in chapter 5 collected data on three MFIs delivering credit through the classic Grameen group lending scheme. A survey of close to 1,500 MFIs across 85 countries conducted by Lapenu and Zeller (2001) showed that approximately 68% of microfinance borrowers obtained their loans through group lending schemes.

In the particular case of India, the majority of microfinance is still delivered through the SHG model (explained in more detail in the next section) except for non-farm enterprises which much rather prefer individual lending schemes (this applies to the case of SEWA Bank which is discussed in chapter 4) or the Grameen model, as illustrated by Table 4.

Table 4: Delivery of microfinance in India by economic activity and delivery model, figures are in %

Model/Sector	Agriculture	Animal Husbandry	Non-Farm Enterprise
Grameen	16	26	57
Individual banking	19	13	68
SHG	34	48	17

Source: Nair (2005, p. 1696).

The next section exclusively focuses on the concept of group lending which still dominates the area of microfinance and illustrates its workings using the Indian context as a backdrop.

2.3.1. Group lending

The main innovation that is commonly attributed to microfinance is the concept of group lending, which attempts to resolve the problem of information asymmetries by encouraging the formation of JLGs. Members of JLGs are collectively responsible for the timely repayment of the loan. If one member defaults on its loan, the other members will have to step in to ensure repayment unless they want to risk losing access to future loans (Chowdhury, 2010). The formation of JLGs thus creates an incentive scheme for each member to screen and monitor the respective members of the JLG (Armendáriz de Aghion and Morduch, 2005 and 2010). In other words, the JLG concept has an effect on maintaining high repayment rates by relying on peer screening, monitoring and enforcement of contracts (Hermes and Lensink, 2007). These effects are confirmed by a range of theoretical models which are reviewed in the next section. To begin with, in order to understand how JLGs work let us take a look at the Indian microfinance sector which is dominated by two different group lending models, i.e. the SHG model and the MFI or Grameen model (Misra, 2010), both explained in more detail in subsequent paragraphs.

SHGs are defined as informal associations consisting of 5 to 20 members. These members are mostly from a similar socio-economic background. SHGs work on the principle of solidarity, no collateral is required. Should a group member default, the other members will have to step in. The concepts of joint liability and peer pressure are at the centre of making the SHG model work. They help to keep transaction and monitoring costs low and ensure that the loans taken out are being repaid (Chakrabarti, 2004).

According to Harper (2002), SHGs and their working mechanisms can vary to a certain degree but they usually operate as follows: first, a group is formed, and then regular savings contributions are made by its members. These savings contributions are pooled and kept by the SHG's elected head or in a bank. In a next step, credit is accessed either by borrowing individually from the savings pool which has been formed by the SHG on terms and conditions which have been decided by the group or by formally linking up with a MFI or a bank. To formally link up with a MFI or a bank, the SHG is required to open a savings account with the respective institution in order to qualify for any of their loans. After that, the MFI or bank extends the loan to the SHG and the group then decides what the money is used for. SHGs effectively function as micro-banks; the MFI or bank only deals with the group as a whole and not with its members individually (Harper, 2002).

As mentioned earlier, the microfinance movement in India began with the promotion of the SHG system. However, the SHG model is rather unique to Indian microfinance. The model that is in fact dominating microfinance in other countries and which is slowly taking over Indian microfinance is the so-called MFI or Grameen model (Misra, 2010; Shah, Rao and Shankar, 2007). As suggested by its name, the Grameen model was developed by GB in Bangladesh and has now been replicated by MFIs across the world which adapted the Grameen-style model to their respective local contexts.

Both approaches, the SHG and the Grameen model, utilise the concept of JLGs. However, the Grameen model differs from the SHG model in many ways. The groups that are formed are smaller, usually consisting of around 5 members. Groups are not linked up with a bank; instead 5 to 7 such groups are organised into centres which are sponsored by either for-profit or not-for-profit MFIs (Harper, 2002). Also, the focus is more on loan provision and recovery and not so much on savings (Shah, Rao and Shankar, 2007). However, with regard to savings the MFIs' strategies can differ tremendously. Some institutions establish compulsory savings schedules while others leave it entirely up to the borrower to decide whether he or she wants to accumulate savings. Moreover, Grameen-type groups generally meet on a weekly basis; these meetings are usually supervised by an employee of the MFI who also collects repayments (Harper, 2002). It appears that Grameen group members are more tightly controlled by the MFI and less autonomous than SHG group members. Moreover, Shah, Rao and Shankar (2007) argue that the organisations following the Grameen approach see the provision of microfinance as an emerging business opportunity with a focus on maintaining high repayment rates rather than a development initiative. Some Indian MFIs have even made headlines with abusive collection practices driving many borrowers to find money elsewhere, i.e. at moneylenders, in order to ensure timely repayment of the microfinance loan. Due to the concepts of joint liability and peer pressure, many borrowers fear to default on a microfinance loan because the other group members are disqualified from obtaining another loan until the previous loan has been fully repaid (Shah, Rao and Shankar, 2007).

Group lending schemes have disadvantages. Many borrowers complain that attending weekly group meetings is too time-consuming and that the idea of joint liability places too much pressure on them; it even discourages creditworthy clients from joining their groups. Hence, as a response to the needs of their clients MFIs have become more flexible and started to provide individual liability schemes (Hermes and Lensink, 2007; Armendáriz de Aghion and Morduch, 2005 and 2010) – as mentioned in the previous section. The literature on comparing group liability with individual liability schemes, however, is still rather underdeveloped and it is not clear which one of these methodologies is superior and will dominate the future (Hermes and Lensink, 2007). The studies by Gine and Karlan (2007 and 2009) and Madajewicz (2004) are some of the few making a credible attempt at investigating the issues of group versus individual liability schemes in more depth. In this study I investigate an individual lending scheme (chapter 4) as well as a group lending scheme (chapter 5) but without attempting to directly compare these two lending methodologies.

2.3.2. Theoretical models

The concept of group lending is commonly heralded as the main innovation of microfinance and claims to provide an answer to the shortcomings of imperfect credit markets, in particular to the challenge of overcoming information asymmetries

(Armendáriz de Aghion and Morduch, 2005 and 2010). As mentioned in section 2.2., information asymmetries may lead to the distinct phenomena of adverse selection and moral hazard. To recap, in the case of adverse selection, the lender lacks information on the riskiness of its borrowers. Riskier borrowers are more likely to default than safer borrowers, and thus should be charged higher interest rates to compensate for the increased risk of default. Accordingly safer borrowers should be charged less provided each type can be accurately identified. Since the lender has incomplete information about the risk profile of its borrowers, higher average interest rates are passed on to all of them irrespective of their risk profile (Armendáriz de Aghion and Morduch, 2005 and 2010). In brief, moral hazard generally refers to the loan utilisation by the borrower, i.e. the lender cannot be certain that once a loan is disbursed, that it is used for its intended purpose, or that the borrower applies the expected amounts of complementary inputs, especially effort and entrepreneurial skill, that are expected and are the basis for the agreement to provide the loan; if these inputs are less than expected then the borrower may be less able to repay it (Ghatak and Guinnane, 1999). As discussed in section 2.2., in addition to adverse selection and moral hazard, high transactions costs, the provision of incentives to borrowers for timely repayment as well as the design and enforcement of adequate loan contracts are further challenges that play a role in explaining the failure of rural credit markets. Given this context, this is where microfinance and its group lending approach steps in. Microfinance advocates claim that the formation of JLGs with its focus on peer pressure and monitoring responds to these challenges. As a result, the theoretical microfinance literature has focused on developing models that explain the workings of the JLG concept and its success in particular in overcoming information asymmetries.

The standard model of lending commonly contains two mechanisms which address the issue of information asymmetries: assortative matching[7] or screening to deal with adverse selection and peer monitoring to overcome moral hazard (Ghatak and Guinnane, 1999). In this widely cited paper, Ghatak and Guinnane (1999) review how

[7] In the event of joint liability group lending where individuals are faced with endogenously forming their own groups (Chowdhury, 2010), safer borrowers commonly form groups with safer borrowers rather than with riskier ones, while riskier borrowers have no choice but to form groups with riskier ones. This is referred to as 'positive assortative matching' (Ghatak, 1999).

the principle of group lending facilitates assortative matching or screening and peer monitoring. Early models were developed by Stiglitz (1990) and Varian (1990) and Banerjee, Besley and Guinnane (1994). These models examined how group liability schemes resolve moral hazard and monitoring problems. Other models developed by Ghatak (1999 and 2000), Gangopadhyay, Ghatak and Lensink (2005) and Armendáriz de Aghion and Gollier (2000) were inspired by Stiglitz and Weiss (1981) and focused on adverse selection and screening mechanisms. Moreover, social ties among group members, i.e. social connections in the language of Karlan (2007), also referred to as social capital, appear to play an important role in the context of group liability schemes in terms of enhancing repayment behaviour, as theorised by Besley and Coate (1995) and Wydick (2001).

The overall thrust of the literature is that the concept of JLGs does indeed overcome adverse selection by introducing better screening mechanisms. In addition, peer monitoring helps to overcome moral hazard and provides group members with incentives to repay loans resulting in high repayment rates (Ghatak and Guinnane, 1999). In spite of that, Hermes and Lensink (2007) argue that MFIs are gradually abandoning the group liability scheme in favour of individual liability schemes; and as briefly mentioned in section 2.3., the literature on theorising individual liability schemes is surprisingly scant. Thus it seems that theory has lagged behind recent developments in the sector and requires some attention. Having said that, the last few years, however, have seen an increase in empirical studies employing experimental techniques, i.e. RCTs, to evaluate the impact of microfinance. These developments have triggered a wave of new theoretical thinking; see Banerjee and Duflo (2010) and Fischer (2010) for recent contributions to the theoretical microfinance literature. It is beyond the scope of this thesis to further contribute to the theoretical discussion since the focus is on the methodological debate that currently dominates the impact evaluation arena.

Hermes and Lensink (2007) further claim that

> "in spite of the abundance of theoretical literature [*Author's note: which models group liability schemes*], there has been surprisingly little empirical evidence of whether and how microfinance actually helps to reduce existing information asymmetries" (p. F3).

The authors further argue that this is

> "due to the difficulty of obtaining reliable data on the working of these programmes and the behaviour of their participants" (ibid, p. F3).

Putatively, convincing empirical evidence is provided by Ahlin and Townsend (2007), Karlan (2007) and Cassar, Crowley and Wydick (2007). The studies by Karlan (2007) and Cassar, Crowley and Wydick (2007) are particularly interesting as they focus on the role of social ties in the context of group lending. Both studies investigate whether the strength of social ties among group members positively affects the repayment behaviour of the group. Chapter 4 elaborates on the topic of social capital in the context of a microfinance impact evaluation conducted in India and further discusses the literature in this area.

In summary, it appears that the concept of joint liability contributes to resolving issues such as adverse selection and moral hazard but whether microfinance is the magic bullet as advocated by many remains to be seen. Hence, the next section takes a closer look at the evidence of microfinance impact more broadly with the objective to establish whether the evidence of microfinance impact in India and Bangladesh presented in this thesis (chapters 4 and 5) is atypical compared to the evidence provided by studies from other countries. Also, the context-specific challenges of measuring microfinance impact are discussed in the next section.

2.4. Microfinance impact evaluations

Measuring the impact of microfinance is a challenging task mainly because of problems with controlling for selection and non-random programme placement bias as briefly mentioned in chapter 1. Adam and von Pischke (1992) argue that most impact studies lack credibility because of weak methodologies and data inconsistencies. Existing microfinance impact evaluations have so far provided rather mixed results and there have been relatively few rigorous statistical analyses. A number of studies suggest social and economic benefits of microfinance such as examinations by PnK, Khandker (1998 and 2005), Rutherford (2001) and Morduch and Haley (2002). However, on the other end are studies by Banerjee et al (2009), RnM, Karlan and Zinman (2009) and Bateman and Chang (2009) that are more cautious and indicate that microfinance is not always beneficial. RnM conclude that

> "30 years into the microfinance movement we have little solid evidence that it improves the lives of clients in measurable ways" (p. 4).

There have been many microfinance impact evaluations, including those employing predominantly qualitative methodologies, with careful selection of comparison groups that have explored the effects of microfinance on incomes. Microfinance practitioners, donors and policymakers assume that microfinance increases borrower incomes (Armendáriz de Aghion and Morduch, 2005). However, this is not always the case as a study conducted by Snodgrass and Sebstad (2002) has shown. They examined the impact of microfinance for SEWA Bank in India, for Zambuko Trust in Zimbabwe and for Mibanco in Peru. The study found that net income gains could only be found in India and Peru. As a study by Mosley (1996, p. 19) on BancoSol in Bolivia shows, only 25% of all microfinance clients showed significant income gains, while 60-65% did not report any changes in income; 10-15% actually went into bankruptcy. It thus appears that microfinance does not work for everyone. In addition, Coleman (2006) and Alexander (2001) find that wealthier households are more likely to be involved in participating in microfinance; this is not surprising considering the prevalence of selection into microfinance programmes by self, others and/or microfinance loan officers. Moreover, measuring income gains is not the only way by which to assess the impact of microfinance; this is partly because of the difficulties with measuring income and interpreting its effects on well-being (Armendáriz de Aghion and Morduch, 2005). There are a number of other channels through which microfinance can influence household outcomes and a number of other variables that can be beneficially affected even if income is not significantly improved. Hence, several authors use other indicators of income such as assets, education and health (examples include PnK; Chen and Snodgrass, 2001, Banerjee et al, 2009). Apart from influencing changes in income and consumption patterns, microfinance can also affect other economic, social and political outcomes such as asset accumulation, business profits, health, nutrition, education and women's empowerment (Armendáriz de Aghion and Morduch, 2005). In particular the impact of microfinance on women's empowerment has received a lot of attention; well-known studies were conducted by Schuler and Hashemi (1994), Hashemi, Schuler and Riley (1996) and Goetz and Sen Gupta (1996), all examining microfinance programmes in Bangladesh.

While the microfinance sector in Bangladesh has been the focal point of many impact evaluations, few rigorous evaluations have been conducted on India's largest and most well-known MFIs (Fisher and Sriram, 2002). These studies on microfinance impact in India vary tremendously with regard to their approach and scope. Moreover, their findings are mixed and range from very positive to almost no impacts at all (Fisher and Sriram, 2002).

As mentioned in section 1.3., the studies by Sebstad and Chen (1996) and Gaile and Foster (1996) both together reviewed 43 microfinance impact assessment studies. While many of these 43 studies stressed the importance of finding solutions to the bias problems, most did not attempt to remedy it. Few impact evaluations have addressed the issue of biases thoroughly mainly because of time and cost constraints. Also, controlling for biases commonly requires advanced statistical skills and data from a suitably designed empirical study. As most studies have the objective of being practitioner-friendly with a focus on rapidly assessing the impact of microfinance programmes, rigorous statistical impact studies usually did not fit these objectives. More thorough studies addressing selection bias were conducted by Hulme and Mosley (1996), PnK and Coleman (1999). Some of these studies did not find conclusive evidence of the beneficial effects of microfinance, and the arguments of others have been contested. For example, Hulme and Mosley (1996) imply that microfinance does on average have positive impacts on the poor but that it does not reach the poorest of the poor. Furthermore, Morduch (1999) criticises the Hulme and Mosley (1996) studies on the basis that they do not succeed in controlling for self-selection bias and that their data lack quality and accuracy. Morduch (1999) is also concerned about the representativeness of their control groups and inconsistencies in their calculations. As an example, he refers to the study on BancoSol in Bolivia, which based its findings on only 24 borrowers (op. cit.). Furthermore, Morduch (1999) claims that the Indonesian study presented by Hulme and Mosley (1996) displays inconsistencies with regard to the quality of the control groups. Thus, he recommends drawing conclusions from these studies with caution.

The most authoritative microfinance impact evaluation to date was conducted by PnK and Khandker (2005) on three MFIs in Bangladesh; for related papers see Pitt et al (2006, 2003, 2002, and 1999). PnK argue that microfinance has significant benefits for the poor, especially when women are targeted (p. 987). However, Morduch (1998) questions the reliability of these results on the basis of a flawed implementation of the research design. Morduch's criticism was vigorously rebutted by Pitt (1999) but neither of their papers was published. Although Goldberg (2005) clearly views PnK as unreliable, he writes that Khandker (2005) "is much less controversial" (p. 19). The matter rested until RnM replicated the main PnK papers in this controversy. After reviewing all these studies, including Khandker (2005), RnM find that

> "decisive statistical evidence in favor of [*Author's note: the idea that microcredit alleviates poverty, smoothes household expenditure and lessens the pinch of hunger especially when women are involved in borrowing*] is absent from these studies" (RnM, 2009, p. 40).

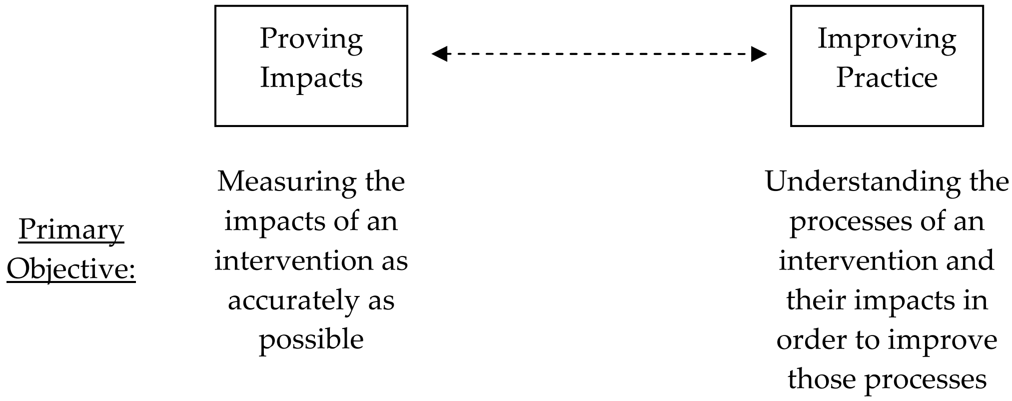
This debate is discussed in depth in chapter 5 which re-analyses, replicates and extends PnK and related studies.

2.4.1. Measuring microfinance impact

Ultimately, all microfinance impact studies are trying to find out how the lives of the poor would have turned out if microfinance had not been introduced. As briefly discussed in chapter 1, this is the problem of measuring the counterfactual which cannot be observed. Thus, every programme evaluation can only make an attempt at creating an estimate of such a counterfactual. These estimates are then used to estimate the effect of the programme (Bryson, Dorsett and Purdon, 2002). The process of estimating counterfactuals commonly introduces biases which adversely affect the reliability of impact evaluation results. Most impact assessors agree that future impact assessments should control for biases because ignoring them can greatly distort impact assessment results (see Sebstad and Chen, 1996; Armendáriz de Aghion and Morduch, 2005; PnK; Coleman, 1999). These are the challenges every impact evaluation has to grapple with and they are not unique to the context of microfinance impact evaluations. However, there is an added component which makes measuring the impact of microfinance especially challenging compared, for example, to measuring the impact of labour training or educational programmes, and that is the issue of fungibility, i.e. a loan can be transferred to other sources or purposes once it has been taken out (Hulme, 2000).

In brief, the challenge of a microfinance evaluation is to demonstrate that a microfinance intervention has led to a specific change. Attributing the effects to a cause requires knowledge of all the other influences that occurred while the intervention took place (Ledgerwood, 1999). Hulme (2000) argues that the objective of an impact evaluation should be to 'prove impact' as well as to 'improve practices', as illustrated by Figure 2. Early microfinance impact evaluations in the 1980s were predominantly used to 'improve practices'. This, however, has now changed and current impact studies focus on 'proving impact' of a microfinance intervention as well as 'improving practices' (Hulme, 2000).

Figure 2: The objectives of an impact assessment



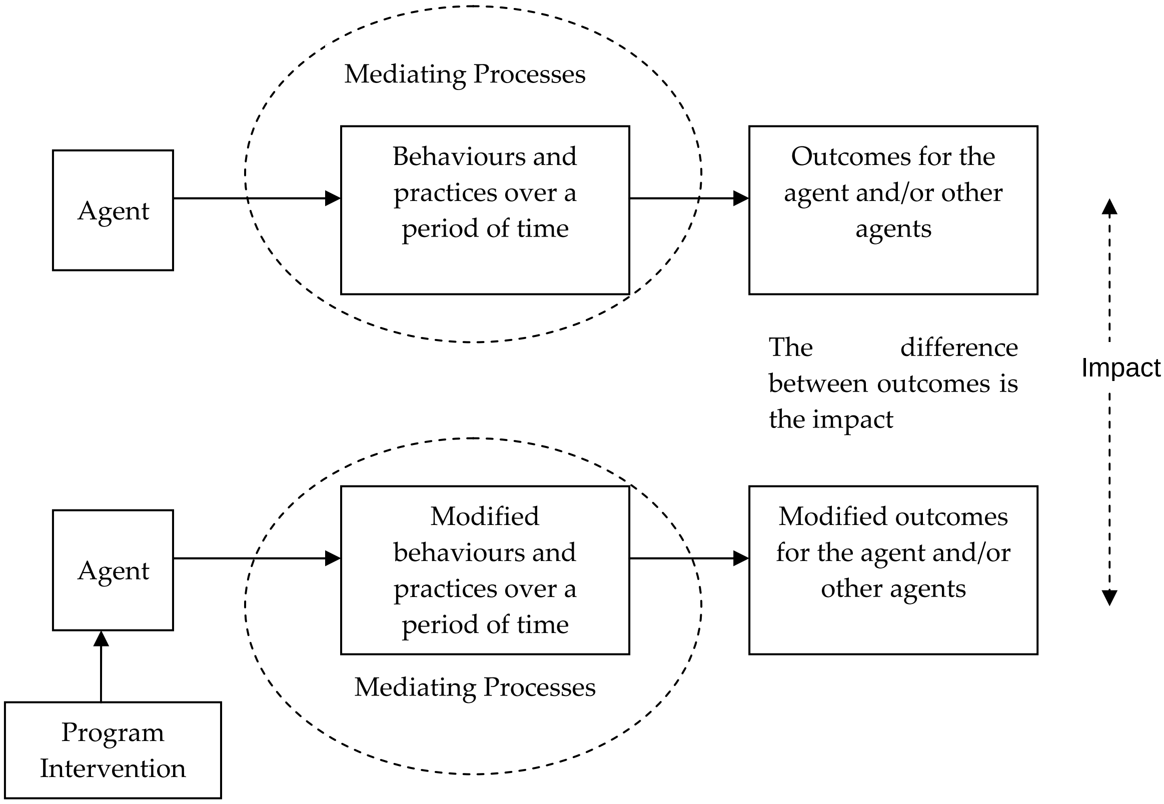
Source: Adapted from Hulme (2000, p. 80).

Hulme (2000) explains that the underlying assumption of virtually all interventions is that it will change patterns of human behaviour in such a way that the probability of achieving certain outcomes is increased. Hulme (ibid) mentions the issue of the counterfactual and argues that impact evaluations are trying to measure

> "the difference in the values of key variables between the outcomes on 'agents' (individuals, enterprises, households, populations, policymakers, etc.) which have experienced an intervention against the values of those variables that would have occurred had there been no intervention" (Hulme, 2000, p. 81).

All these changes in human behaviour are, according to Hulme (2000), brought about by so called 'mediating processes'. He defines 'mediating processes' as processes that are influenced by the particular characteristics of an agent and the environment he or she lives in. According to Sebstad et al (1995), examples of such 'mediating processes' include client characteristics such as gender or ethnicity, household characteristics such as dependency ratios, economic factors such as inflation, price shocks, market structures, etc. It is virtually impossible to predict how these 'mediating processes' affect human behaviour patterns and the outcome of interventions (Sebstad et al, 1995).

Figure 3: The conventional impact chain model



Source: Hulme (2000, p. 81).

Hulme (2000) uses Figure 3 to illustrate the impact chain model and further explains that

> "a more detailed conceptualization would present a complex set of links as each 'effect' becomes a 'cause' in its own right generating further effects" (p. 81).

As an example, he reasons that a microfinance institution provides financial services to an agent. As a result, the agent modifies his behaviour which then influences his income in positive or negative ways. These income modifications can lead to changes in household economic security which in turn influences the household members in terms of level of education, changes in health and potential economic opportunities. On a higher level, this can then have a potential impact on political relations and structures (Hulme, 2000).

This example illustrates that the impact chain is a rather complex structure confronting the impact assessor with a dilemma, namely which link in the impact chain should he or she focus on? Hulme (2000) argues that this question can be answered by looking at two main schools of thought, which he calls the 'intended beneficiary' school and the 'intermediary' school.

As its definition suggests, the 'intended beneficiary' school attempts to measure the impact of microfinance on the intended beneficiary such as individuals and households by moving down the impact chain as far as possible (Hulme, 2000). Makina and Malobola (2004) argue that this school reflects the conventional project life cycle approach which is donor driven with the main concern of justifying the continued funding of the programme in question. According to the 'intended beneficiary' school,

> "the impact of aid-funded projects needs to be measured and attributed in order to justify the effect of the intervention through its direct impact on the poor" (Makina and Malobola, 2004, p. 801).

This can be achieved, for example, by using assets and net worth of a beneficiary as indicators for assessing the impact of microfinance (Makina and Malobola, 2004). According to Johnson (1998), the 'intended beneficiary' school looks at micro-credit as a separate productive input with significant effects on improving the livelihoods of the poor. These improvements are assumed to be achieved thanks to the ability of credit to increase incomes, reduce vulnerabilities and provide alternatives to moneylenders. However, not everybody agrees with the claims of the 'intended beneficiary' school. Adams (1988) for example, is sceptical. He argues that credit is highly fungible and thus it is difficult to identify the impact of credit through a particular beneficiary. Rather than using credit to make investments in the microenterprise, the poor have a tendency to use it for consumption instead.

The 'intermediary' school follows a different approach and is more concerned with changes occurring in the microfinance institution and its operations. It focuses on the beginning of the impact chain and is primarily preoccupied with parameters such as institutional outreach and institutional sustainability. The 'intermediary' school is linked to the so-called ‘Ohio School’[8], which heavily criticises the provision of subsidised credit both through the state in form of rural financing schemes and NGOs. It is said that subsidised credit distorts the market (von Pischke, Adams and Donald, 1983; Adams, Graham and von Pischke, 1984; Ellis, 1992). Moreover, the most well-known proponents of the ‘Ohio School’, Adams and von Pischke (1992), argue that debt is not an effective tool for improving livelihoods. Subsidised credit unnecessarily burdens the poor and increasingly leaves them to deal with a mountain of bad debt. However, the ‘Ohio School’ is very supportive of building financial institutions that aim for financial self-sustainability (von Pischke, 1991).

Generally speaking, the ‘intermediary’ school believes that an intervention has had positive impacts when institutional outreach and sustainability both improved as this leads to the financial market being widened in a sustainable way. The underlying assumption of this belief is that such institutional impacts provide additional alternatives to already existing credit and savings services which in turn influence household economic security, etc. (Hulme, 2000). Both schools have certain advantages and disadvantages. According to Hulme (2000), the ‘intended beneficiary’ school, for example, is better suited for identifying who benefits and how, while not making too many assumptions with regard to the impact chain. He further argues that following this school is methodologically rather demanding and costly. While the ‘intermediary’ school is not very strong with regard to identifying who benefits and how, it certainly presents a sound methodological impact assessment framework, which can be used for already existing data sets and which includes the concept of sustainability (Hulme, 2000). The discussion of the ‘intended beneficiary’ school and the ‘intermediary’ school has shown that there may be no perfect approach to conducting an impact assessment. Recent years have seen a movement in the industry that attempts to combine the different approaches and create more practitioner-friendly assessment tools. This movement gave rise to the Imp-Act project based at the Institute of Development Studies at the University of Sussex (http://www.ids.ac.uk/go/idsproject/impact; Armendáriz de Aghion and Morduch, 2005 and 2010). Furthermore, organisations such as USAID and the Consultative Group to Assist the Poor (CGAP) followed suit and

---

[8] The ‘Ohio School’ is commonly associated with the Rural Finance Program at Ohio State University. The Rural Finance Program is noted for its research on development finance.

developed their own practitioner-friendly methods (Barnes and Sebstad, 2000). While these assessments are very useful for the institutions themselves, they are not rigorous quantitative measures of impact (Armendáriz de Aghion and Morduch, 2005 and 2010). As a consequence, more recently, there has been growing emphasis on more rigorous impact evaluations combining these approaches by emphasising both empirical intended beneficiary impacts and 'theory based' evaluations. This trend is manifested in 3ie - The International Initiative for Impact Evaluation - and the World Bank's DIME initiative.

2.4.2. Measurement challenges

Apart from the anecdotal evidence provided by qualitative studies (e.g. Todd, 1996), the majority of microfinance impact evaluations have applied a quasi-experimental design. As mentioned earlier, well-known examples include Hulme and Mosley (1996) and PnK as well as Coleman (1999) who used a pipeline quasi-experiment design which was innovative at the time, to address selection bias.

In a quasi-experiment the outcomes of an intervention are typically compared

> "with a simulation of what the outcomes would have been, had there been no intervention" (Hulme, 2000, p. 84),

thus allowing the measurement of the counterfactual. As in the experimental design, quasi-experiments identify a treatment and a control group. However, the treatment group already participates in a microfinance intervention, whereas the control group should be as identical as possible to the treatment group in terms of economic and social set-up but minus the microfinance intervention (Hulme, 2000). There are problems with using a quasi-experimental design, for example the identification of identical control groups, in other words, the challenge of overcoming selection bias.

In the case of microfinance, selection bias typically comes in the form of self-selection bias and non-random programme placement bias (PnK; Coleman, 1999). According to Coleman (1999), self-selection bias refers to microfinance programme members that have self-selected into a programme. The decision to participate may have been influenced by certain unobservable characteristics such as entrepreneurial skills, organisational abilities and motivation (Armendáriz de Aghion and Morduch, 2005 and 2010) which tend to increase the likelihood of individuals to self-select into a programme. Coleman (1999) further argues that prospective borrowers will not only have to make a decision on programme participation but they will also have to gain acceptance from incumbent borrowers who have also self-selected into the programme. As a result, programme members may significantly differ from non-members in terms of motivation or wealth. Also, programme participation is voluntary and individuals with more entrepreneurial drive or business skills are more likely to participate. In other words, programme participants usually self-select, meaning they choose to enter a microfinance programme in a non-random way or are selected by their peers. In addition, microfinance loan officers play a role in selecting borrowers. Hence, impact studies need to address this problem because estimates obtained in the presence of selection bias will most likely be invalid.

Moreover, programme placement can also be biased; MFIs assign new programme locations in non-random ways based, for example, on considerations for infrastructure or wealth (Hulme, 2000; Coleman, 1999). Some programmes, for example, are placed in areas which are easily accessible, although the opposite could also be possible; PnK report that the microfinance programmes in Bangladesh at the time of their study were placed in particularly flood-prone areas.

The issue of selection bias is further set out and conceptualised in chapter 3 where it becomes clear that selection bias needs to be controlled for by appropriate evaluation strategies. An additional challenge that needs to be tackled in the specific context of microfinance impact evaluations is the issue of fungibility, as discussed above. According to Ledgerwood (1999), money is considered to be fungible within the household. This means that once a loan has been taken out by the borrower, it is difficult to track in which way the loan has actually been used. For example, a borrower could transfer his loan to someone else or the loan could be used in ways which were originally not intended (Feder et al, 1990; Adams and von Pischke, 1992; Ledgerwood, 1999; Hulme, 2000). The loan might be tied to investments in the microenterprise but instead it is being used for household consumption. Borrowers generally do not have an incentive to be honest about the utilisation of their loans; they might be afraid that they will not qualify for future loan disbursements since the loan is not used for the originally intended purpose (Johnson and Rogaly, 1997; Ledgerwood, 1999). Furthermore, according to Ledgerwood (1999), there are additional reasons for borrowers not wanting to be honest about the (mis-)utilisation of their loans, e.g. shame, embarrassment that others in the community might find out about it and concerns with regard to taxation, just to name a few. Thus, it is very difficult to gather reliable data on this issue. According to Hulme (2000), there are currently no studies that have successfully remedied the problem of fungibility. He suggests using case study material which would allow comparing actual loan use with intended loan use and then attempt an estimation of potential 'leakage' (Hulme, 2000, p. 85). He further argues that the problem of fungibility might not be relevant after all (Hulme, 2000, p. 85). Hulme (ibid) reiterates that studies which solely focus on the enterprise level will have to deal with the issue whereas studies focusing on the household or the community level will find this problem to be irrelevant. Furthermore, Hulme (ibid) claims that fungibility can be beneficial for enhancing microfinance impact. Armendáriz de Aghion and Morduch (2005) agree with Hulme (2000) and argue that fungibility is indeed not such a major problem after all. They further claim that despite the fact that money is fungible within the household, and that

> "… a given loan cannot be attached to a given change in profit, it is still possible to evaluate how profits change with capital (i.e. measure the marginal return to capital) and how borrowing affects household-level variables such as income, consumption, health, and schooling" (Armendáriz de Aghion and Morduch, 2005, p. 223).

A further problem is that most microfinance impact assessments neglect to account for the issue of drop-outs. It is a microfinance reality that clients exit programmes once they have exhausted the utility of the products and services available. Armendáriz de Aghion and Morduch (2005) report drop-out rates between 3.5% and 60% in a wide range of microfinance programmes worldwide. Another example is the study conducted by Khandker (2005) which provides evidence that the drop-out rate in the examined programme in Bangladesh is around 30%. A study by Alexander-Tedeschi and Karlan (2007) on Peru finds that the drop-out rate is about 56%.

Typically, there are two different types of clients exiting a programme, namely graduates and drop-outs. Graduates exit a microfinance programme because they have successfully utilised the microfinance institution's products and have outgrown the programme. They have 'graduated' and are hopefully qualified to move into the realm of the formal financial sector and apply for commercial banking services. Drop-outs, on the other hand, were not satisfied with the products of the microfinance programme and thus decided to exit. They could have also had other reasons for leaving, e.g. they could not afford the microfinancial services, they did not benefit from them, they failed to utilise microfinance to their advantage maybe due to a lack of skills and thus defaulted or they left the programme because of personal events such as death of the household head, wedding of a family member, birth, etc. (Karlan and Goldberg, 2006). In this context, Karlan (2001) argues that it is also important to analyse the alternatives of credit and savings available in the communities. Some households might exit a programme because the alternatives are better suited for them. Many households in fact have multiple sources of borrowing, i.e. microfinance and non-microfinance loans, which they use interchangeably for investments, consumption and/or repayment of debts (Coleman, 1999). The implications of this for measuring microfinance impact are discussed in more detail in chapter 5. For the sake of simplicity, the remainder of this section uses the term drop-out bias or biases due to drop-outs, both terms referring to both types of drop-outs, i.e. graduates and drop-outs.

Drop-out bias can be positive or negative depending on the reasons for dropping out. For example, the impact of the programme will be underestimated when richer clients are more likely to leave the microfinance programme, thus leaving behind only the poorer ones. The impact of the programme will be overestimated when poorer clients predominantly leave, thus leaving behind only the richer ones (Karlan, 2001). This illustrates that biases due to drop-outs can have significant effects on the conclusions of microfinance impact evaluations. Thus, evaluators should be aware of this problem and control it as best as they can.

However, the literature on microfinance impact evaluations has mostly neglected drop-out bias so far. Gaile and Foster (1996) reviewed 11 impact studies and Sebstad and Chen (1996) examined 32 impact assessments, none of which took biases due to drop-outs into account. Even the studies by PnK and Coleman (1999), which were innovative in many respects and the first to make a serious attempt at rigorously controlling for biases, did not account for drop-out bias. Karlan (2001) and Alexander-Tedeschi and Karlan (2007) are some of the few authors that stress the importance of further researching drop-out bias. They argue that neglecting biases due to drop-outs can lead to distorted impact evaluation results.

One of the few studies that accounts for drop-out bias was conducted by Alexander-Tedeschi and Karlan (2007) and Tedeschi (2008) in their work on Mibanco borrowers in Peru. First, they measure the impact of microfinance without accounting for drop-out bias and in a second step re-calculate the data set to correct for drop-outs. They find that microenterprise profit showed an increase of USD 1,200 when drop-out bias was not controlled for, i.e. using only remaining borrowers. The same figure, however, fell dramatically when drop-out bias was accounted for and suddenly showed a decrease of approximately USD 170.

Karlan (2001) argues that the problems associated with drop-out bias can be solved. He suggests that better sampling techniques can help to solve this issue. Drop-outs should be included in the treatment group even though this might take an extra effort to track down those who have dropped out. Armendáriz de Aghion and Morduch (2005), however, argue that this method is too time-consuming and costly. Instead, they suggest using an econometric approach to deal with drop-out. In a first step, the authors recommend estimating predictors of drop-outs. These predictors are based on observable characteristics from the treatment group. In a next step, a prediction is formed on who among the new borrowers in the control group is likely to remain in the programme, and then the new borrowers in the control group are weighted according to their probability of remaining.

Moreover, Alexander-Tedeschi and Karlan (2007) add that systematically interviewing drop-outs can help to control the quality of the programme and improve the microfinance services and products. The authors recommend further analysis of drop-out bias in order to get a better picture on the pattern of drop-outs and to what extent not including them in the analysis will affect impact evaluation results. Hence, chapter 4 made an attempt at sampling drop-outs, i.e. graduates as well as drop-outs, and include them in the analysis. However, there were problems in sampling and then finding drop-outs but this is discussed in more depth in chapter 4.

Partly in response to critical reviews of evaluations using observational (qualitative and quantitative) data there has been a trend towards using experimental methods, i.e.

conducting RCTs of many development interventions including microfinance (see microfinance RCTs by Banerjee et al, 2009; Karlan and Zinman, 2009). RCTs claim to resolve the issue of selection bias. However, they are vigorously debated and many microfinance interventions lack crucial characteristics necessary for valid RCTs but this is discussed in more detail in chapter 3.

2.5. Conclusion

The purpose of this chapter was to provide the reader with the contextual background of rural credit markets and microfinance with a focus on India and Bangladesh. The chapter began with a brief chronological overview of the various policies and strategies that have been initiated in particular on the Indian subcontinent with the aim to financially include the rural poor. It became clear that many of the earlier subsidised credit programmes were not having the desired effects and did not meet the needs of the poor. The recent emergence and rapid growth of microfinance should be understood in this context. The discussion in this chapter then moved away from the particular context of India and Bangladesh to outline the concept of group lending, i.e. the formation of JLGs, which has shown that microfinance has some impact on reducing information asymmetries and transaction costs as well as on enforcing contracts but that it is not the magic bullet as advocated by many. This chapter also reviewed existing microfinance impact evaluations more broadly to establish whether the evidence presented in chapters 4 and 5 on microfinance impact in India and Bangladesh is atypical compared to the evidence provided by microfinance impact evaluations from other countries. Furthermore, this chapter pointed out that issues such as fungibility, drop-outs and the particular characteristics of the unobservables in the microfinance context are challenges that need to be kept in mind when assessing the impact of microfinance. Hence, the next chapter takes a look at the recent methodological debate on experimental versus observational studies and links this debate to the specific challenges of measuring microfinance impact. The following chapter illustrates that selection on unobservables plays a central role and has effects on the robustness of impact estimates. Therefore, it explains the theoretical issues of impact evaluations and introduces the various methods of evaluating impact by using examples from microfinance where appropriate.

3. Impact evaluation methods – a dead end?

3.1. Introduction

As outlined in the previous chapter, accounting for selection bias due to observable as well as unobservable characteristics is one of the main tasks every impact evaluation has to grapple with and many studies claim to have mastered this task with the application of advanced econometric techniques. However, I argue that, despite the application of these techniques, controlling for selection bias and observing the unobservables remains a major challenge. The subsequent empirical chapters which assess the impact of microfinance interventions in India and Bangladesh further scrutinise the usefulness of these econometric techniques in reliably assessing impact. Furthermore, I illustrate that not acknowledging selection on unobservables can adversely affect impact estimates and that the quality of the underlying data as well as an understanding of the contextual background of the programme that is evaluated are crucial to the reliability of impact evaluation results.

This chapter introduces the extensive literature dealing with the evaluation of social and economic programmes and critically discusses some of the techniques that statisticians and econometricians have developed to account for selection on observables as well as unobservables. As briefly mentioned in chapter 1, since the early 1970s, researchers have become more and more interested in this area of work and the evaluation of labour market programmes attracted a great deal of attention (Imbens and Wooldridge, 2008). Econometricians such as Heckman (1974, 1976, 1978 and 1979) typically focused on issues such as endogeneity and self-selection and proposed non-experimental methods to deal with these issues, while statisticians such as Fisher (1935) – commonly heralded as the inventor of randomised experiments - Neyman (1923), and Cox (1958) promoted the use of experimental methods (Imbens and Wooldridge, 2008).

Experimental designs are still relatively rare in the area of development economics because of some obvious drawbacks which are discussed later in this chapter. However, they are steadily gaining in popularity in particular in the area of microfinance which has seen an increase in RCTs; e.g. noteworthy studies were conducted by Banerjee et al (2009) and Karlan and Zinman (2009). In spite of this, observational studies still dominate but are not free from debate either due to issues such as confounding variables and selection on observables as well as unobservables (RnM). Econometricians have put forward various methods for dealing with these challenges such as PSM, instrumental variables (IV), regression discontinuity designs (RDD) and DID, etc. which continue to be strongly advocated, and are discussed in depth in this chapter.

Thus, recent years have seen a surge in publications extensively reviewing econometric programme evaluation techniques (Blundell and Costa Dias, 2000, 2002 and 2008; Caliendo, 2006; Caliendo and Hujer, 2005; Imbens and Wooldridge, 2008). Hence the objective of this chapter is not to provide another review of these techniques but to critically shed light on their usefulness in particular with regard to controlling for selection on unobservables. Also, a discussion on the suitability of different methods in relation to the evaluation questions posed, the quality of the data available and the assignment mechanism that allocated individuals to a particular programme or intervention is provided. In order to select convincing methods for analysis, it is crucial to understand the underlying assignment mechanisms and their reliability (Blundell and Costa Dias, 2008), and this requires a deeper understanding of the social, economic and political context of the programme.

As briefly mentioned in the previous chapter, it is not clear how we are trying to measure impact, i.e. what is the appropriate counterfactual? Therefore, this chapter begins by setting out the evaluation problem and the challenges of causal inference and counterfactual modelling. It then proceeds to discussing the most commonly used experimental and non-experimental techniques in the area of impact evaluation and assesses how well suited they are to controlling for selection on observables as well as unobservables especially when heterogeneous treatment effects can be assumed. The recent evaluation literature has given a lot of attention to the challenges of dealing with heterogeneous treatment effects and hence this chapter further examines how the evaluation techniques presented here are able to address heterogeneity (Caliendo, 2006; Caliendo and Hujer, 2005).

The issues discussed in this chapter are related to the area of development economics by using the case of microfinance wherever appropriate to further illustrate the challenges of obtaining reliable impact estimates in this particular context.

3.2. The evaluation problem

The main concern of evaluations is to understand how programme participation affects the outcomes of individuals. Evaluators are trying to understand how outcomes differ when an individual participates in a programme versus had this person not participated (Caliendo, 2006; Caliendo and Hujer, 2005). In other words, individuals can either participate or not in a given intervention but they cannot do both at the same time. Constructing a counterfactual that would allow observing the potential outcomes of programme participants had they not participated is the main challenge of every evaluation study (Blundell and Costa Dias, 2008, Heckman and Vytlacil, 2007a). This requires finding an adequate control group which would allow a comparison of programme participants with non-participants. This, however, is a major challenge because programme participants commonly differ from non-participants in many ways and not just in terms of programme participation status. A simple comparison between participants and non-participants, i.e. analysing the mean differences of their outcomes after treatment, will give rise to the issue of selection bias and will not provide any convincing impact estimates (Caliendo, 2006; Caliendo and Hujer, 2005). In summary, the evaluation problem is essentially that of missing data because only one side of the comparison can be observed (Ravallion, 2001; Blundell and Costa Dias, 2008; Imbens and Wooldridge, 2008; Heckman and Vytlacil, 2007a). This gives rise to the problem of causal inference. Imbens and Wooldridge (2008) state that

> "in order to evaluate the effect of the treatment we therefore always need to compare distinct units receiving the different levels of the treatment" (p. 1).

The authors further argue that the literature so far has mainly focused on dealing with binary treatments, though Morgan and Winship (2007), Imbens (2000), Lechner (2001) and Hirano and Imbens (2004) briefly discuss approaches to matching with multi-valued treatments.

The evaluation problem is pervasive, in particular in the context of microfinance as briefly described in chapter 2. Existing microfinance impact evaluations have so far provided rather mixed results and lack rigor and credibility mainly due to not adequately controlling for selection bias. Few microfinance impact evaluations have addressed the issue of biases thoroughly (e.g. Sebstad and Chen, 1996; Gaile and Foster, 1996) and the few that did are not uncontested (PnK; Hulme and Mosley, 1996; and Coleman, 1999).

Furthermore, Pritchett (2002) argues that it is not surprising that there are so few rigorous impact studies. Not only is that a phenomenon in the area of microfinance but health and education interventions are met with the same fate. Pritchett (2002) claims that programmes usually have few incentives to be assessed seriously. Conducting a thorough impact evaluation is costly, time intensive and requires statistical skills which not everybody possesses. Most programmes conclude that the costs of an impact study by far exceed its benefits.

The next section conceptualises causal effects in the context of observational studies and introduces the widely acclaimed 'potential outcomes framework'.

3.3. Causal inference and counterfactual modelling

This section begins by providing an understanding of the identification problem and by explaining the necessity to address it successfully since this has implications for the reliability of causal models (Heckman and Vytlacil, 2007a). In an influential book devoted to the topic of identification and statistical inference, the economist Charles F. Manski who has made numerous contributions to the subject matter since the mid 1970s explains that the identification problem arises when we are trying

> "to learn what conclusions can and cannot be drawn given specified combinations of assumptions and data" (Manski, 1995, p. 3).

In other words, to what extent can we build theoretical constructs and determine models given the empirical evidence that is available (Heckman and Vytlacil, 2007a)? The identification of appropriate counterfactuals and the corresponding treatment parameters is the main challenge when discussing causal inference. The unobservables drive the identification problem, i.e. researchers commonly do not have sufficient information explaining the link between programme participation decision and outcomes (Heckman and Vytlacil, 2007a). In other words, researchers are not aware of the unobservable characteristics that drive the decision of individuals to participate.

Hence, these unobservables need to be controlled for by applying appropriate techniques aimed at eliminating selection bias (Heckman and Vytlacil, 2007a). These techniques are outlined in more depth later in this chapter.

3.3.1. The potential outcomes framework

In a next step, to enhance our understanding of causal effects in observational studies, the 'potential outcomes framework' is introduced. In the early 1970s, the statistician Donald B. Rubin put forward the 'potential outcomes framework' which was derived from a series of papers he had published (Rubin, 1973a, 1973b, 1974, 1977, and 1978). It is also called 'Rubin's Model' by Holland (1986) and Heckman, Lochner and Taber (1999) and is now dominating the econometrics and statistics literature. This chapter uses the term 'potential outcomes framework' throughout to avoid any misunderstandings. The 'potential outcomes framework' is also applied in sociology, psychology and political science (Morgan and Winship, 2007). Heckman and Vytlacil (2007a) remark that this framework has its origins in the statistics literature that dealt with experimental designs and which can be traced back to Fisher (1935), Neyman (1923) and Cox (1958). Essentially, the 'potential outcomes framework' is an extension of Neyman's (1923) work according to Rosenbaum (2002).

The 'potential outcomes framework' is commonly used to understand causal effects in the context of observational studies. It proposes to compare potential outcomes to allow analysing causal effects (Imbens and Wooldridge, 2008). Imbens and Wooldridge (2008) further explain that

> "pairs of outcomes [*Author's note: are*] defined for the same unit given different levels of exposure to the treatment. Models are developed for the pair of potential outcomes rather than solely for the observed outcome" (p. 2).

In other words, the 'potential outcomes framework' revolves around individuals' programme participation or non-participation and potential outcomes. It suggests that two potential outcomes exist for each individual, one for each treatment state: Y^1 and Y^0. Y^1 represents participation in the programme and Y^0 stands for non-participation (Caliendo, 2006; Caliendo and Hujer, 2005). Hence, Y^1 denotes the treatment state and Y^0 denotes the control state. Individuals that are assigned to the treatment state are referred to as the treatment group and individuals in the control state commonly form the control group (Imbens and Wooldridge, 2008). In addition, a binary variable reflecting actual programme participation is included, where D = 1 when an individual participates and D = 0 when an individual does not participate. The difference of potential outcomes between participants and non-participants is then the treatment effect as denoted by equation (1)[9], where i stands for each individual (Caliendo, 2006; Caliendo and Hujer, 2005):

(1) $$\Delta_i = Y_i^1 - Y_i^0$$

However, there is a problem because the observed outcome for each individual i is denoted as follows:

(2) $$Y_i = D_i Y_i^1 + (1 - D_i)\, Y_i^0$$

As mentioned earlier, only one of these outcomes can be achieved because an individual i either participates or not but it cannot do both at the same time. Both outcomes, Y_i^1 and Y_i^0 , are potentially observable before an individual i is assigned to a treatment state, hence the model refers to *potential* outcomes (Imbens and Wooldridge, 2008). An observed outcome Y_i^1 will be achieved when individual i participates in the programme and this makes Y_i^0 the unobservable counterfactual outcome. Similarly, if Y_i^0 is observed, then Y_i^1 is the unobservable counterfactual (Morgan and Winship, 2007). Since individual i either participates or not, estimating equation (1) is a challenging task because one component, i.e. the counterfactual, cannot be observed (Caliendo, 2006; Caliendo and Hujer, 2005).

According to Imbens and Wooldridge (2008), the main advantage of the 'potential outcomes framework' is that it allows researchers to define the causal effect at an individual level by taking the difference of Y_i^1- Y_i^0 without having to make any other further assumptions about endogeneity or exogeneity. Furthermore, it engages with the various assignment mechanisms "defined as the conditional probability of receiving treatment" (Imbens and Wooldridge, 2008, p. 7). Classically assignment mechanisms can be distinguished as follows: (i) randomised experiments; (ii) unconfounded assignment referring to selection on observables, exogeneity and conditional independence; and (iii) all other assignment mechanisms not covered by (i) and (ii) (Imbens and Wooldridge, 2008).

[9] Notations in this section follow Caliendo (2006).

3.3.2. The Stable-Unit-Treatment-Value-Assumption

Caliendo (2006) cautions that

> "the concentration on a single individual requires that the effect of the intervention on each individual is not affected by the participation decision of any other individual, i.e. the treatment effect Δ_i for each person is independent of the treatment of other individuals" (p. 12).

The statistics literature refers to this as the Stable-Unit-Treatment-Value-Assumption (SUTVA) which will have to be observed in order to avoid biases such as externalities, spill-overs and network effects. The 'potential outcomes framework' assumes that there is no interaction between individuals. For example, it is assumed that whether an individual receives a new painkiller or not will not have any effects on the health outcomes of other individuals (Imbens and Wooldridge, 2008). In other words, no spill-over effects from the treatment of individuals to the society as a whole can be expected (Heckman, Lochner and Taber, 1999). Observing SUTVA, however, is a challenging task and its violation is a main concern and indeed a common occurrence in reality. Imbens and Wooldridge (2008) provide examples from agricultural and epidemiological interventions that illustrate that interactions cannot be avoided altogether, i.e. treated individuals can indeed have an impact on the outcomes of non-treated individuals. Despite these challenges of maintaining SUTVA, few studies have in fact dealt with this problem. An attempt is made by Heckman, Lochner and Taber (1999) who analyse policies on tax reforms and tuition subsidies and resort to a simulation approach that provides evidence that biases arise when ignoring these interactions between individuals.

3.4. Parameters of interest & selection bias

An intervention can have homogeneous or heterogeneous treatment effects across individuals. In other words, does participation in a particular programme evoke the same responses across individuals? If not, are there any systematic differences within these responses? (Blundell and Costa Dias, 2008). A central theme in the evaluation literature is to provide solutions to dealing with heterogeneous treatment effects. The early literature in this area had assumed that treatment effects are homogeneous and hence believed that the impact of an intervention can be described by one single parameter (Imbens and Wooldridge, 2008). Blundell and Costa Dias (2008) reiterate that in the case of homogeneous models

> "there is only one impact of the program and it is one that would be common to all participants and non-participants alike" (p. 8).

The literature has progressed and recognised that this is not the case. In fact, heterogeneous models are the norm where programme benefits differ between participants and non-participants. This will have an affect on the average treatment effect on the treated which is expected to be different to the average treatment effect on the untreated (Blundell and Costa Dias, 2008, p. 8). Hence, a set of parameters is indeed needed to describe the impact of an intervention since heterogeneous treatment effects prevail (Imbens and Wooldridge, 2008).

As mentioned earlier, estimating equation (1) is a challenging task, and hence alternative ways for calculating treatment effects are needed. The literature refers to two parameters that are popular for estimating treatment effects[10]. The (population) ATE is one of them. The ATE estimates

> "… the difference of the expected outcomes after participation and non-participation" (Caliendo, 2006, p. 13)

and can be denoted[11] as follows:

$$\Delta_{ATE} = E\ (\Delta) = \ E\ (Y^1) - E\ (Y^0) \tag{3}$$

In other words, ATE includes impact estimates for treated and non-treated individuals (Caliendo, 2006; Imbens and Wooldridge, 2008). Therefore, the ATE might not always be the appropriate choice and a more suitable parameter was developed, namely the (population) ATT (Imbens and Wooldridge, 2008). The ATT includes estimates only for those individuals the programme was specifically meant to target, as described by equation (4):

---

[10] There are other parameters of importance such as the marginal treatment effect (MTE) introduced by Björklund and Moffitt (1987) and further developed and discussed by Heckman and Vytlacil (1999, 2000, 2001, 2005, 2007a and 2007b) and the local average treatment effect (LATE) introduced by Imbens and Angrist (1994). In addition, Blundell and Costa Dias (2008) refer to the average treatment effect on non-participants (ATNT). The focus here is on ATE and ATT since those parameters will be of relevance in later chapters. MTE and LATE are briefly introduced in section 3.6.2.

[11] Notations in section 3.4. follow Caliendo (2006) unless stated otherwise.

$$(4) \qquad \Delta_{ATT} = E\,(\Delta|D=1) = E\,(Y^1|D=1) - E\,(Y^0|D=1)$$

The ATT allows the estimation of actual programme benefits by concentrating on programme participants (Caliendo, 2006; Caliendo and Hujer, 2005). Hence, this parameter appears to be more interesting than the ATE when trying to assess the success of an intervention (Imbens and Wooldridge, 2008). Essentially, the choice of the appropriate treatment parameter depends on the information that is accessible to the researcher, i.e. the availability and quality of the data, and on the assumptions that are made based on these data (Heckman and Vytlacil, 2007b). For example, researchers

> "…using matching make strong informational assumptions in terms of the data available to them. In fact, all econometric estimators make assumptions about the presence or absence of informational asymmetries…" (Heckman and Vytlacil, 2007b, p. 4885).

Caliendo (2006) mentions that not all the components in equation (4) can be observed, i.e. $E\,(Y^0|D=1)$ illustrates potential outcomes in the event of no treatment for individuals that were in fact treated. Furthermore, the author argues that non-participants would form a suitable control group when the following assumption holds: $E\,(Y^0|D=1)$ = $E\,(Y^0|D=0)$. This, however, will only be the case for experimental studies but will not apply to non-experimental studies. Hence, comparing the mean values of participants $E\,(Y^1|D=1)$ and non-participants $E\,(Y^0|D=0)$ using ATT commonly introduces selection bias (Caliendo, 2006; Caliendo and Hujer, 2005). Heckman and Vytlacil (2007b) state

> "that there is selection bias with respect to Y^0, Y^1, so persons who select into status 1 or 0 are selectively different from randomly sampled persons in the population" (p. 4882).

Further to this, Caliendo and Hujer (2005) explain that

> "selection bias arises because participants and non-participants are selected groups that would have different outcomes, even in absence of the programme. It might be caused by observable factors, like age or skill differences, or unobservable factors like motivation" (p. 4).

Imbens and Wooldridge (2008) argue that it becomes even more difficult to estimate these parameters if selection bias due to unobservable characteristics exists. A classic solution to this is the application of the IVs approach. IVs are commonly used to identify treatment parameters such as ATE, ATT, etc. However, the reliability of the IV estimates depends on the quality of the underlying instruments that are employed. Moreover, when heterogeneous treatment effects exist, the common assumptions of the IV approach cannot be maintained. IVs are still useful for defining treatment parameters but their identification is made with difficulties (Heckman and Vytlacil, 2007b). IVs are discussed in more detail later in this chapter.

3.4.1. Econometric specification of potential outcomes framework

Caliendo (2006) proposes to express the 'potential outcomes framework' in common econometric language to allow a better understanding of homogeneous and heterogeneous treatment effects and suggests following the notation of Blundell and Costa Dias (2002):

(5) $$Y_{it}^1 = g_t^1(X_i) + U_{it}^1 \text{ and } Y_{it}^0 = g_t^0(X_i) + U_{it}^0$$

Where:

i = the individual

t = time period

g_t = link between potential outcomes and observable characteristics

X = regressor X

U_{it} = error terms which are not correlated with regressors X

Furthermore, Caliendo and Hujer (2005) assume that programme participation occurs in period k, and hence the treatment effect for any X_i can be described as follows:

(6) $$\Delta_{it}(X_i) = Y_{it}^1 - Y_{it}^0 = [g_t^1(X_i) - g_t^0(X_i)] + [U_{it}^1 - U_{it}^0]$$

with $t > k$

Consequently, the ATT which is estimated after the treatment, i.e. when $t > k$, is expressed as:

(7) $$\Delta_{ATT} = E\ (\Delta_{it} | X = X_i, D_i = 1)$$

It is assumed that individuals are assigned to treatment groups in non-random ways. To illustrate this point, we can take the example of microfinance again. Microfinance participants differ from non-participants in terms of education or wealth; in addition programme participation is voluntary and individuals with more entrepreneurial drive or business skills are more likely to participate (Coleman, 1999). In other words, programme participants usually self-select or are selected by their peers or loan officers based on these observable and/or unobservable characteristics, i.e. they enter a microfinance programme in a non-random way which gives rise to the problem of selection bias. In more technical terms, selection bias arises when the following occurs: $E\ (U_{it}\ D_i) \neq 0$ (Heckman and Robb, 1985). Heckman and Robb (1985) use an index function framework to further describe this non-random treatment assignment[12]:

$$IN_i = Z_i\gamma + V_i \quad (8)$$

Where:

IN_i = Index of benefits to the individual making a favourable treatment decision

Z_i = Function of observed variables

V_i = Function of unobserved variables

Further to this:

D_i = 1 if $IN_i > 0$

= 0 otherwise.

And hence, Heckman and Robb (1985) argue that the reason for selection on observables or unobservables can be stochastic dependence between the error term (U_{it}) and the function of the variables that are observed (Z_i) for the former or stochastic dependence between the error term (U_{it}) and the function of the variables that are unobserved (V_i) for the latter – as outlined in equation (8).

A final word with regard to homogeneous and heterogeneous treatment effects before proceeding to discussing experimental approaches. It was mentioned earlier that the evaluation literature initially assumed that treatment effects are homogeneous across

[12] Notation for equation (8) follows Heckman and Robb (1985).

individuals. It is now recognised, however, that treatment effects are in fact heterogeneous (Imbens and Wooldridge, 2008).

Homogeneous treatment effects can be expressed by the following equation:

$$(9) \qquad \Delta_t = \Delta_{it}(X_i) = g_t^1(X_i) - g_t^0(X_i) \qquad \text{with } t > k$$

In which

> "g^1 and g^0 are two parallel curves that differ only in the level and, furthermore, that participation-specific error terms are not affected by the treatment status" (Caliendo, 2006, p. 16).

Caliendo (2006) therefore suggests re-arranging the outcomes equation (5) as outlined by Blundell and Costa Dias (2002):

$$(10) \qquad Y_{it} = g_t^0(X_i) + \Delta_t D_{it} + U_i$$

In the case of heterogeneous treatment effects, however, equation (5) changes as follows:

$$(11) \qquad Y_{it} = g_t^0(X_i) + \Delta_t(X_i)D_{it} + [U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)],$$

where

$$(12) \qquad \Delta_t(X_i) = E[\Delta_{it}(X_i)] = g_t^1(X_i) - g_t^0(X_i)$$

which is the estimated treatment impact for individuals characterised by X_i at time t (Caliendo and Hujer, 2005; Blundell and Costa Dias, 2002). Equations (10) and (11) are used again in sections 3.6.4. and 3.6.2. respectively to illustrate selection models and IVs.

Having introduced the parameters of interest and conceptualised selection bias, the next section provides an introduction to experimental methods before outlining the various non-experimental methods. It is commonly assumed that experimental designs have the ability to resolve the issue of selection bias and hence provide more robust results than those obtained using non-experimental methods. The following section discusses whether this assumption holds and outlines the challenges of implementing experimental studies.

3.5. Experimental approaches

At the heart of every experimental design lies a natural or an artificially formulated experiment which attempts to attribute the effects of an intervention to its causes (Hulme, 2000). Evaluations applying a randomised design are generally believed to provide the most robust results. There is a long tradition of experimental methods in the natural sciences. As mentioned earlier, Fisher (1935), Neyman (1923) and Cox (1958) were early proponents of randomised experiments. However, few randomised experiments have been conducted in the social sciences in the past. Many of these early experiments were regarded with suspicion as to their credibility of establishing causality and their importance for researchers and policy-makers alike (Imbens and Wooldridge, 2008). Moreover, in many cases the availability of experimental data are limited and hence, observational studies continued to capture the attention of many researchers and this culminated in the development of the 'potential outcomes framework' by Rubin (1973a, 1973b, 1974, 1977 and 1978) which was discussed earlier.

3.5.1. Comparison of randomised experiments and analysis of observational data

However, in the 1980s researchers started to seriously question the reliability of methods commonly used to analyse observational data and their ability to establish causal effects. For example, an influential study by LaLonde (1986) which evaluated the impact of the National Supported Work (NSW) programme showed that the econometric and statistical methods used in observational studies could not replicate the estimates obtained from experiments. As a result, experiments apparently steadily gained in popularity again. Nonetheless, experimental studies are not uncontested either, and there has been renewed interest in and advocacy of observational methods (Deaton, 2009; RnM). For example, the study of LaLonde (1986) was re-analysed by Dehejia and Wahba (1999 and 2002) who employed PSM and illustrated that PSM can in fact approximate the results obtained from an experimental setting. The author's argue that their results

> "are close to the benchmark experimental estimate" (Dehejia and Wahba, 1999, p. 1062), i.e. "[a] researcher using this method [*Author's note: propensity score matching*] would arrive at estimates of the treatment impact ranging from $1,473

> to \$1,774, close to the benchmark unbiased estimate from the experiment of \$1,794" (ibid, p. 1062).

However, Smith and Todd (2005) argue that the PSM estimates calculated by Dehejia and Wahba (1999 and 2002) are sensitive to their choice of a particular sub-sample of LaLonde's (1986) data and find evidence that a DID approach is in fact more appropriate as an evaluation strategy in this context than PSM as proposed by Dehejia and Wahba (1999 and 2002). Dehejia (2005) responds to Smith and Todd (2005) and corrects some of their claims (further discussed in the PSM section later in this chapter). Overall, the outcome of this debate remains inconclusive with strong evidence provided by all parties involved. Nonetheless, recent years have seen a revival of the debate that raged in the 1980s on the value of experimental versus observational studies in particular in the field of development economics; this debate was resumed by researchers like Burtless (1995) who argues that randomised studies provide more credible estimates and Heckman and Smith (1995) who are somewhat more critical. Apart from the LaLonde – Dehejia & Wahba – Smith & Todd debate, there have not been many studies comparing experimental with observational studies; the few studies that convincingly did so were conducted by Meier (1972), Fraker and Maynard (1987) and Friedlander and Robins (1995).

3.5.2. Randomisation in development

Notwithstanding the critique of randomised studies, there has been a move towards RCTs in development economics driven by the so-called 'randomistas' (Banerjee et al, 2009; Duflo and Kremer, 2005; Miguel and Kremer, 2004). A classic example from a developing country setting is the randomised evaluation of PROGRESA in Mexico which triggered a wealth of publications, e.g. see studies by Behrman, Sengupta and Todd, 2005; Gertler, 2004; Gertler and Boyce, 2001; Hoddinott and Skoufias, 2004.

To begin with, the nature of randomised experiments is explained. Applying a randomised study design requires to randomly assign potential clients to so-called treatment and control groups, whereby both groups must be drawn from potential clients whom the programme has yet to serve so that the impact of an entire programme can be evaluated (Karlan and Goldberg, 2006). This random assignment to either treatment or control group ensures that potential outcomes are not contaminated by self-selection into treatment (Blundell and Costa Dias, 2008). In other words, the potential outcomes or effects of the treatment are independent from the treatment assignment. Proper randomisation ensures those individuals in treatment and control groups are equivalent in terms of observable and unobservable characteristics with the exception of the treatment status, assuming that no spill-over effects exist (Blundell and Costa Dias, 2000, 2002 and 2008). Hence, the mean differences in the outcomes of these individuals are understood to be the effects of the treatment (Caliendo and Hujer, 2005).

There are various methods of randomisation: oversubscription, randomised phase-in, within-group randomisation and encouragement designs. These methods are not discussed in depth here but are outlined in detail by Duflo, Glennerster and Kremer (2008) and Khandker, Koolwal and Samad (2010).

Many scientists believe that randomisation is the only method that can convincingly establish causality (Imbens and Wooldridge, 2008). It is claimed that social experiments provide an accurate counterfactual and control for self-selection bias provided that the experiment is properly implemented and that individuals are randomly allocated to either treatment or control groups (Blundell and Costa Dias, 2008). Furthermore, the analysis of experimental data is usually rather simple. Researchers commonly analyse the differences in the mean values by treatment status. Alternatively, a regression-based approach can generate an unbiased estimator for the average treatment effect of a programme (Imbens and Wooldridge, 2008).

However, limitations exist in the case of randomised experiments, i.e. double-blinding, ethical issues, pseudo-random methods, attrition and the fact that behavioural changes caused by the experiment itself such as Hawthorne and John Henry effects cannot be ruled out. Also, spill-over effects cannot be eliminated (Blundell and Costa Dias, 2000 and 2002).

In more detail, Imbens and Wooldridge (2008) argue

> "that even randomized experiments rely to some extent on substantive knowledge. It is only once the researcher is willing to limit interactions between units that randomization can establish causal effects" (p. 15).

The authors further claim that the identification problem that generally occurs when trying to establish causality cannot be solved by randomisation alone; in particular when interactions between individuals or units are prevalent which is often the case as discussed earlier. In many medical studies, such interactions are limited or non-existing and hence randomised studies are an appropriate choice providing robust results. The so-called double-blinding can commonly be enforced in medical studies, i.e. individuals participating in the experiment are generally not aware of their treatment status. This further enhances the robustness of the studies' results as well as improves external validity. However, double-blinding cannot necessarily be guaranteed in social science experiments and this raises serious concerns about the external validity of the results (Imbens and Wooldridge, 2008).

Scriven (2008) emphasizes that double-blinding is a prerequisite for a robust RCT; this is further reiterated by Goldacre (2008). As argued by Imbens and Wooldridge (2008), most medical RCTs can ensure double-blinding but the case is different for studies in the area of the social sciences. For example, RCTs evaluating the impact of education, social services or microfinance programmes are usually not even single-blinded and essentially 'zero-blinded' (Scriven, 2008, p. 12). In other words, individuals usually discover whether they belong to treatment or control groups, which undermines the notion of double-blindedness. Hence, individuals in the treatment group may benefit from the programme

> "due either to the experimental treatment, or to the sum of that effect plus the effect of any interaction of that treatment with the psychological impact of knowing that one is part of an experiment…" (Scriven, 2008, p. 14).

If non-interaction can be assumed, then the benefits reaped are due to the experimental treatment alone. However, RCTs in the social sciences generally do not assume non-interaction and hence the challenge to separate out the causal effects of a programme from all the other factors that occur at the same time remains (Scriven, 2008).

In addition, ethical questions are raised (Imbens, 2009). The implementation of randomised studies is not always feasible, e.g. on which grounds can it be justified that certain individuals are assigned to treatment while others are excluded from a potentially beneficial treatment. However, it could be argued that these ethical concerns are not valid considering treatment will eventually become available to individuals in the control groups after a certain time delay.

Furthermore, Goldacre (2008) argues that pseudo-random methods are often used during the process of assigning individuals to the various treatment and control groups. It pays to investigate how exactly individuals were assigned to their respective groups; was the underlying process truly random? Many studies fail to accurately describe their assignment process. This can have consequences for the reliability of the estimates obtained from a RCT.

Blundell and Costa Dias (2008) add to the limitations as follows:

> "[F]irst, by excluding the selection behaviour, experiments overlook intention to treat. However, the selection mechanism is expected to be strongly determined by the returns to treatment. In such case, the experimental results cannot be generalized to an economy-wide implementation of the treatment. Second, a number of contaminating factors may interfere with the quality of information, affecting the experimental results. One possible problem concerns drop-out behavior" (p. 19).

Drop-out behaviour - or attrition - refers to individuals that have been assigned to either treatment or control groups but then decide not to proceed with the experiment. It is not clear why those individuals drop out and this behaviour can have adverse effects on the results of the experiment (Blundell and Costa Dias, 2008). Goldacre (2008) and Duflo, Glennerster and Kremer (2008) argue that the individuals dropping out would have been worse off than the ones remaining and hence a risk of overstating impact estimates exists. To sum up, drop-outs change the composition of treatment and control groups thereby influencing the results of the experiment since their outcomes cannot be observed (Blundell and Costa Dias, 2008). Duflo, Glennerster and Kremer (2008) argue that attrition can be managed by tracking individuals that dropped out which would allow gathering information on them. However, this is a costly undertaking and might not always be feasible. More importantly, all randomised studies should report the level of attrition and compare drop-outs with the individuals that remain in the study to gauge whether there are systematic differences between these two groups – at least in terms of observable characteristics (Duflo, Glennerster and Kremer, 2008).

Duflo, Glennerster and Kremer (2008) further argue that the generalisation and replication of randomised studies is further hampered by behavioural changes in the treatment and control groups. To give an example, Hawthorne effects[13] refer to behavioural changes in the treatment group while John Henry effects[14] relate to behavioural changes in the control group. For example, individuals in the treatment group might positively change their behaviour for the duration of the study as they feel thankful for receiving treatment and as a response to being observed. The same behavioural changes might apply to members in the control group who might positively or in fact negatively alter their behaviour (Duflo, Glennerster and Kremer, 2008). However, a recent study by Levitt and List (2009) raises doubts about the existence of these Hawthorne effects. The authors' claim that the evidence is not as convincing as previously thought. In fact, it cannot be said with certainty that changes in lightning led to an increase in workers' productivity. According to Duflo, Glennerster and Kremer (2008), Hawthorne and John Henry effects can be circumvented by continuing to collect data even after the termination of the experiment to confirm whether any behavioural changes were due to Hawthorne or John Henry effects or due to the intervention itself.

As already mentioned earlier by Blundell and Costa Dias (2000 and 2002), Duflo, Glennerster and Kremer (2008) reiterate that spill-over effects can have adverse effects on the impact estimates obtained from a randomised study. Spill-over effects refer to individuals in the control groups that are affected by the treatment in physical ways or in the form of prices changes, learning or imitation effects. If spill-over effects are expected to be significant, then the experimental design can account for them. For

[13] In 1924 a series of experiments were conducted in the Hawthorne plant belonging to the Western Electric Company of Chicago. The aim of those experiments was to find out whether the productivity of workers could be improved with better lightening in the plant. Researchers found that workers increased their productivity irrespective of the lightening conditions which led to the conclusion that workers made an extra effort during their work precisely because of the knowledge of being observed (Duflo, Glennerster and Kremer, 2008; Levitt and List, 2009).

[14] John Henry effects refer to the "Ballad of John Henry" who was a rail worker and American folk hero. The ballad tells a tale of competition between rail workers and technical innovation which ultimately replaced rail workers. This tale can be related to the case of experimental design as discussed in Saretsky (1975).

example, the level of treatment exposure within groups can be adjusted to assess the magnitude of these spill-over effects (Duflo, Glennerster and Kremer, 2008).

These issues do not exhaust the limitations of experimental studies yet. In addition, extensive cooperation from the programmes that are being evaluated is required. This is time and cost intensive. Researchers need to obtain the institution's consent for randomising the implementation of their microfinance services (Montgomery, 2005). Moreover, for an experiment to work, the environment needs to be rigorously controlled, so that any difference in outcomes between the two groups can be adequately attributed to the impact of the intervention (Ledgerwood, 1999).

3.5.3. Critique of randomisation in development

The discussion so far has shown that there are threats to the internal and external validity of randomised studies, i.e. can the estimated impact be attributed to a particular intervention? Technical deficiencies such as a lack of double-blinding, pseudo-random methods as well as issues such as attrition and spill-over effects question the internal validity of experiments while Hawthorne and John Henry effects commonly have consequences for the external validity.

Hence, is the recent drive towards RCTs just a fad or here to stay? To recap, the 'randomistas' (Banerjee et al, 2009; Duflo and Kremer, 2005; Miguel and Kremer, 2004) claim that randomised experiments control for the issue of selection bias. Critics (Deaton, 2009; Heckman and Urzua, 2009; Imbens, 2009, Scriven, 2008 and Pritchett, 2009), however, argue that there are doubts about the adequate execution of randomised studies. As discussed earlier, in many cases pseudo-random methods are applied and respondents and implementers in fact know the treatment and control group status of participants, i.e. double-blinding is generally not ensured, or indeed possible. Moreover, issues of attrition are commonly neglected and drop-outs would have done worse than the ones remaining in the study. Consequently, not accounting for drop-outs tends to overstate impact estimates (Scriven, 2008). There are also Hawthorne and John Henry effects to consider which weaken the validity of RCTs especially in the absence of blinding. Moreover, ethical questions are often raised. Thus, applying experimental methods is in many cases not desirable or feasible. Moreover, Heckman and Vytlacil (2007a) emphasize that

> "even under ideal conditions, randomization cannot answer some very basic questions such as what proportion of a population benefits from a program. And in practice, contamination and cross-over effects make randomization a far from sure-fire solution even for constructing ATE." (p. 4836).

RnM argue that the present drive towards encouraging RCTs also renews calls for taking a closer look at the value of observational studies which collect data through non-random processes. Observational studies are not uncontested either as there are threats to both internal and external validity that arise in observational data as well. There is a risk of confounding, i.e. confounding variables are both related to the outcome that is being measured and the exposure. Typically, observational data require the application of more complex econometric techniques, i.e. PSM, IV and DID estimations. However, many of these econometric techniques cannot deal adequately with selection bias due to unobservable characteristics as the next section will argue.

3.6. Non-experimental approaches

The discussion of experimental versus non-experimental approaches reveals that evaluation results heavily depend on the quality of the underlying data (Heckman, LaLonde and Smith, 1999). Data quality, for example, refers to the availability of a rich set of appropriate variables that are related to participation as well as outcomes (Smith and Todd, 2005). Also, data on control groups located in the same environment as the treatment groups greatly improve quality (Heckman et al, 1998). Many evaluations in the past provided not particularly meaningful results precisely because of the non-availability of rich data sets (Caliendo and Hujer, 2005). Caliendo and Hujer (2005) further argue that researchers have no control over the origination of the data in the case of observational studies and can thus only observe outcomes for participants and non-participants after the intervention was implemented. In other words, the task of non-experimental techniques is often to restore comparability between treatment and control groups to allow solving the evaluation problem. Rosenbaum (2002) expands on this and argues that researchers should ideally be involved in the design stages of an observational study and take part in the data collection process in order to be able to avoid many of the pitfalls that later transpire in the analytical process. Rosenbaum (2002) further argues that ethnographic or other qualitative tools can be of great help to design an observational study. In a paper he published with Silber in 2001, he emphasizes the importance of making use of qualitative tools in particular in the design or pilot stages of a study to improve data collection procedures and hence the overall quality of the quantitative study. Well-known statisticians and econometricians such as Heckman, LaLonde and Rosenbaum advocate the collection of better quality data since this could possibly be the solution to the evaluation problem and not necessarily the introduction of further even more sophisticated evaluation techniques (Heckman, LaLonde and Smith, 1999; Rosenbaum, 2002; Rosenbaum and Silber, 2001; Caliendo and Hujer, 2005). Caliendo and Hujer (2005) further reiterate

> "that the problem of selection bias is a severe one and cannot be solved with more data, since the fundamental evaluation problem will not disappear" (p. 6).

Subsequent chapters will support this view and illustrate that not more data but better data is in fact needed.

Furthermore, two main aspects of non-experimental techniques can be observed according to Caliendo and Hujer (2005): first, the choice of the technique depends on the characteristics of the underlying data and whether it is cross-sectional or longitudinal. Secondly, each of these techniques manages the issue of selection bias differently. Some techniques depend on the assumption that selection on observables or unconfoundedness is observed and hence techniques such as matching, regression approaches or RDDs are appropriate. However, in the case of poor data quality, this assumption might not hold and thus approaches such as IVs, selection models and DID are more suitable since they also allow for selection on unobservables (Caliendo and Hujer, 2005). The following sections outline some of the methods commonly used in the evaluation arena which all aim to construct a counterfactual outcome by applying different identification strategies. As discussed in section 3.5.1., Dehejia and Wahba (1999 and 2002) conclude that PSM results are a good approximation to those obtained under an experimental approach. In addition, PSM is the technique of choice in subsequent chapters and hence I start with introducing PSM. Later sections outline IVs, selection models and DID approaches which claim to control for selection on unobservables. In addition, RDDs are introduced which are essentially a variant of IVs but only control for selection on observables, as does PSM.

3.6.1. Propensity score matching

Matching has become a very popular technique in the area of development economics in recent years and has its roots in the experimental literature beginning with Neyman (1923). Rubin (1973a, 1973b, 1974, 1977 and 1978) expands on this literature and essentially laid the conceptual foundations of matching. The technique was further refined in particular by Rosenbaum and Rubin (1983 and 1984). Econometricians got involved in advancing matching techniques in the mid 1990s; see studies by Heckman, Ichimura and Todd (1997, 1998), Heckman et al (1998) and Heckman, LaLonde and Smith (1999).

The basic idea of matching is to compare a participant with one or more non-participants who are similar in terms of a set of observed covariates X (Caliendo and Kopeinig, 2005 and 2008; Rosenbaum and Silber, 2001). In a next step, the differences in the outcome variables for participants and their matched non-participants are calculated, i.e. the ATT is the mean difference between participants and matched non-participants (Morgan and Harding, 2006). The objective of this method is to account for selection on observables. The drawback of this technique is that selection on unobservables remains unaccounted.

Caliendo and Kopeinig (2008) explain that

> "since conditioning on all relevant covariates is limited in the case of a high dimensional vector X ('curse of dimensionality'), Rosenbaum and Rubin (1983) suggest the use of so-called balancing scores $b(X)$, i.e. functions of the relevant observed covariates X such that the conditional distribution of X given $b(X)$ is independent of assignment into treatment. One possible balancing score is the propensity score, i.e. the probability of participating in a programme given observed characteristics X" (p. 32).

To begin with, the main assumptions of PSM are outlined below before proceeding to discuss the challenges that occur during the implementation of PSM.

3.6.1.1. Theory

The central assumption of PSM that needs to be observed is referred to as the Conditional Independence Assumption (CIA) or unconfoundedness. This assumption is denoted as follows, where the notation is taken from Heckman et al (1998) and Caliendo (2006) who in turn have taken it from Dawid (1979)[15]:

(13) $$Y^0, Y^1 \perp\!\!\!\perp D \mid X \quad \text{(Unconfoundedness)}$$

Where $\perp\!\!\!\perp$ represents independence. If this is correct, it follows that

(14) $$E(Y^0 | X, D = 1) = E(Y^0 | X, D = 0)$$

and

(15) $$E(Y^1 | X, D = 1) = E(Y^1 | X, D = 0)$$

which implies that the outcomes of non-participants would have the same distribution as the outcomes of participants had they not participated given conditionality on X (Caliendo, 2006; Caliendo and Hujer, 2005). Caliendo (2006) explains that

> "… matching balances the distributions of all relevant, pre-treatment characteristics X in the treatment and comparison group" (p. 31)

which makes it comparable to a randomised approach. As a result of this, independence between potential outcomes and treatment assignment is accomplished. Assuming the following holds

(16) $$E(Y^0 | X, D = 1) = E(Y^0 | X, D = 0) = E(Y^0 | X)$$

and

(17) $$E(Y^1 | X, D = 1) = E(Y^1 | X, D = 0) = E(Y^1 | X)$$

then the counterfactual outcomes can be deduced from the outcomes obtained from participants and non-participants.

In addition, the assumption of common support or overlap will have to be met and applies to all X (Caliendo, 2006; Caliendo and Hujer, 2005):

(18) $$0 < \Pr(D = 1 \mid X) < 1 \qquad \text{(Overlap)}$$

According to Caliendo (2006), this assumption of overlap indicates that treatment and control groups provide equal support of X. It further ensures that X is not a perfect predictor that identifies a corresponding match for each participant from the pool of

[15] The remaining notations in section 3.6.1.1. follow Caliendo and Hujer (2005).

non-participants and the other way round. The literature encourages matching over the common support region only when

> "… there are regions where the support of X does not overlap for the treated and non-treated individuals …" (Caliendo, 2006, p. 31).

Rosenbaum and Rubin (1983) introduce the term 'strong ignorability' which applies when CIA can be maintained and when there is in fact overlap between treatment and control groups. If 'strong ignorability' is the case, then ATE and ATT as described in equations (3) and (4) respectively can provide valid estimates for all X. However, the notion of 'strong ignorability' is often difficult to observe in practice and can be relaxed to a certain degree when the focus is on estimating ATT only. In this case, it is sufficient to assume $Y^0 \amalg D \mid X$ and $P(D = 1 \mid X) < 1$ and hence ATT is denoted as follows (Caliendo and Hujer, 2005):

$$\Delta_{ATT}^{MAT} = E(Y^1|X, D = 1) - E_x[E(Y^0|X, D = 0)| D = 1] \tag{19}$$

Where $E(Y^1|X, D = 1)$ calculates the mean outcomes of treated individuals and $E_x[E(Y^0|X, D = 0)| D = 1]$ provides the calculation for the matches from the control group (Caliendo and Hujer, 2005). The treatment effects can be estimated by comparing the mean outcomes of the matches; the differences obtained are the estimates of the programme impact for these particular observations (Ravallion, 2001).

3.6.1.2. PSM in practice

How does PSM work in practice? Implementing PSM essentially requires the researcher to follow the 6 steps outlined below (Caliendo and Kopeinig, 2005 and 2008):

1) Choice of statistical model

2) Estimation of propensity scores

3) Choice of matching algorithm

4) Check of overlap/common support region

5) Check of quality of matches

6) Sensitivity analysis

3.6.1.2.1. Model

To recap, PSM is performed by matching participants to non-participants based on the predicted probability of programme participation or the "propensity score" (Ravallion, 2001). Hence, for step 1 and 2, a representative sample of participants and non-participants has to be sampled which are then pooled to estimate programme participation using a logit or probit model (Caliendo and Kopeinig, 2005 and 2008). The choice of the model is straightforward in the case of binary treatments while in the case of multi-valued treatments multinomial logit or probit models or a series of binary logit or probit models should be used (Caliendo and Kopeinig, 2005 and 2008). The predicted values obtained from the logit or probit models are then the propensity scores. Each of the sampled participants and non-participants has a propensity score.

As mentioned earlier, the CIA must be satisfied when choosing explanatory variables, i.e.

> "outcome variable(s) must be independent of treatment conditional on the propensity score", (Caliendo and Kopeinig, 2008, p. 38).

In fact the choice of the appropriate explanatory variables is crucial for the quality of the matches (Smith and Todd, 2005). When designing the model, a researcher will have to consider all variables that could determine participation as well as potential outcomes. Caliendo and Kopeinig (2005 and 2008) argue that economic theory and previous empirical findings should drive the choice of these variables. Various search procedures are available to help build a model such as the 'hit or miss' method as advocated by Heckman et al (1998) or the 'leave-one-out cross-validation' method (Caliendo and Kopeinig, 2005 and 2008) just to name a few. To recap, Smith and Todd (2005) emphasize that PSM only provides robust estimates when a rich set of variables, i.e. variables that provide information related to participation as well as outcomes, are included in the model. The authors, along with Heckman et al (1998), further argue that the control group will have to be sampled from the same environment as the treatment group. In addition, Smith and Todd (2005) stress that the dependent variable will have to be measured in exactly the same way across both groups. If these conditions are maintained, then matching can be a success (Smith and Todd, 2005).

3.6.1.2.2. Algorithm

In a next step, a matching algorithm has to be chosen. Participants are matched to one or more non-participants with similar propensity scores by applying a matching algorithm. In other words, the role of these matching algorithms is to select appropriate matches for each household or individual. The literature in this area is not yet very developed. Various algorithms exist such as nearest neighbour matching, caliper and radius matching, stratification and interval matching, kernel and local linear matching and weighting but their reliability varies and there are no clear recommendations as to which matching algorithm should be preferred. Heckman, Ichimura and Todd (1997 and 1998), Smith and Todd (2005), Morgan and Winship (2007), Rosenbaum (2002), Rosenbaum and Silber (2001), Gu and Rosenbaum (1993) have discussed the topic of matching algorithms in depth. Generally speaking, the two main differences between the various matching algorithms are related to

> "… (1) the number of matched cases designated for each to-be-matched target case and (2) how multiple matched cases are weighted if more than one is used for each target case" (Morgan and Harding, 2006, p. 29-30).

This section only provides a brief overview of the main distinguishing features of selected algorithms, i.e. nearest neighbour matching, interval matching and kernel matching.

According to Caliendo and Kopeinig (2005 and 2008), nearest neighbour matching is the most easily understood matching algorithm. It works by matching an individual or household from the treatment group with an individual or household from the control group. The basis for the matching is the closeness of their respective propensity scores. Nearest neighbor matching is possible 'with' and 'without replacement'. 'With replacement' is recommended as it increases the quality of the matches (Caliendo and Kopeinig, 2005 and 2008; Morgan and Winship, 2007). It requires that individuals or households from the control group are used more than once for the matching procedure. In the case of 'without replacement', individuals or households are only used once. This matching algorithm has a risk of producing bad matches, i.e. the distance to the closest neighbour is large. Hence, caliper matching is commonly preferred since a caliper or tolerance level for the maximum distance of the propensity scores is imposed. This has a similar effect to running nearest neighbour matching 'with replacement' and increases the quality of the matches (Caliendo and Kopeinig, 2005 and 2008).

Interval matching works by splitting the common support of the propensity scores into intervals. The impact is then calculated separately in each of the intervals by comparing the mean difference of the outcome variables between participants and non-participants (Caliendo and Kopeinig, 2005 and 2008; Rosenbaum and Rubin, 1983).

Generally speaking, most of the matching algorithms construct a counterfactual by using only a few observations from the control group. Kernel matching, however, works slightly different, i.e. weighted averages of individuals or households in the control group are used to construct a counterfactual (Caliendo and Kopeinig, 2005 and 2008). Smith and Todd (2005) describe kernel matching as a weighted regression approach where weights depend on the distance of the propensity scores between treated and untreated individuals or households.

Finally, Smith and Todd (2005) find that the quality of the matches is unaffected by the matching algorithms employed. In other words, the various matching algorithms do not make much of a difference in terms of producing unbiased impact estimates. What makes a difference is the quality and availability of the underlying data, i.e. rich and high quality data are required (Smith and Todd, 2005).

3.6.1.2.3. PSM in microfinance

In the few microfinance studies that employ PSM (see studies by Abou-Ali et al, 2009; Arun, Imai and Sinha, 2006; Chemin, 2008; Deininger and Liu, 2009; Imai, Arun and Annim, 2010; Imai and Azam, 2010; Setboonsarng and Parpiev, 2008; Takahashi, Higashikata and Tsukada, 2010), nearest neighbour matching, kernel matching as well as stratification matching are preferred. Heckman, Ichimura and Todd (1997 and 1998), Heckman et al (1998) and Smith and Todd (2005) recommend kernel matching as its results appear to be the most robust. In summary, there is no best practice approach with regard to choosing a matching algorithm; it is recommended to try a number of algorithms since their performance varies from case to case and heavily depends on the underlying data structure (Zhao, 2003) as well as the statistical package that is being used (Morgan and Winship, 2007)[16]. It is worrying that the various algorithms and software packages produce slightly contradictory results but one can only hope that this will improve as work on matching algorithms and statistical programming continues.

3.6.1.2.4. Common support

As to step 4, setting a common support region implies that some non-participants will have to be excluded since their propensity scores are outside the range found for participants. Failure to do so will lead to biases (Caliendo and Kopeinig, 2005 and 2008). For example, individuals in the treatment group will only be compared to individuals in the control group when their propensity scores are close enough and within the common support region. Treated individuals with high propensity scores shall be excluded from the analysis since insufficient matches from non-treated individuals are available. Treatment effects can only be measured within the common support region. Caliendo and Kopeinig (2005 and 2008) argue that there are two different methods that assist researchers in defining the common support region more accurately, the minima and maxima comparison and the trimming procedure. However, it is recommended to start off with a visual analysis of the distribution of the propensity scores before applying these two methods (Caliendo and Kopeinig, 2005 and 2008). Defining an accurate common support region is crucial because attempts to maximise the number of exact matches by narrowing down the common support region may lead to incomplete matching since cases might be excluded. On the other hand, a widening of the common support region may lead to a maximisation of cases but may result in inexact matching (Caliendo and Kopeinig, 2005 and 2008).

3.6.1.2.5. Quality of matches

In a next step, researchers will have to assess the quality of their matches

> "since we do not condition on all covariates but on the propensity score, it has to be checked if the matching procedure is able to balance the distribution of the

[16] STATA routines that allow running PSM are pscore developed by Becker and Ichino (2002); psmatch2 developed by Leuven and Sianesi (2003) and nnmatch developed by Abadie et al (2004). Results vary depending on the matching algorithm and PSM routine applied; this is briefly discussed in Morgan and Winship (2007).

relevant variables in both the control and treatment group" (Caliendo and Kopeinig, 2008, p. 47).

Caliendo and Kopeinig (2005 and 2008) briefly outline the various procedures that are available to assess matching quality, e.g. t-tests and stratification tests to investigate whether the mean outcome values for both treatment and control groups are significantly different from each other. While these tests are useful, they cannot help to improve the quality of the matches. In order to improve the quality of the matches one essentially has to revisit the model that was originally used to estimate the propensity scores and re-think its specification to improve matching quality (Caliendo and Kopeinig, 2005 and 2008).

3.6.1.2.6. Sensitivity analysis

A final step in the matching process is to test the sensitivity of the results obtained with respect to selection on unobservables. As discussed earlier, PSM was not designed to control for selection on unobservables (Smith and Todd, 2005). In fact, the technique is heavily dependent on the CIA or unconfoundedness assumption, i.e. selection on observable characteristics (Caliendo and Kopeinig, 2005 and 2008). This assumption needs to be maintained in order to produce unbiased PSM estimates. Selection on unobservables or 'hidden bias' as Rosenbaum (2002) calls them exist without a doubt. They are driven by unobserved variables that influence treatment decisions as well as potential outcomes (Becker and Caliendo, 2007). Matching estimators are commonly not robust enough to deal with selection on unobservables.

Hence, to test the likelihood that one or more unobservables could play a role in selection, which would explain unobserved differences, sensitivity analysis has become increasingly important. In other words, sensitivity analysis attempts to gauge how vulnerable the assignment process into treatment to unobservables is, and hence assess the quality of the matching estimates (Becker and Caliendo, 2007). Few approaches for sensitivity analyses have been developed, the most well-known method is the bounding approach developed by Rosenbaum (2002). Rosenbaum's (2002) approach does not directly assess the CIA itself but tests how sensitive the impact estimates are in view of a possible violation of this identifying assumption. In the case that the matching results are indeed sensitive to possible violations of CIA, alternative estimation strategies will have to be considered (Becker and Caliendo, 2007). Furthermore, Caliendo and Kopeinig (2005) introduce another approach which they call 'Lechner bounds' developed by Lechner (2000) which assesses the sensitivity of the results with regard to the common support or overlap assumption. Few studies have extensively dealt with sensitivity analysis; Rosenbaum (2002) provides an in depth discussion on the topic. Becker and Caliendo (2007), Ichino, Mealli and Nannicini (2006) and Nannicini (2007) provide further insights as well[17]. In the context of microfinance, none of the studies that applied PSM (e.g. Abou-Ali et al, 2009; Arun, Imai and Sinha, 2006; Chemin, 2008; Deininger and Liu, 2009; Imai, Arun and Annim, 2010; Imai and Azam, 2010; Setboonsarng and Parpiev, 2008; and Takahashi, Higashikata and Tsukada, 2010) used sensitivity analysis to assess the robustness of their matching estimates and hence chapters 4 and 5 further expand on this topic and apply sensitivity analysis.

3.6.1.3. PSM compared to randomised experiments: LaLonde, Dehejia & Wahba and Smith & Todd

Finally, an overview of PSM is not complete without mentioning the debate that questions the overall robustness of PSM results versus those obtained from randomised experiments. As mentioned in an earlier section, Dehejia and Wahba (1999 and 2002) – as a response to a study by LaLonde (1986) - tested how well PSM does in comparison to a randomized experiment and concluded that PSM's overall contribution towards reducing biases is positive. In other words, the results provided by PSM are a good approximation to those obtained under an experimental approach (as discussed in section 3.5.1.). Smith and Todd (2005) re-investigate the analysis of Dehejia and Wahba (1999 and 2002) and conclude that their (Dehejia and Wahba's) results

> "are highly sensitive to their choice of a particular subsample of LaLonde's (1986) data" (p. 306),

[17] Various STATA commands were developed to implement sensitivity analyses such as mhbounds (developed by Becker and Caliendo, 2007) which executes the bounding approach developed by Rosenbaum (2002) or sensatt (Nannicini, 2007) which executes the approach developed by Ichino, Mealli and Nannicini (2006).

and that

> "the changing [*Author's note: of*] the set of variables used to estimate the propensity scores strongly affects the estimated bias in LaLonde's original sample" (p. 306-307).

Smith and Todd (2005) further argue that the DID estimator they employed performed much better with regard to eliminating biases than Dehejia and Wahba's (1999 and 2002) PSM approach. In fact, combining PSM with DID estimators (Khandker, Koolwal and Samad, 2010) is a popular approach which will be discussed in a later section in this chapter. Smith and Todd (2005) conclude that

> "[I]nstead of engaging in a hopeless search for a magic bullet estimator, the goal of theoretical and empirical investigation should be to develop a mapping from the characteristics of the data and institutions available in particular evaluation contexts to the optimal nonexperimental estimator for those contexts" (p. 347).

To conclude, the quality of the underlying data is what matters (Bryson, Dorsett and Purdon, 2002). Matching is a good choice when high quality data sets are available but might not be an appropriate evaluation strategy if that is not the case (Smith and Todd, 2005). Dehejia (2005) responds to the claims made by Smith and Todd (2005) and argues that Smith and Todd (2005) essentially made some mistakes in their re-analysis of Dehejia and Wahba (1999 and 2002). It is beyond the scope of this chapter to outline this debate in depth. In brief, Dehejia (2005) concludes that PSM is indeed not the panacea for solving the evaluation problem but it has never claimed otherwise. Furthermore, Dehejia (2005) points out that the correct specification of the propensity score is crucial, i.e. the balancing properties of the propensity score should be satisfied – as emphasized by Caliendo and Kopeinig (2005 and 2008) and Smith and Todd (2005) – and that the sensitivity of the results require testing – as advocated by Rosenbaum (2002), Becker and Caliendo (2007), Ichino, Mealli and Nannicini (2006) and Nannicini (2007).

This line of thought is developed later in the empirical chapters which expand on the challenges of employing PSM, its potential drawbacks and its implications for the robustness of the impact estimates obtained. It is argued that selection on unobservables that drive the selection or screening process in the context of microfinance cannot be controlled for by the application of PSM. Hence, the next section turns its attention to a technique that claims to control for these unobservables, i.e. IVs.

3.6.2. Instrumental variables

The IV approach is widely used in the evaluation arena and claims to control for selection on observables as well as unobservables (Heckman and Vytlacil, 2007b; Basu et al, 2007) which is in contrast to PSM which tries to construct an appropriate set of counterfactual cases to counteract selection on observables only. The main goal of the IV method is to identify a variable or a set of variables, i.e. instruments, that influence the decision to participate in a programme but at the same time do not have an effect on the outcome equation. Only when adequate instruments can be identified, then the IV approach is an effective strategy for estimating causal effects (Morgan and Winship, 2007). In the words of Caliendo (2006),

> "the instrumental variable affects the observed outcome only indirectly through the participation decision and hence causal effects can be identified through a variation in this instrumental variable" (p. 25).

Imbens and Angrist (1994); Angrist, Imbens and Rubin (1996); Heckman (1997); Angrist and Krueger (2001); Basu et al (2007); Heckman and Vytlacil (2007b); Heckman and Urzua (2009) and Imbens (2009) have discussed IV approaches in depth. Hence, this section only briefly outlines the IV method and provides an overview of the challenges using this approach.

Basically, a regressor qualifies as an instrument for Z*, which represents programme participation, when it is uncorrelated with the error terms and is not entirely influenced by X (Caliendo, 2006; Caliendo and Hujer, 2005). The IV estimator for a binary instrument $Z^* \in \{0,1\}$ can thus be denoted as follows[18]:

$$\Delta^{IV} = \frac{E(Y|X, Z* = 1) - E(Y|X, Z* = 0)}{P(D = 1|X, Z* = 1) - P(D = 1|X, Z* = 0)} \qquad (20)$$

The main challenge of the IV method is to identify an adequate instrument which influences programme participation but at the same time does not influence the

---

[18] Notation for equation (20) follows Caliendo and Hujer (2005).

outcome equation. In the words of Imbens and Angrist (1994), the variable Z* should be

> "…independent of the responses Y_i^0 and Y_i^1, and correlated with the participation indicator D_i" (p. 468).

In many cases weak instruments are employed which can have adverse effects on the accuracy of IV estimates (Caliendo, 2006; Caliendo and Hujer, 2005).

As mentioned earlier, there are problems with the implementation of IV methods in the case of heterogeneous treatment effects. For illustrating this point, Caliendo (2006) suggests returning to equation (11) where the following error term was provided: $[U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)]$. It is assumed that there is no correlation between Z* and U_{it} but the same assumption cannot hold for $[U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)]$ precisely because D_i is determined by Z* and hence applying IV in the context of heterogeneous treatment effects can be ineffective (Caliendo, 2006; Caliendo and Hujer, 2005; Blundell and Costa Dias, 2000; Heckman and Vytlacil, 2007b).

However, despite the drawbacks of IVs, Imbens and Angrist (1994) argue that IVs can in fact identify adequate treatment parameters even in the case of heterogeneous treatment effects when certain conditions are met. The authors develop an alternative parameter which they term the LATE that

> "identifies the treatment effect for those individuals (with characteristics X) who are induced to change behaviour because of a change in the instrument" (Caliendo and Hujer, 2005, p. 10).

LATE is essentially an approximation of the MTE which was introduced by Björklund and Moffitt (1987) and further developed by Heckman and Vytlacil (1999, 2000, 2001, 2005, 2007a and 2007b). The MTE applies to continuous instruments and is defined by Heckman and Vytlacil (2007b) as

> "…a choice-theoretic parameter that can be interpreted as a willingness to pay parameter for persons at a margin of indifference between participating in an activity or not" (p. 4878).

The authors further argue that most of the standard treatment parameters can be derived from the MTE since all evaluation methods such as PSM, IVs, selection models and RDDs make assumptions about this particular parameter (Heckman and Vytlacil, 2007b). Basu et al (2007) suggest alternatives to the application of conventional IV estimates that expand on the IV-LATE approach and argue that these alternatives would be better suited to deal with heterogeneity and self-selection bias. At the same time, the authors stress that care should be exercised about the choice of the instruments (Basu et al, 2007, p. 1154).

Overall, it can be concluded that IV estimates are only as good as the underlying instruments they employ. Heckman and Vytlacil (2007b) argue that IV estimates are not necessarily better than simple ordinary least square (OLS) estimates; they might even be more biased.

3.6.3. Regression discontinuity designs

The RDD is essentially a variant of the IV approach and was first established by Thistlethwaite and Campbell (1960) and Campbell (1969). The RD approach assesses causal effects by introducing a cut-off point or discontinuity that determines programme participation. In more general terms, RDDs follow a deterministic rule, i.e. $D = 1$ if $Z < \acute{Z}$ and $D = 0$ if $Z > \acute{Z}$ where Z is an observed variable which shapes the decision of individuals to select into a programme or not (Heckman, LaLonde and Smith, 1999). This is the case of a 'sharp' design, which is the ideal case but cannot often be observed in practice (Blundell and Costa Dias, 2008). The more common case is the 'fuzzy' design where the decision to participate in a programme is not an entirely deterministic function of Z (Heckman, LaLonde and Smith, 1999). In a 'fuzzy' design, participants and non-participants exist on either side of the threshold of variable Z and discontinuity cannot be observed (Blundell and Costa Dias, 2008). The variable Z is central since it has an effect on outcome variable Y directly as well as indirectly through D. The indirect effect through D is the causal effect that is of interest and shall be assessed. The publication by Hahn, Todd and van der Klaauw (2001) is widely cited and discusses RDDs in the context of LATE, while Heckman and Vytlacil (2007b) illuminate the RDD from the perspective of MTE. For a more practical guide, see Imbens and Lemieux (2008).

PnK use landownership as the variable Z to determine whether a household is eligible to participate in microcredit. If $Z < \acute{Z}$, then D = 1 and D = 0 if not. In other words, households are eligible to participate in the programme if they own less than 0.5 acres of land and are not eligible if they own more than 0.5 acres. However, there is a debate surrounding the study of PnK based on that argument that their RDD was not strictly enforced and that the underlying design was ‘fuzzy’ (Chemin, 2008). Furthermore, Ravallion (2008) argues that the study by PnK followed a DID approach (discussed in more depth in chapter 5).

It is commonly argued that the ‘sharp’ design is able to control for selection on observables (Blundell and Costa Dias, 2008). This does not apply in the event of a ‘fuzzy’ design where the problem becomes one of selection on unobservables since participation occurs for various levels of the eligibility variable whose values reflect the unobservables. In the study by PnK land cultivated plays this role, but the ‘unobservables’ are inferred to be quality of land, so that the true variable is value of land equivalent to the value of 0.5 acres of land of average value (this point is further developed and explained in chapter 5).

Much of the appeal and simplicity of RDDs disappear when the criterion is ‘fuzzy’ (Heckman, LaLonde and Smith, 1999; Blundell and Costa Dias, 2008), and other methods that deal with selection on unobservables such as IV, selection models or DID will have to be considered. Heckman, LaLonde and Smith (1999) point out that non-participation by individuals who are in fact eligible, i.e. their values of Z satisfy the eligibility rules, is a serious concern in RDDs since unbiased estimates of the mean impact of treatment of eligible individuals can no longer be obtained (Heckman, LaLonde and Smith, 1999).

3.6.3.1. Applications of regression discontinuity designs

Furthermore, earlier sections mentioned the studies conducted by Dehejia and Wahba (1999 and 2002) that re-investigated LaLonde’s (1986) study. They employed PSM claiming to illustrate that it can approximate the results obtained from an experimental study. Cook and Wong (2008) did something similar with RDDs applying within-study comparisons commonly associated with LaLonde (1986) to analyse how well the RD results compare to those obtained from a randomised experiment. They found “considerable correspondence between the experimental and RD results” (Cook and Wong, 2008, p. 32). The RD estimates Cook and Wong (2008) obtained, seemingly using a 'sharp' design, were generally robust and a good reproduction of the results of the experimental study (p. 33). A related inquiry was conducted by Green et al (2009).

In addition, further examples that are widely used in the literature to illustrate the application of RDDs are provided by Angrist and Lavy (1999) on schooling and Hahn, Todd and van der Klaauw (1999) on anti-discrimination laws of minority workers. In brief, Angrist and Lavy (1999) in a widely quoted paper, investigate the impact of class size on students' test scores in Israeli public schools. Israeli schools typically restrict their class size to 40 students. The authors argue that this cap is based on the so-called 'Maimonides' Rule' which was established by a rabbinic scholar called Maimonides in the 12th century. This rule states that a class is split into two if the number of its students exceeds 40, i.e. with the 41st student the class size drops to an average of 20.5 students. Angrist and Lavy (1999) exploit the 'Maimonides' Rule' as an eligibility criterion to estimate the impact of class size on test scores. Students that have been in a class that has been split into two qualify for treatment. As expected, the students' test scores are directly affected by class size, i.e. the smaller the size of the class, the higher the test scores of the students, in particular in their 4th and 5th grades. In addition, an indirect effect of class size through *D* can also be observed (Caliendo, 2006; Caliendo and Hujer, 2005). Hahn, Todd and van der Klaauw (1999) present another well-known example that assesses the impact of anti-discrimination laws of minority workers. These laws are only applicable to firms with more than 15 employees. This threshold functions as the discontinuity rule and hence allows assessing the causal effects of this intervention.

Finally, Heckman, LaLonde and Smith (1999) and Caliendo (2006) conclude that RDDs are adequate for controlling for selection on observables as long as a deterministic rule can strictly be observed. In other words, van der Klaauw (2002) argues that

> "the validity of the RD approach relies on an appropriate specification of the relationships between the selection and treatment variable, and between the selection and outcome variable"(p. 1284).

If this specification is not done correctly, RDDs fail to provide unbiased impact estimates of programme participation and selection on unobservables becomes an issue. This then requires a different evaluation strategy, e.g. matching or IVs. However, as mentioned earlier, matching requires common support for treated and non-treated individuals and if this requirement cannot be fulfilled, then matching is also not an appropriate strategy (Caliendo, 2006; Caliendo and Hujer, 2005). Moreover, an earlier section has shown that matching does not work well when there is selection on unobservables. Hence, the next section takes a closer look at selection models which claim to provide solutions to selection on unobservables.

3.6.4. Selection models

Selection models were first developed by Heckman (1974, 1976, 1978 and 1979) and are also referred to as the Heckman selection estimator. According to Caliendo (2006) and Caliendo and Hujer (2005), this method is more sophisticated and robust than the IV approach since more assumptions with regard to the structure of the underlying model are made.

According to Heckman and Vytlacil (2007a), selection models

> "do not start with the experiment as an ideal but start with well-posed, clearly articulated models for outcomes and treatment choice where the unobservables that underlie the selection and evaluation problem are made explicit" (p. 4835).

This is in contrast to the 'potential outcomes framework', which is strictly speaking and by definition, based on the ideal of randomisation (Heckman and Vytlacil, 2007a).

The first assumption of the Heckman selection estimator states that an additional regressor with a non-zero coefficient and with independence from the error term has to be part of the decision rule (Caliendo, 2006; Caliendo and Hujer, 2005; Blundell and Costa Dias, 2000). The second assumption asserts that the distribution densities of the error terms should be known or it should be possible to estimate them (Caliendo, 2006; Caliendo and Hujer, 2005; Blundell and Costa Dias, 2000).

It is assumed that the outcome equation contains an error term (U_{it}) of which a part is correlated with the dummy variable that represents programme participation (D_i) (Caliendo, 2006). The objective of selection models is to control for this part of the error term. The impact of programme participation is then assessed by estimating the correlated part of U_{it} and including it in the outcome equation (Caliendo, 2006). The result is an error term that is not correlated with D_i any longer (Blundell and Costa

Dias, 2000). Caliendo (2006) illustrates this point by expressing equation (10) in linear terms: $Y_i = X_i\beta_0 + \alpha D_i + U_i$. In addition, it is assumed that the error terms follow a joint normal distribution and hence the conditional outcome expectation can be written as follows (Caliendo, 2006; Caliendo and Hujer, 2005)[19]:

$$(21) \qquad E(Y_i|D_i = 1) \;=\; \beta + \alpha + \rho\frac{\theta\,(Z_{i\gamma})}{\Theta\,(Z_{i\gamma})}$$

and

$$(22) \qquad E(Y_i|D_i = 0) \;=\; \beta - \rho\frac{\theta\,(Z_{i\gamma})}{1-\Theta\,(Z_{i\gamma})}$$

Following Blundell and Costa Dias (2000), Caliendo (2006) reiterate that

> "the new regressor includes the part of the error term that is correlated with the decision process in the outcome equation, allowing us to separate the true impact of the treatment from the selection process. Thus it is possible to identify α as outlined above, by replacing γ with γ^ and running a least-squares regression on the conditional outcome expectations" (p. 29).

Moreover, selection models appear to be suitable for estimating ATT even in the case of heterogeneous treatment effects (Blundell and Costa Dias, 2000). However, Puhani (2000) is critical of Heckman's selection estimator and runs Monte Carlo simulations to illustrate his point. He argues that selection models are mainly driven by narrow assumptions about functional form and error distributions, in particular when no exclusion restrictions are enforced. In other words, impact estimates obtained from selection models are only as good as these assumptions on distributional and functional form (Vytlacil, 2002). This last point is similar to IV estimates whose reliability heavily depends on the quality of the underlying instruments (Caliendo and Hujer, 2005). Overall, the econometric techniques introduced so far have all drawbacks in one way or another. Hence, the choice of the correct evaluation strategy heavily depends on the particularities of the programme that is being evaluated and the availability of data.

[19] Notations for equations (21) and (22) follow Caliendo (2006) and Caliendo and Hujer (2005).

3.6.5. Differences-in-differences

The DID method is an advance on the standard before and after approach which simply compares the outcomes of individuals before programme participation with the outcomes of the same individuals after participation (Bertrand, Duflo and Mullainathan, 2004). The DID approach is not only a simple before and after comparison but a combination of both before and after and with and without comparisons (Khandker, Koolwal and Samad, 2010). In other words,

> "the DID estimator eliminates common time trends by subtracting the before-after change in non-participant outcomes from the before-after change for participant outcomes" (Caliendo and Hujer, 2005, p. 11).

Longitudinal data is typically required for implementing DID methods but the approach can be adjusted to a cross-section approach (Khandker, Koolwal and Samad, 2010) as it will be seen later.

I first explain the DID approach in more technical language which is followed by an example from the area of microfinance to practically illustrate this method. Equation (23)[20] illustrates how the DID approach works. To begin with, mean values for outcome variable Y for participants and non-participants are calculated in pre-treatment period t'. These values are then compared with the ones obtained for participants and non-participants after the treatment represented by period t (Caliendo and Hujer, 2005). In other words, outcome variable Y for treated and non-treated individuals is compared over time (Heckman et al, 1998; Caliendo and Hujer, 2005, p. 11).

$$(23) \qquad \Delta^{DID} = [Y_t^1 - Y_{t'}^0 | \, D = 1] - \; [Y_t^0 - Y_{t'}^0 | \, D = 0]$$

Where the identifying assumption is expressed as follows:

$$(24) \qquad E(Y_t^0 - Y_{t'}^0 | \, D = 1) = \; E(Y_t^0 - Y_{t'}^0 | \, D = 0]$$

It is argued that DID controls for selection bias since it assumes "time-invariant linear selection effects" (Caliendo and Hujer, 2005, p. 11). For the sake of clarity, the outcome equation for an individual *i* in period *t* is re-arranged:

$$(25) \qquad Y_{it} = \; \pi_{it} + D_{it} * Y_{it}^1 + \; (1 - D_{it}) * \; Y_{it}^0$$

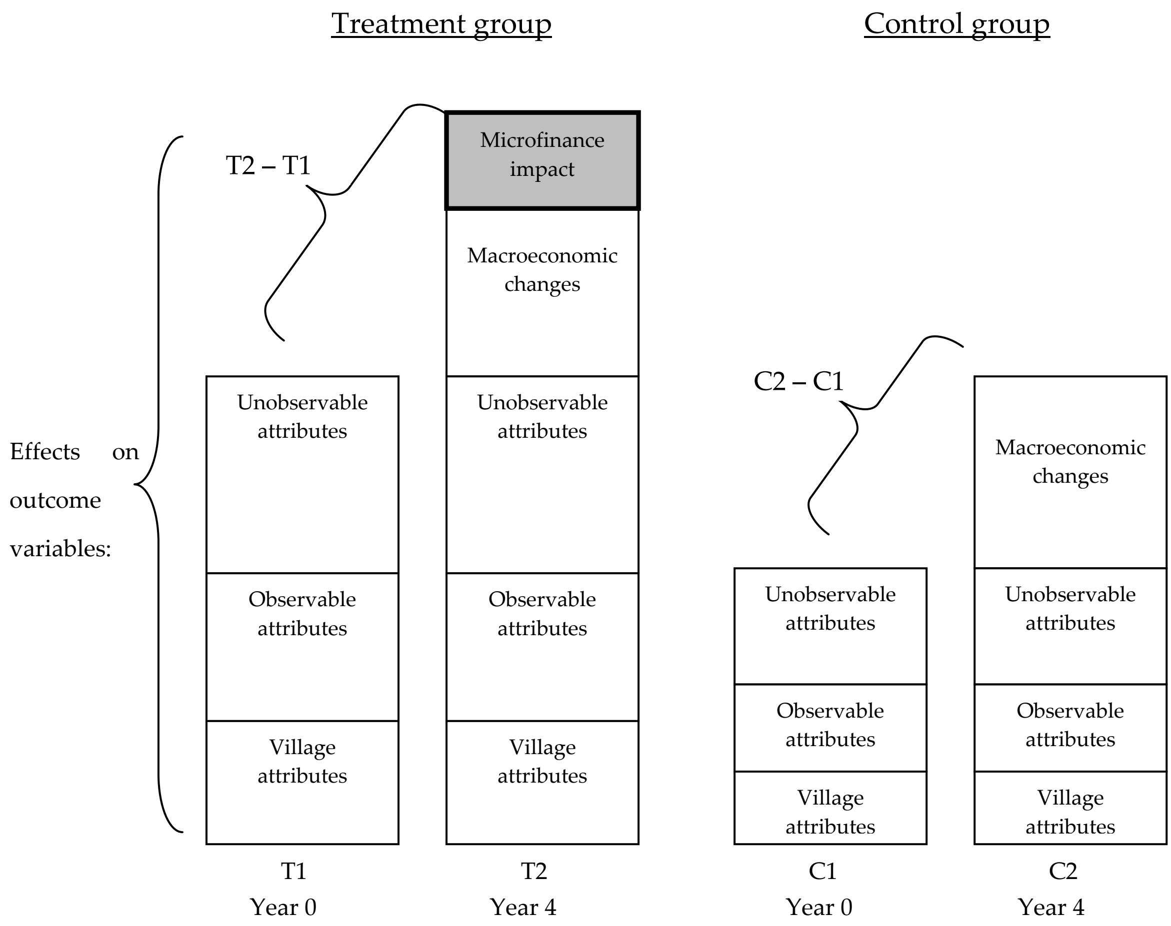
[20] Notations for equations (23), (24) and (25) follow Caliendo and Hujer (2005).

The effects of selection on unobservables is represented by the term π_{it} . For the DID method to be valid, the following assumption should hold: $\pi_{it} = \pi_{it'}$. In other words, the bias is assumed to remain constant over time and therefore can be controlled for using the DID approach (Heckman et al, 1998).

3.6.5.1. Differences-in-differences in microfinance

The DID approach can be explained with the help of an example from the area of microfinance. As noted in earlier chapters, isolating the effects of microfinance participation from all the changes or events that occur during participation is a challenge (Johnson and Rogaly, 1997; Armendáriz de Aghion and Morduch, 2005). Furthermore, microfinance impact evaluations are confounded by placement and selection bias. Assuming an adequate data set is available, DID can help to isolate microfinance treatment effects by eliminating certain observed and unobserved attributes. Armendáriz de Aghion and Morduch (2005) argue that there are village attributes, observable and unobservable attributes and macroeconomic changes. The authors further explain that village attributes refer to the particulars of where a person lives, e.g. access to markets which affects the likely returns to microfinance. For individuals, observable attributes can be age, education and experience, while unobservable attributes are for example entrepreneurial skills, organisational abilities, motivation, etc. all of which play a role when assessing the impact of microfinance. Hence, the aim is to isolate the microfinance impact controlling for all these attributes and measuring what is seen in the shaded box in Figure 4. However, things are more complicated because these attributes also influence the decision of microfinance participants to join the programme in the first place. Armendáriz de Aghion and Morduch (2005) suggest that there is a high correlation between entrepreneurial skills, age and microfinance participation. Also, if microfinance participants are wealthier than their non-participating peers before joining the programme, as suggested by studies conducted by Coleman (2006) and Alexander (2001), then they will have more potential for income growth. In addition, as the model introduced by Banerjee and Duflo (2010) shows, they will experience lower interest rates and get bigger loans.

Figure 4: Illustration of DID approach, microfinance context



Source: Adapted from Armendáriz de Aghion and Morduch (2005, p. 204).

In summary, it can be expected that participants differ from non-participants due to unobservable characteristics. These differences can lead to contrasting reactions in the event of macroeconomic changes and hence macroeconomic effects can have diverse impacts across both groups (Blundell and Costa Dias, 2002). Further to this, Blundell and Costa Dias (2002) provide a detailed discussion of the set up of π_{it} and break it down into individual-level effects that are either fixed or temporary and macroeconomic effects. The authors further claim that for a DID estimator to be unbiased, the decision to self-select into treatment needs to be independent from any temporary individual-level effects. It is further argued that any fixed individual-level and macroeconomic effects will eventually even out during the differencing procedure (Blundell and Costa Dias, 2002).

Returning to Figure 4, Armendáriz de Aghion and Morduch (2005) suggest comparing T1 with T2. By doing this, village attributes, observable and unobservable attributes that are assumed to be time-invariant are netted out, thus enabling the microfinance impact to be captured. However, the macroeconomic changes occurring between the years of observation and which are independent of the microfinance impact are also displayed but not yet controlled for. Thus, attributing the entire difference of T2 – T1 to the impact of microfinance would be misleading. This problem cannot be solved without the introduction of a control group (Armendáriz de Aghion and Morduch, 2005).

Hence, Figure 4 identifies a control group consisting of individuals that never had access to microfinance. This control group, however, is clearly not identical to the treatment group because of observable and unobservable differences. In a next step, Armendáriz de Aghion and Morduch (2005) suggest comparing T2 with C2 as this will address the biases arising from macroeconomic changes. This comparison appears to be adequate as these economic changes are felt in the same way by the control group as well as in the treatment group. To isolate the true impact of microfinance, however, the single difference of T2 – T1 will have to be compared to the difference of C2 – C1; this is the DID approach. This approach would work well in terms of accurately measuring the causal impact of microfinance if only the underlying assumptions would hold. As mentioned earlier, it is assumed that village attributes, observable and unobservable attributes in treatment and control group are time-invariant. As a result, their effects net out when analysing T2 – T1 and C2 – C1. This assumption, however, does not hold in practice since these attributes are bound to change over time and thus negatively effect the quality of the DID estimates (Armendáriz de Aghion and Morduch, 2005).

An innovative and well-known study that attempts to control for self-selection and non-random programme placement bias in the context of microfinance was conducted by Coleman (1999). The author deliberately applied DID in combination with a pipeline approach and used cross-section data from a quasi-experiment carried out in Northeastern Thailand in 1995 – 1996. He conducted a survey on 455 households and identified a treatment group consisting of participating and non-participating households in a programme village and a control group consisting of participating and non-participating households in a future programme village which had already been identified but which had not yet received any loans. The control group was surveyed one year before it received its first loan. In other words, the roll-out of the microfinance programme was delayed by one year for participants in the control group. This is a so-called pipeline comparison where the timing of programme roll-out is delayed (Khandker, Koolwal and Samad, 2010). Coleman (1999) uses a DID estimator and compares the difference of incomes between participants and non-participants in programme villages with the difference of incomes between participants and non-participants in the control villages. Coleman's (1999) study concludes that the microfinance programme in Northeastern Thailand has little impact. However, the author points out that his results should be read critically because Thailand is already fairly rich and less credit constrained compared to other developing nations. Overall, combining the DID approach with pipeline methods might be an appropriate strategy to account for selection on unobservables and can also be effective in the case of heterogeneous treatment effects (Khandker, Koolwal and Samad, 2010). Khandker, Koolwal and Samad (2010) further suggest that pipeline methods can also be combined with RDDs or even with PSM where PSM is applied before implementing the pipeline comparison (Galasso and Ravallion, 2004). However, this assumes that expansion areas are in all relevant aspects similar to already incorporated areas, which may not be the case.

Finally, another common problem in the application of DID estimators is the so-called Ashenfelter's dip (Blundell and Costa Dias, 2008). This phenomenon was first observed in labour market programmes where it was found that employment and earnings of potential participants decreased compared to individuals not intending to join shortly before the participation in an employment programme. It is argued that this could be due to the anticipation of participating (Ashenfelter, 1978). Caliendo and Hujer (2005) claim that

> "if the 'dip' is transitory and the dip is eventually restored even in the absence of participation in the programme, the bias will not average out" (p. 12).

Overall, the DID approach is immensely popular but not without flaws; e.g. section 3.6.5. argued that DID controls for selection bias since time-invariant selection effects are assumed (Caliendo and Hujer, 2005; Blundell and Costa Dias, 2008). This assumption, however, cannot always be observed (Khandker, Koolwal and Samad, 2010). Hence, better impact estimates may be obtained by either looking to alternatives or combining DID methods with other techniques. For example, Heckman, Ichimura and Todd (1997) and Khandker, Koolwal and Samad (2010) advocate combining DID with matching methods as this would account for selection on observables as well as on unobservables by comparing the outcomes of participants before and after an intervention with the before and after outcomes of matched non-participants (Caliendo, 2006; Caliendo and Hujer, 2005). This, however, assumes that the observable and unobservable attributes in treatment and control groups are indeed time-invariant, which is often not the case as shown earlier. I use and further discuss the combination of DID and PSM in chapter 5 when re-analysing the panel data set discussed by Khandker (2005), which builds upon the data collected by PnK.

3.7. Conclusion

This chapter has critically reviewed impact evaluation strategies. The literature provides a number of pointers as to which strategy to choose when there is no 'best practice' approach.

For example, Heckman and Vytlacil (2007a and 2007b) mention that randomised experiments are considered to be the 'gold standard' by evaluators. This, however, is not the case as discussed in section 3.5.3. and as pointed out by Deaton, 2009; Imbens, 2009; Scriven, 2008 and Pritchett, 2009. At the same time, observational studies are not without flaws either. The debate as to which evaluation strategy should be preferred is ongoing (see studies by Banerjee et al, 2009; Duflo and Kremer, 2005; Miguel and Kremer, 2004 supportive of experimental techniques and studies by Deaton, 2009; Imbens, 2009; Scriven, 2008 and Pritchett, 2009 for a critique). In essence, the choice of the evaluation strategy is guided by a set of assumptions as well as the background of the study design and data (Caliendo and Hujer, 2005).

Despite the well-known drawbacks of the various econometric evaluation methods - in particular with regard to their limited ability to controlling for selection on unobservables - many studies continue to claim that their impact estimates are robust and provide definite answers to the evaluation problem (PnK; Pitt, 1999). Moreover, many studies fail to acknowledge that the techniques they employ cannot compensate for the weaknesses of the underlying data (PnK; Pitt, 1999; Chemin, 2008). On the whole, researchers should critically scrutinise the impact estimates provided by quantitative impact evaluations. The characteristics of the data that have been used require further inspection as to their quality and availability of rich sets of variables. Also, certain data are simply not adequate for particular evaluation techniques, e.g. they lack variables yielding information on participation as well as outcomes and/or do not have adequate control groups (Smith and Todd, 2005; Heckman et al, 1998); this last point is further explored in chapter 4. These issues and others motivate the following empirical chapters which mainly draw on secondary data from India and Bangladesh which are used to replicate the findings presented by previous studies.

Researchers should be involved in the evaluation design and the data collection process as early as possible in order to be able to select an appropriate evaluation strategy and control for potential biases due to observable and unobservable characteristics as early as possible (Rosenbaum, 2002; Heckman, LaLonde and Smith, 1999; Caliendo and Hujer, 2005). However, in many cases researchers are not part of the initial stages of a study and will have to adjust their evaluation strategy ex-post according to the design of the programme, the selection processes thought to be at play and the quality and availability of data.

Ideally, an understanding of the contextual background of the programme under investigation should be gained (as discussed in sections 3.1. and 3.6.), i.e. through field visits and the collection of qualitative data which might help gathering information on the unobservables (Rosenbaum, 2002; Rosenbaum and Silber, 2001). This is of significance particularly for the area of microfinance where the success of a programme heavily depends on the quality and repayment behaviour of the groups that borrowers self-select into or are selected into by their peers or microfinance loan officers. This group formation process is driven by the unobservables such as access to social networks, entrepreneurial abilities and organisational skills (Armendáriz de Aghion and Morduch, 2005; Coleman, 1999). In addition, the workings of the microfinance institution in question and its lending strategy need to be understood as well as the country-specific context and the economic conditions at the time of the evaluation.

The next chapter presents the empirical findings from a microfinance programme in India with the objective to illustrate the theoretical and practical challenges of conducting rigorous impact evaluations and to further explore the role of selection on unobservables along with related topics such as social capital.

4. Smoke and mirrors: evidence of microfinance impact from an evaluation of SEWA Bank in India[21]

4.1. Introduction

The earlier chapters have outlined the challenges of conducting rigorous impact evaluations of microfinance interventions. We have seen that the ultimate goal of all impact evaluations is to assess how the lives of the poor would have turned out if a particular programme had not been introduced, which requires estimating the counterfactual - a process which commonly introduces selection bias (as shown in chapter 3). The latter chapter conceptualised the issue of selection bias and related the challenges of eliminating it to microfinance. Numerous econometric techniques are available to control for selection bias and obtain estimates based on counterfactuals but these techniques and their appropriate application are not free from controversy. Particular measurement challenges exist in the context of microfinance which need to be taken into account.

This chapter re-visits the evidence of the impact evaluation of SEWA Bank conducted by USAID in India in 1998 and 2000 to illustrate the broader challenges of measuring the impact of microfinance and to contribute to the methodological debate that was outlined in chapter 3. In particular, the challenges of controlling for selection bias and the role of the unobservables in this context are discussed in depth. Existing panel data are subjected to PSM and panel data techniques which purport to eliminate selection bias in impact evaluations[22].

---

[21] A revised and condensed version of this chapter has been published in the DEV Working Paper Series and is forthcoming in T. Nair (ed.) *Development Promise of Indian Microfinance* (provisional title), New Delhi: Routledge. I would like to thank Richard Palmer-Jones for his time, patience and invaluable feedback throughout the analysis and write-up of this chapter. I would also like to thank Arjan Verschoor and Ben D'Exelle for critical comments.

[22] This is not the first attempt to re-examine the SEWA Bank study. Augsburg (2006) appears to be the first to re-investigate SEWA Bank's cross-section results. However, her study focuses on examining merely three household-level income-related outcome variables. She applies PSM DID. Her results could not be fully replicated due to differences in the data re-construction; I contacted the author for her STATA do-files in order to comprehend her data analysis but she could not make them available. Her matching estimates were on average lower than the ones presented here in this study. I will not further refer to her study since this paper casts its net wider and re-analyses all of USAID's outcome variables on the household, enterprise and individual level.

The matching results are subjected to sensitivity analysis to assess their robustness, this is rather novel. Sensitivity analysis of PSM was examined extensively by Rosenbaum (2002) and taken further by Ichino, Mealli and Nannicini (2006), and others, as discussed later in this chapter. The few studies that have applied PSM to microfinance (e.g. Abou-Ali et al, 2009; Arun, Imai and Sinha, 2006; Chemin, 2008; Deininger and Liu, 2009; Imai, Arun and Annim, 2010; Imai and Azam, 2010; Setboonsarng and Parpiev, 2008; and Takahashi, Higashikata and Tsukada, 2010) have not given sensitivity analysis much attention.

The analysis of the survey data is supported by direct observation; taken together the sensitivity analysis and the direct observation suggest that selection processes driven by unobservables are highly likely to have influenced who became a participant in microfinance (and progression from saver to borrower), and, as is well known, unobservables cannot be controlled by econometric techniques. In addition, the role of social capital (as briefly referred to in chapter 2) in those selection processes is further investigated and evidence is presented suggesting that informal networks drive self-selection into microfinance. Further doubt is thrown on the impact claims by the sampling strategy of the control group for the original USAID study, which is not sufficiently described in the literature. Also, the panel design is problematic because it does not have a 'true' baseline which would allow a before and after comparison, with the control and treatments groups shown to be equivalent before joining SEWA Bank, since the treatment groups had already joined SEWA Bank well before the baseline period.

SEWA Bank members start as savers and the majority never progress to borrowing. The literature on the impact of savings on the well-being of the poor is scarce (see studies by Aportela, 1999; Ashraf, Karlan and Yin, 2006; Berg, 2010; and Devaney, 2006) since few MFIs offer savings products only. SEWA Bank is one of those MFIs that focuses on a savings approach and having more savers than borrowers (Chen and Snodgrass, 2001); thus a further selection process segregates SEWA Bank members into borrowers and savers, in which unobservables may also well play a role. Thus this chapter also contributes to the literature by conducting various sub-group comparisons between borrowers, savers and controls in order to potentially shed some light on the impact of savings versus credit. Finally, I draw conclusions as to what these findings imply for the robustness of the original impact estimates provided by USAID. I find that while the results presented by USAID cannot be contradicted, there are doubts about the quality of the PSM impact estimates, and sensitivity analysis indicates how much selection on unobservables there would have to be for the measured impact to become insignificant. The USAID control group sampling procedure and the panel design leave questions of comparability with the treatment group unanswered. As a result, it is also not possible to reject the hypothesis that unobservable differences in characteristics account for some at least and possibly most of the observed impact. I draw conclusions as to appropriate impact evaluation procedures, and the need for qualification to even apparently robust estimates of impact.

This chapter begins by linking the selection bias problem with social capital in the specific context of microfinance; this is followed by a brief description of SEWA Bank, its products, services and workings, drawing on my qualitative fieldwork during compilation of a repeat cross-section data set. Next, the research design and data as well as the sampling procedures are described together with the estimation strategies; the analyses of USAID are replicated and PSM is applied to their data as well as the new cross-section data and sensitivity analysis are conducted. Finally, the results are discussed in the light of my own and other qualitative research on MFIs, before drawing conclusions.

4.2. Social capital and microfinance

Social capital has received a lot of attention in the microfinance literature, in particular when group lending schemes which were discussed in chapter 2 are employed (Ito, 2003). As mentioned in chapter 2, Karlan (2007), Besley and Coate (1995) and Wydick (2001) argue that the presence of social capital in group lending schemes play an important role in improving repayment behaviour and reducing the number of defaults. Furthermore, Ito (2003) argues that

> "many of the successful microfinance schemes are seen to rely upon social capital in the form of small groups within which information sharing takes place" (p. 324).

This statement supports the widely held view that group lending schemes resolve all information asymmetries and hence overcome imperfections in credit markets (Banerjee, Besley and Guinnane, 1994).

The academic literature distinguishes between three main concepts of social capital. Putnam (1993) provides the narrowest definition of social capital and describes it as horizontal associations that link a variety of actors (Ito, 2003). The definitions provided by Olson (1982) and North (1990) are broader and include the entire social and political spectrum that influence a society as well as its institutional framework such as government, laws and regulations. Coleman's (1988) definition lies somewhere in between. He acknowledges the existence of horizontal as well as vertical associations between actors, where vertical associations refer to hierarchical and unequal relationships between those actors (World Bank, 1998). Thus, the concept of social capital appears to be rather fuzzy as implied by Harriss and de Renzio (1997) and Molyneux (2002) and there is no consensus on which one of these three definitions should be preferred. The lack of a clear definition of social capital also has implications for identifying its role in enhancing the effectiveness of development interventions and for accurately measuring it in the sense that the choice of empirical proxy variables representing social capital is rather arbitrary.

Social capital first gained prominence in development circles through the social capital initiative started by the World Bank in 1996. The World Bank argues that social capital influences the outcomes of development programmes. Hence, the initiative was launched with the objective to provide a better understanding of the role of social capital on the effectiveness and sustainability of development interventions (World Bank, 1998).

The World Bank describes social capital as follows:

> "The social capital of a society includes the institutions, the relationships, the attitudes and values that govern interactions among people and contribute to economic and social development. Social capital, however, is not simply the sum of the institutions which underpin society, it is also the glue that holds them together. It includes the shared values and rules for social conduct expressed in personal relationships, trust, and a common sense of 'civic'

> responsibility, that makes society more than a collection of individuals. Without a degree of common identification with forms of governance, cultural norms, and social rules, it is difficult to imagine a functioning society" (World Bank, 1998, p. 1).

Social capital also plays a role in individual lending schemes as this chapter demonstrates. The discussion in chapters 2 and 3 argued that microfinance participants commonly self-select into a programme or are selected by their peers which often monitor their performance. This process is mainly driven by unobservable characteristics of borrowers such as access to social networks, entrepreneurial drive or business skills (Coleman, 1999). As a result, microfinance participants significantly differ from non-participants who in principle also have access to microfinance but whose (difficult to observe) lack of social capital means they are not selected into microfinance. Relationships between borrowers who have formed groups of their own accord are defined as horizontal social capital (Ito, 2003) or horizontal associations in Putnam's (1993) terminology. Furthermore, potential microfinance borrowers not only self-select into microfinance or are selected by their peers but are also encouraged to participate in microfinance by the loan officers they interact with. Those loan officers form special relationships with certain individuals from the communities they frequently visit (Ito, 2003), something observed during my own fieldwork in Ahmedabad in autumn 2008. The closeness of this relationship can lead to preferential treatment with regard to identifying borrowers and approving loan applications. Ito (2003) refers to this as vertical social capital, which is characterised by hierarchical relationships. She also provides evidence from GB in Bangladesh showing that those close relationships between borrowers and loan officers indeed exist (see also Fernando, 1997). Moreover, Ito (2003) argues that some of GB's loan officers

> "exercised personal discretion to get around Grameen's rules, often to the benefit of the borrowers" (p. 327).

Ito (ibid) further argues that the power of GB's loan officers influences the loan approval process. Irrespective of the loan approval procedures that exist within an MFI, loan officers use their information advantage about particular loan applicants and their power to either grant or reject loan proposals, thus deciding over microfinance participation. Ito (ibid) and Holmes, Isham and Wasilewski (2005) reiterate this point and argue that the stronger the relationship (measured in terms of length of participation in microfinance and the client's credit history) between the MFI as represented by the loan officer and an individual the more likely is the individual to obtain a loan. Direct observations made during my research with SEWA Bank support the view that vertical social capital of this type is indeed present, and increases the likelihood of individuals to be selected into microfinance (see Box 1 for evidence).

Box 1: Vertical social capital at SEWA Bank

The aim of the qualitative part of this study was to understand the selection processes from an ethnographic point of view. I wanted to understand how potential microfinance clients are recruited into SEWA Bank. To do this, I interviewed SEWA Bank staff and shadowed them in the field for several days in addition to interaction with them over design and fielding of a survey questionnaire.

The SEWA Bank staff I talked to explained that the Bank's 'recruitment' process works as follows: potential clients are generally recommended by people they know, i.e. family, friends or neighbours and are then referred to so-called Bank Saathi. The Bank Saathis are SEWA Bank's voice in the field; they are the first point of contact for the clients and function as advisors and mediators. In other words, they are the link between SEWA Bank and the clients. Saathi literally means companion. Bank Saathis are clients themselves and live in the same neighbourhoods as ordinary SEWA Bank clients. According to information from SEWA Bank, an individual can become a Bank Saathi when she has been saving and borrowing with SEWA Bank for several years, has displayed impeccable financial behaviour, is honest, trustworthy and good at managing relationships as well as has a certain social standing in the community. Bank Saathis are responsible for savings and loan collections but also for recommending and identifying future clients. Bank Saathis are paid on a commission basis, i.e. the more clients they 'recruit' the more they earn. Unfortunately, I could not obtain any further information on the nature of these payments and their size in proportion to the Bank Saathis' income. Once a Bank Saathi has identified a potential client, a member of SEWA Bank staff visits the potential client to initiate the loan approval procedure. I shadowed SEWA Bank staff on some of those visits and was surprised by the informality of the loan approval process. It was more like an informal conversation with the potential client and was completed within ten minutes. These visits were often not documented and the loan was usually granted after such a visit – i.e. on the basis of information produced in this process and trust in the network that led to it and without further investigations.

Cont.

I formed the impression that SEWA Bank staff were usually inclined to follow the recommendations of the Bank Saathi, mainly because they (the Bank Saathi) are presumed to know the potential clients, their family, friends and neighbours as they are living in the same communities (and can thus observe those variables that are unobservable to formal data production techniques). As a result, there is a lot of room for the Bank Saathis to abuse their information advantage and power as suggested by Ito (2003). My enumerators saw evidence of such abuse and observed that some Bank Saathi demanded 10% to 15% of the loan amount granted as an additional commission from the client. It appears that this informal screening or selection process does indeed play a role in explaining microfinance participation. However, the econometric tools commonly applied in the context of impact evaluations do not seem to be able to control for those unobservables that seem to be driving the screening or selection process. Based on my observations I conclude that an ethnographic approach could possibly be more appropriate for providing further insights and this would be a recommendation for future research in this area.

The World Bank (1998) argues that understanding social capital is important in explaining the effectiveness and sustainability of development interventions; what is meant by this is difficult to pinpoint (as noted in my discussion above of the three definitions of social capital). Nevertheless, in the particular case of microfinance, indicators of social capital may provide insights into unobservables and explain why some, among otherwise indistinguishable individuals, are selected into microfinance by themselves, their peers or by loan officers.

The analysis in this chapter seeks to understand whether analysis of the original USAID observational data, supplemented by my own re-survey can throw light in particular on the existence and effects of selection on unobservables. Before outlining the specifics of the research design, data, sampling procedure and estimation strategy, the next section introduces SEWA Bank and its microfinance programme.

4.3. The SEWA Bank context

SEWA Bank, a cooperative bank headquartered in the Indian city of Ahmedabad, is a sister organisation of the Self Employed Women's Association (SEWA). SEWA was established in 1972 as a trade union with the objective to organise self-employed women working in the informal sector. SEWA is not a mere trade union but a women's movement with its origins in Ghandian philosophy based mainly on principles of truth, non-violence and self-sufficiency (http://www.sewabank.com/aboutus-origin.htm). Various sister organisations grew out of the SEWA movement such as SEWA Bank which provides microfinance products, Vimo SEWA which provides insurance services, SEWA Academy, which is responsible for training and research, and a number of other organisations which offer a range of services to its female members (http://www.sewabank.com/aboutus-origin.htm).

Based on an initiative by Ela Bhatt, SEWA Bank was established in 1974 with the help of 4,000 SEWA members. The aim of SEWA Bank is to provide financial services such as savings and loan products to self-employed women. The bank has its base in urban Ahmedabad where it mainly operates individual savings and lending programmes but in the early 1990s it has also expanded into Gujarat's rural areas where it provides its services through self-help groups (SHGs) (http://www.sewabank.com/rural-activities.htm).

SEWA Bank emphasizes the provision of savings over credit (Chen and Snodgrass, 2001). That is illustrated by the following figures: as of fiscal year (FY) 2007, SEWA Bank had 163,187 clients out of which 143,806 were savers and the remainder of 20,011 were borrowers. SEWA Bank only targets women, not all of them are micro-entrepreneurs, i.e. women that sell goods and services on their own account, but many work as casual labourers or sub-contractors. Furthermore, SEWA Bank targets minorities; e.g. roughly 25% of its clients are Muslims and the remainder are from scheduled castes and tribes and other backward castes[23]. To relate those figures to the overall population of Ahmedabad: according to the Census of India (2001), Ahmedabad has an overall population of close to 4.7 million out of which 82.1% are

[23] This information is taken from SEWA Bank's internal management information system (MIS) which I had access to.

Hindus (out of which 13% are scheduled castes and 1% are scheduled tribes), 13% are Muslims and the remainder are Christians, Buddhists, Sikhs and Jains.

SEWA Bank offers a range of savings products such as current deposit, fixed term deposit and ordinary savings accounts as well as loan products. Loans can be secured, which requires physical collateral such as jewellery or a savings account, or unsecured, which requires a guarantor as 'social' collateral. In addition, housing loans are offered as well as emergency loans which many clients use to pay for weddings, funerals or other consumption-related expenditure. As of FY 2007, approximately 49% of the loans disbursed were secured, 24% were unsecured and 27% were given as housing loans[24]. Loan sizes vary from 5,000 Rupees to 50,000 Rupees. All loans are provided under an individual lending scheme, SEWA Bank does not operate any group lending schemes in its urban operations.

The loan application process works as follows: every potential SEWA Bank borrower is required to first open a savings account and SEWA Bank staff then monitors the savings behaviour of those potential borrowers, i.e. the size and regularity of their savings. Potential borrowers qualify for loans once they have regularly deposited money in the savings account for at least 6 months (Chen and Snodgrass, 2001). However, those rules are often relaxed. SEWA Bank loan officers, so-called Bank Saathi (as explained in Box 1), are living in the communities they service and are responsible for recommending future clients whose creditworthiness they then assess as well. If a Bank Saathi feels that a future client is bankable, then SEWA Bank commonly does not reject the loan application; this suggests that informal networks around these Bank Saathi drive the selection and loan approval processes as described in Box 1. In addition, if a future client can provide collateral, either physical or 'social', i.e. in the form of a guarantor, a loan is usually granted without the need of having to open a savings account.

SEWA Bank is unique in many ways. It is one of India's oldest and most established microfinance providers with a strong ideological base rooted in Ghandian traditions using struggle and development as a strategy to strengthen their member's position in society. In addition, SEWA Bank also prefers to work in a cooperative structure and

---

[24] This information is taken from SEWA Bank's internal MIS.

extend its financial services to individuals without the need for group formation. This cooperative structure allows SEWA Bank to focus on a savings approach. Typically, microfinance in India is offered by microfinance-NGOs (MF-NGOs) which are registered as non-profit organisations. The registration as a non-profit organisation limits their scope for providing financial services. The RBI, for example, prohibits all non-profit organisations from taking savings (Fisher and Sriram, 2002). Strictly speaking the MF-NGOs that do take savings are operating illegally. Thus, many organisations do refrain from mobilising savings because their organisational set-up simply does not allow it. Recently for-profit MFIs have emerged, so-called Non-Banking Financial Companies (NBFCs); examples include organisations like BASIX and SHARE. However, NBFCs, although regulated by the RBI, are also not allowed to take savings (Fisher and Sriram, 2002). Ghate (2007) as well as many practitioners (Karlan and Morduch, 2009) argue that this is a major drawback because the poor need savings more than credit as the next section will elaborate in more detail.

4.4. Impact of savings

As mentioned in section 4.3., SEWA Bank emphasizes savings over credit and had on average seven times more savers than borrowers as of FY 2007. SEWA Bank views credit merely as a complementary tool to savings, and hence I assess the impact of credit as well as savings to account for SEWA Bank's distinctive approach. The objective of this section is to briefly introduce the savings literature and to review some of the key studies that assessed the impact of savings.

Policy makers assumed for a long time that the poor are too poor to save and hence savings mobilisation was low on the agenda of many governments. This assumption has been questioned by Adams (1978) and von Pischke (1983) and further by Rutherford (2001) and Collins et al (2009). Rutherford (2001) claims that the poor have the capacity to save and traditionally used rotating savings and credit associations (ROSCAs) or other informal mechanisms to satisfy their savings needs. Indeed, savings are crucial for accumulating assets which in turn are used to finance future investments and consumption (von Pischke, 1983). Following Keynes (1936) and Browning and Lusardi (1996), Karlan and Morduch (2009) explain that individuals have various motives that encourage them to save such as

> "precautionary, life-cycle (to provide for anticipated needs), intertemporal substitution (to enjoy interest), improvement (to enjoy increasing expenditure), independence, enterprise, bequest, avarice, and downpayment" (p. 39).

It is beyond the scope of this thesis to discuss the motivations of individuals to save in detail and hence I will refer the interested reader to a comprehensive review of the savings behaviour of individuals in developing countries which is provided by Rosenzweig (2001).

Studies evaluating the impact of savings are scarce; some notable exceptions include the studies by Aportela (1999), Ashraf, Karlan and Yin (2006), Berg (2010) and Devaney (2006) for a review. Dupas and Robinson (2009) conducted the first and so far only RCT assessing the impact of savings products. Most MFIs focus on providing credit as well as savings and a range of other services which makes it rather challenging to disentangle the impact of savings from all the other products and services that clients use at the same time; e.g. Burgess and Pande (2005) showed that financial access can reduce poverty but they could not separate the impact of savings from the impact of credit.

Devaney (2006) reviewed eight impact studies - including the SEWA Bank study discussed in this chapter - that focused on the impact of savings on the poor. The aim of most of these studies was to provide evidence that the poor have the capacity to save in the first place and to justify the need for savings products in addition to loan products. Moreover, the majority of these studies reviewed by Devaney (2006) investigated the impact of a particular savings product on the savings rate of the poor and found that access to a savings product had indeed a positive impact on the households' savings rate. An exception is the study by Berg (2010) which uses the SEWA Bank data to assess the impact of savings on household income and consumption. When controlling for selection bias, she finds no significant impacts of participation in SEWA Bank's savings programme and concludes that naïve impact estimates that neglect selection bias simply overstate programme effects. Another recent study on savings impact by Dupas and Robinson (2009) is based on a field experiment in Kenya, testing for the existence of savings constraints, and concluded that access to savings has positive impacts on income and productive investments.

Furthermore, Devaney (2006) claims that borrowers are more likely to save than non-borrowers. However, most of these studies mentioned here did not compare the impact of saving versus the impact of borrowing versus not saving or borrowing at all. Only the study by Rogg (2000) and the SEWA Bank study under discussion in this chapter are exceptions in this regard. It can be concluded that the literature in particular on the impact of saving versus borrowing is still rather underdeveloped, but is mainly positive.

In the Indian context, the RBI frequently turns a blind eye to the MFIs illegally mobilising savings because it recognised the importance of microfinance and savings in particular (Basu, 2006). A solution to the savings dilemma, i.e. the fact that MFIs are not officially allowed to mobilise savings but unofficially do so at times, is to register a MFI as a mutual benefit organisation, which allows it to be classified as a cooperative. SEWA Bank is one of the few microfinance providers that is registered as an urban cooperative bank which means that savings can legally be mobilised (Fisher and Sriram, 2002). This organisational set-up is suitable for SEWA Bank's activities since it allows its clients to save rather than only access credit. This chapter returns to the issue of savings in section 4.8.6. when comparing the impact estimates of the various sub-groups, i.e. borrowers, savers and controls. The question is whether the data support the view that a savings approach - as advocated by Ghate (2007) and as implemented by SEWA Bank - is justified and desirable. After this brief introduction to SEWA Bank, the next section presents the research design and describes the data.

4.5. Research design and description of data

The study discussed here is one of three longitudinal USAID microfinance impact evaluations that were carried out between 1997 and 2000 on Mibanco in Peru, Zambuko Trust in Zimbabwe and SEWA Bank in India. All three studies share a similar research design and aim to examine the socio-economic impact of microfinance participation (Snodgrass and Sebstad, 2002). For the purpose of this study, the impact evaluation conducted on SEWA Bank is discussed in more detail in this section.

The original SEWA Bank study (henceforth USAID) assesses the impact of SEWA Bank's microfinance services on urban client households (Chen and Snodgrass, 2001). It examines hypotheses at the household, enterprise and individual level. The study hypothesized that microfinance participation at the household level leads to an increase in household income, more diversified income sources, housing improvements, an increase in household assets, better education of the household's children, an increase in food expenditure and improved mechanisms for coping with shocks. At the enterprise level, microfinance participation leads to an increase in informal sector income, an increase in revenues and fixed assets, employment generation as well as better transactional relationships. At the individual level, microfinance clients might gain more control over the household's resources and incomes, increase their self-esteem and self-confidence, and increase personal savings and improve their ability to deal with the future (Chen and Snodgrass, 2001). Table 5 outlines the details of the hypotheses tested and the corresponding impact variables.

Table 5: Hypotheses and impact variables of USAID study

Household level	
Hypotheses	**Impact variable**
H1: increase of household income	• Total annual household income • Household income per capita
H2: more diversified income sources	• Inverse Simpson's index
H3a: housing improvements	• Expenditure on housing improvements and repairs in terms of material and labour
H3b: increase of household assets	• Expenditure on household assets, e.g. appliances, vehicles, jewellery
H4: better education of the household's children	• Net enrolment ratios
H5: increased food expenditure	• Per capita expenditure per day for food and beverages
H6: better coping with shocks	• Mechanisms used for dealing with shocks
Enterprise level	
E1: increase of informal sector income	• Microenterprise income of previous month from household head and respondent
E2: increase of revenues	• Gross sales revenue of previous month
E3: increase of fixed assets	• Value of all fixed assets used in microenterprise
E4: more employment generation	• Hours worked in previous week • Days worked in previous month
E5: better transactional relationships	• Types of suppliers and customers
Individual level	
I1: client gains more control of the household's resources and income	• Who took decision to take last loan? • Who took decision how to spend loan amount? • Who took decision how to spend income?
I2: increase in self-esteem and self-confidence	• Respondent's feelings with regard to her contribution to household • Is this contribution respected by other household members?
I3: increase in personal savings	Existence of personal savings
I4: better ability to deal with the future	• Respondent's feelings with regard to preparedness to deal with future • How does respondent prepare herself to deal with future?

Source: Chen and Snodgrass (2001, p. 58).

In order to test the hypotheses outlined in Table 5, researchers collected baseline data on 900 women from low-income households across ten wards in Ahmedabad. The sampling criterion required the selection of women who were above 18 years and economically active. An economically active person is defined as somebody who engages in informal economic activities in the home, on the street or on business premises and who is either self-employed, a dependent producer or a wage worker (on an irregular basis without written contracts and/or fixed wages) (Chen and Snodgrass, 2001). Out of the 900 sample women, 600 were SEWA Bank clients – consisting of borrowers and savers - and 300 non-clients. The sampling procedure was based on a three-step process (Chen and Snodgrass, 2001). First, a geographical area was selected. Ahmedabad is split into 43 wards and USAID limited its survey to 10 of those 43 wards due to budget constraints. The sample was drawn from the following ten wards in Ahmedabad: Behrampura, Jamalpur, Bapunagar, Rakhial, Asarwa, Khadia, Amraiwadi, Saraspur, Raikhad and Dudheshwar. The wards were selected based on the number of SEWA Bank clients residing in them. Almost half of all current clients live in those 10 wards (Chen and Snodgrass, 2001). Next, a random sample of borrowers and of savers was selected from a list provided by SEWA Bank which contained all its current borrowers and savers as of FY 1997, listed by ward. Savers should have made at least one deposit in a SEWA Bank savings account during FY 1997. Moreover, savers should not have taken out any loans in FY 1997. Replacements were made when the respondent could not be located, was not economically active, e.g. not self-employed anymore, or did not want to participate. Also, replacements were needed when respondents from the sample of current savers were not actively saving anymore or had taken out loans in FY 1998.

The rationale for sampling borrowers as well as savers is explained by SEWA Bank's emphasis on savings over credit - at the time of the USAID study there were ten savers for every borrower - hence USAID decided to gather a separate sample of savers. Chen and Snodgrass (1999) explain

> "that those clients who are savers only will benefit from having a secure place to deposit their savings. Since all borrowers have to save, it is hypothesized that there will be greater impact on those who borrow as well as save" (p. 16).

Finally a non-client sample was chosen. USAID carried out a pre-survey

> "in the neighbourhood [*Author's note: it is not clear whether neighbourhood and ward are used interchangeably or whether neighbourhood refers to something else*] of each of the 300 sample borrowers to identify 50 households in which there were economically active women over age 18 who were not SEWA members" (Chen and Snodgrass, 2001, p. 53).

From those 15,000 households a random sample of 300 non-clients was drawn.

Rosenbaum (2002) argues that the sampling of an appropriate control group is crucial in observational studies and in view of this the robustness of the USAID sampling procedure of the control group is explored. Chen and Snodgrass (2001) argue that the neighbourhoods where most of SEWA Bank's clients reside are reasonably homogeneous in terms of caste, occupation and class (p. 53) and hence the control group is relatively similar to the treatment group. However, if the households in the control group are so similar, then why are they not clients of SEWA Bank? This points towards a selection process that is driven by unobservable characteristics which account for why otherwise apparently eligible households did not belong to SEWA Bank. As a consequence, the control group sampling of USAID does not convince. Chen and Snodgrass (2001) admit that SEWA Bank members

> "are not chosen at random but are in fact purposefully selected from a larger population, both by themselves and by SEWA Bank. A woman must first self-select by deciding to open a savings account and later to apply for a loan. Once she does so, SEWA Bank decides whether to provide her with the financial service in question" (p. 60).

The follow-up data I collected copies the USAID sampling procedure as much as possible and this is elaborated in section 4.6. The first round (henceforth Round 1) of the USAID survey was conducted in January 1998 and a follow-up round (henceforth Round 2) was then collected in January 2000. Between survey rounds, a rate of attrition of approximately 11 percent was observed, resulting in a final sample of 798 respondents (Chen and Snodgrass, 2001, p. 56). In addition to the two surveys, twelve case studies of SEWA Bank borrowers were conducted with the objective to provide a better understanding of the issues that SEWA Bank borrowers commonly have to deal with on a daily basis and how microfinance has helped them in the process (Chen and Snodgrass, 2001).

The data collected from both survey rounds was then subjected to Analysis of Variance (ANOVA) in order to examine cross-section differences and Analysis of Covariance (ANCOVA) to evaluate whether any personal characteristics possibly influenced any impact variables. Chen and Snodgrass (2001) argue that ANCOVA would reduce selection bias to a certain degree. In addition, gain score analysis was employed to estimate the degree of change over time between treatment and control groups and to assess whether such changes were significant. The headline findings of the USAID study provide evidence that microfinance leads to changes at the household level, i.e. higher household income in terms of total income and per capita income was observed. In addition, minor positive impacts could be observed on income diversification, food expenditure and the ability to cope with shocks. However, the evidence was rather mixed. Moreover, impact at the enterprise and the individual levels were negligible (Appendix 2 presents the detailed USAID results for the household, enterprise and individual level). Chen and Snodgrass (2001) admit that measuring impacts at the enterprise and individual level were rather challenging due to the fact that SEWA Bank clients are not classical micro-entrepreneurs per se. Most clients do not have micro-enterprises but are dependent sub-contractors or labourers, thus do not require micro-enterprise capital. SEWA Bank provides loans for a range of purpose, e.g. business, housing improvements/repairs, repayment of other debts and consumption but without a particular focus on micro-enterprise development.

As discussed in chapter 2, fungibility of money is a central problem in the context of microfinance impact evaluations and notoriously difficult to control for. To recap, money is considered to be fungible within the household, i.e. once a loan has been taken out by the borrower, it is difficult to track in which way the loan has actually been used (Ledgerwood, 1999). Based on the findings of the USAID study, it appears that measuring impact separately at the enterprise and individual level does not lead to particularly satisfactory results. This is likely to be due, at least to some extent, to fungibility.

As mentioned earlier, the aim of this study is to re-visit the evidence of microfinance impact evaluations. Therefore, the USAID panel data set has been subjected to more advanced econometric techniques, i.e. PSM to account for selection bias. In addition, the new cross-section data set (henceforth Round 3) was produced, with the aim of exploring the potential of social capital indicators to illuminate the role of the unobservables and to compare the USAID panel with Round 3 to get a clearer picture on short-term versus long-term impacts of microfinance. Hence, a social capital section was added to the original USAID questionnaire.

Efforts were made to track the original respondents of the USAID study in order to collect a proper panel round to the existing panel sample but this was not possible due to the unavailability of detailed information of the original households which could not be tracked even with the cooperation of SEWA Bank. As a result, new research respondents were sampled in a comparable way to the USAID study in terms of geographical area and socio-economic characteristics. Table 6 below provides descriptive statistics as evidence of the comparability of the samples across all three data collection rounds. Further descriptive statistics on all outcome variables across rounds 1 - 3 can be found in Appendix 1.

Table 6: Descriptive statistics of female research respondents

		Borrowers			Savers			Controls		
Data round		R1	R2	R3	R1	R2	R3	R1	R2	R3
Sample size		264	264	70	260	260	70	262	262	70
Mean age (years)		37.81	40.28	39	34.55	36.88	40.44	35.36	37.51	39.98
Marital status in %	Married	89.77	88.64	92.86	87.31	85	92.86	80.92	79.39	97.14
	Never married	1.89	1.89	0	5.77	4.62	1.43	5.73	5.73	0
	Divorced	0.76	0.76	1.43	1.15	0.38	0	1.15	0.38	0
	Deserted	0.38	0	1.43	1.54	1.15	1.43	1.15	1.53	0
	Widowed	7.2	8.71	4.29	4.23	8.85	4.29	11.07	12.98	2.86
Religion in %	Hindu	72.35	72.73	68.57	76.54	76.92	68.57	77.1	77.48	75.71
	Muslim	27.27	26.52	28.57	23.46	23.08	28.57	22.52	22.14	24.29
	Other	0.38	0.76	2.86	0	0	2.86	0.38	0.38	0
Caste in %	Upper caste	15.15	14.39	7.14	16.15	15.77	12.86	22.9	23.66	10
	Backward caste	45.45	46.97	10	40.77	43.85	11.43	39.31	40.46	11.43
	Scheduled caste	29.92	31.82	24.29	35.38	35.77	17.14	29.77	32.06	14.29
	Scheduled tribe	9.09	6.44	22.86	7.31	4.62	30	8.02	3.82	32.86
	No response	0	0	35.71	0	0	28.57	0	0	31.43
Education	Never attended school in %	39.77	40.15	54.29	40	41.92	61.43	40.84	44.66	61.43
	Mean highest grade completed	3.9	3.9	3.33	4.3	4.24	3.14	4.2	4.01	2.67

Source: Author's calculations.

As in the case of the USAID study, the majority of impact studies examine the impact of microfinance at multiple levels, i.e. at the household, enterprise and individual level; see Hulme and Mosley (1996), Sebstad et al (1995) and Gaile and Foster (1996) for a comprehensive overview of studies up to the mid 1990s. However, examining the impact at multiple levels requires sufficient funds and time. Moreover, solely looking at the individual, enterprise or community level has a number of disadvantages (see Table 7 for details). In particular the issue of fungibility (as mentioned earlier) has to be considered when assessing impact at the enterprise level. Hulme (2000) argues that "…for all studies except those that focus exclusively on 'the enterprise,' [then] a concern about fungibility may be irrelevant" (p. 85). He further argues that the most promising way to measure impact of microfinance appears to be at the household and institutional level. However, institutional level data is not available in this case. Hence, based on Hulme (2000) and after carefully examining Table 7, I conclude that re-examining the household level hypotheses of the USAID study appears to be the way forward due to issues of fungibility and difficulties of breaking down household level impacts onto the individual level. Therefore, Round 3 focuses on collecting household level data only.

Table 7: Units or levels of evaluation and their advantages and disadvantages

Unit	Advantages	Disadvantages
Individual	• Easily defined and identified	• Most interventions have impacts beyond the individual • Difficulties of disaggregating group impacts on "relations"
Enterprise	• Availability of analytical tools (profitability, return on investments, etc.)	• Definition and identification is difficult in microenterprises • Much microfinance is used for other enterprises and/or consumption • Links between enterprise performance and livelihoods need careful validation
Household	• Relatively easy defined and identified • Permits an appreciation of livelihood impacts • Permits an appreciation of interlinkages of different enterprises and consumption	• Sometimes exact membership difficult to gauge • The assumption that what is good for a household in aggregate is good for all of its members individually is often invalid
Community	• Permits major externalities of interventions to be captured	• Quantitative data is difficult to gather • Definition of its boundary is arbitrary
Institutional Impacts	• Availability of data • Availability of analytical tools (profitability, Subsidy Dependency Indices (SDIs), transaction costs)	• How valid are inferences about the outcomes produced by institutional activity?
Household Economic Portfolio (i.e. household, enterprise, individual and community)	• Comprehensive coverage of impacts • Appreciation of linkages between different units	• Complexity • High Costs • Demands sophisticated analytical skills • Time consuming

Source: Hulme (2000, p. 83).

In the fieldwork for this study the original household questionnaire of the USAID study was administered but with changes, e.g. the food expenditure section was dropped and a social capital section was included with the aim to shed some light onto the role of social and information networks in the case of microfinance. As discussed earlier, horizontal and vertical social capital plays a role in the selection or screening process that determines microfinance participation, in particular vertical social capital plays a role in the SEWA Bank context (as described in Box 1 earlier in this chapter). Hence, social capital type questions which are based on the Social Capital Assessment Tool (SOCAT) developed by the World Bank (Grootaert and Bastelaer, 2002) were added with the objective to understand the role of the unobservables in this context. The questionnaire was pre-tested with a few non-sample client households and then administered to 220 households across the selected 10 wards in Ahmedabad.

Further to adding a social capital section to the questionnaire, eight case study interviews with randomly sampled clients and non-clients chosen from the survey households were conducted with the objective to gain a better understanding of the role of entrepreneurial drive and business skills in microfinance participation. Moreover, drop-outs in addition to the client and non-client groups were sampled in order to understand why households decide to leave microfinance – it is rather common in microfinance that clients exit programmes, once they have exhausted the utility of the products and services available. Armendáriz de Aghion and Morduch (2005) mention drop-out rates between 3.5% and 60% in a wide range of microfinance programmes worldwide. Another example is a study on Bangladesh conducted by Khandker (2003) which provides evidence that the drop-out rate in the examined programmes[25] is around 30%. A study by Alexander-Tedeschi and Karlan (2007) on Peru finds that the drop-out rate is about 56%.

Furthermore, the Round 3 data – which are the data collected in the fieldwork for this study - has certain shortcomings which are discussed before reporting the results in detail. The sample size of 220 households is rather small because of budget and time constraints. As mentioned earlier, a further round of the panel could not be collected as neither the original panel sample SEWA Bank members nor the controls could be

[25] Khandker (2003) examined the microfinance programmes of GB, BRAC and the Bangladesh Rural Development Board (BRDB). His study is discussed in chapter 5.

identified. Hence, the Round 3 data is not an authentic panel follow-up of the original study and comparability to the USAID data is therefore limited. Another point is that PSM requires rich and high quality data sets (Smith and Todd, 2005). However, there are many missing data, in particular in the social capital and the housing improvements section, which will limit the explanatory power of the Round 3 data as sections 4.7. and 4.8. illustrate.

4.5.1. Drop-outs

SEWA Bank does not keep any adequate records of drop-outs. They argue that there are no drop-outs in SEWA Bank which is a rather intriguing statement bearing in mind the drop-out figures presented earlier. This made it very difficult to sample drop-outs. In the end, my enumerators asked current SEWA Bank clients they had interviewed whether they knew of any households that had dropped out of SEWA Bank and who could be approached. Finding those drop-outs was a time-consuming and tedious process and as a result only 10 drop-out households could be surveyed. The lack of microfinance impact evaluations which address the issue of drop-outs might have something to do with the fact that they are difficult to trace and that their existence is often denied. To conclude, it appears that low-cost and small surveys such as the Round 3 survey do not necessarily add value and do not provide accurate impact estimates. The limitations outlined above should be understood before looking at the Round 3 results in more detail.

4.6. Sampling procedure

As mentioned in section 4.5., Round 3 followed the USAID study design as much as possible and collected a cross-section data set on 220 households out of which 70 were borrowers, 70 were savers and 70 were non-clients as a control group. In addition, a drop-out sample of 10 was collected. The sampling criterion required sampling economically active women aged 18 and above. The sample was determined following a three-step process as in the case of USAID.

In a first step, the same geographical area was selected. As mentioned earlier, the USAID study drew its sample from the following ten wards in Ahmedabad: Behrampura, Jamalpur, Bapunagar, Rakhial, Asarwa, Khadia, Amraiwadi, Saraspur, Raikhad and Dudheshwar. As outlined in section 4.5., Chen and Snodgrass (2001) argue that those ten wards represent the area in which SEWA Bank's work first began and in which the majority of SEWA clients still reside. It is assumed that the ward boundaries have not changed much since then and thus Round 3 sampled from the same geographical area.

In a next step, two client samples were selected (i.e. current borrowers and current savers) and an additional sample of drop-outs. The rationale for sampling borrowers as well as savers was discussed earlier in this chapter. At the time of the USAID study there were ten savers for every borrower, however, as of FY 2007, the ratio of borrowers to savers has slightly declined; there are now only seven savers for every borrower.

A random sample of clients was drawn from a list obtained from SEWA Bank providing details of all SEWA Bank borrowers from the ten wards mentioned above who took out loans during FY 2007. Over-sampling was done deliberately since it was assumed that not all of the sampled households would agree to be surveyed. The same procedure was then applied to obtain a random sample of savers. Savers should have made at least one deposit in a SEWA Bank savings account during FY 2007. Moreover, savers should not have taken out any loans in FY 2007. As in the case of USAID, replacements were made when the respondent could not be located, was not economically active, e.g. not self-employed anymore, or did not want to participate. Also, replacements were needed when respondents from the sample of current savers were not actively saving anymore or had taken out loans in FY 2008. In addition, a drop-out sample of ten households was collected. As mentioned earlier, tracing drop-outs was rather challenging and a rather non-random way of identifying them was applied, namely clients who have been interviewed were asked whether they would know of any former SEWA clients.

In a final step, the non-client sample was selected. As discussed earlier, USAID had conducted a pre-survey in the neighbourhoods of the two client samples which resulted in a list of 15,000 households from which a random sample of non-clients was drawn. Round 3 did not have the resources to repeat this exercise, thus the non-client sample was identified by conducting a mini census of the five houses that were adjoining or in the immediate neighbourhood of every client household that was surveyed. The non-client household was then identified by matching it with the client household on the basis of certain key characteristics, e.g. gender, age, primary livelihood activity, marital status, educational background, religion and caste. If the matched non-client household identified through this procedure declined to be interviewed and the next best match from the same cluster also declined to be interviewed, then the census area was expanded by another five houses. This approach generally worked since the ten wards that were sampled were reasonably homogeneous in terms of socio-economic background. This also applies for the sampling of the USAID non-client sample. Chen and Snodgrass (2001) argue that

> "like the client samples, the non-client sample consists of economically active women over age 18 engaged in one or more of a similar range of informal sector activities. Neighbourhoods in the older parts of Ahmedabad City are relatively homogeneous in terms of caste, occupation, and class. Given the homogeneity of the neighbourhoods, the range of economic activities open to non-client women in those neighbourhoods is roughly the same as those open to client women…" (p. 53).

However, this does not exclude the possibility that there was selection on unobservables, a point already made in section 4.5. As for the qualitative part of this study, the eight case studies respondents were randomly selected from the client and non-client sample with the aim to explore the possibility that there were in fact unobservable (to conventional survey instruments) entrepreneurial abilities or access to social networks, as well as to complement the social capital data collected through the survey questionnaire. The respondents mainly talked about their entrepreneurial activities, i.e. reasons for and characteristics of self-employment where applicable.

4.7. Estimation strategy

This study replicates the USAID analysis and subjects these data as well as the Round 3 data to PSM to control for selection bias in the hope of providing more robust impact estimates. As discussed in chapter 3, PSM matches participants to non-participants on the basis of observable characteristics and compares outcomes between the treatment sample and the sample of matches (Caliendo and Kopeinig, 2005 and 2008; Rosenbaum and Silber, 2001). The underlying assumption is that there is no selection bias due to unobservable characteristics, though, whether this assumption holds is questionable. This is examined using sensitivity analysis of the PSM results which can suggest what likelihood of selection on unobservables would be required to render the observed treatment effect statistically insignificantly different from zero (Rosenbaum, 2002). We have seen earlier that in the case of microfinance unobservables are very likely to be present.

To recap, the main drawback of PSM is that it only accounts for selection bias due to observable characteristics and disregards any biases that might occur due to unobservable characteristics (as discussed in chapter 3). Heckman et al (1998) argue that the biases due to observables and unobservables could possibly offset each other. In a well-known study, Dehejia and Wahba (1999) tested how well PSM does in comparison to a randomized experiment and concluded that PSM's overall contribution towards reducing biases is positive. In other words, the results provided by PSM are a good approximation to those obtained under an experimental approach (discussed in more detail in section 3.5.1.). Smith and Todd (2005), however, correct some of those claims made by Dehejia and Wahba (1999) and cast doubts about the appropriateness of PSM (as outlined in chapter 3). The results of sensitivity analysis suggest that the PSM estimates may not be robust to selection on unobservables.

It is argued, however, that PSM combined with DID (as discussed in chapter 3), which compares participants and non-participants before and after the intervention, can eliminate the effects of unobservable characteristics. However, using DID commonly requires a baseline data set which is in many cases not available. In addition, the panel data set has to be a 'true' baseline, i.e. the respondents should not have been microfinance participants at the time of the collection of the baseline data set, which in the instance of the USAID panel is not the case – this is discussed in more detail in the results section 4.8.4. Another approach which helps to control for biases due to unobservable characteristics is the IV approach as discussed in chapter 3. To recap, for this method to succeed instruments are required that influence programme participation without affecting outcomes given participation (Ravallion, 2001) but finding valid instruments is rather difficult as outlined in chapter 3; thus this method is not necessarily reliable for accurately measuring programme impacts (Heckman, 1997).

Nevertheless, the IV method was applied on the USAID study data to assess whether it would be an appropriate tool in this context. This required the identification of an appropriate set of instruments that would influence the decision to participate in a programme without affecting outcomes. Given the variables in the data set and the specific context of this study, caste, religion and age appeared to be adequate instruments (I also experimented with location and family composition, i.e. proportion of family members by sex and age). However, the outcome of the Hausman specification test that examined the differences between the OLS and IV estimates indicated that the latter estimates were insignificant. As a result, it can be concluded that the instruments used were inappropriate and that the IV method is in fact inadequate in the context of the USAID study. For this reason, this method is not further discussed in this chapter.

Therefore, this study proposes to apply PSM to control for selection bias in the hope of providing more robust impact estimates. To account for the drawbacks of PSM, i.e. its inability to control for the unobservables, a social capital section was added to the Round 3 questionnaire with the aim to illuminate the role of the unobservables.

To begin with, the empirical model is outlined. As mentioned earlier, the USAID study collected data on three sub-samples: borrowers, savers and controls. The objective is to assess the socio-economic impact of microfinance participation. The identification strategy can be expressed as follows; *i* stands for household in ward *j*:

$$(26) \qquad y_{ij} = C_{ij}\delta + X_{ij}\alpha + V_j\beta + \varepsilon_{ij}$$

<u>Where:</u>

y_{ij} = outcome on which impact is measured

C_{ij} = level of participation in microfinance, i.e. a membership dummy variable

δ = effect of the microfinance programme, main parameter of interest

X_{ij} = vector of household level characteristics

V_j = vector of ward level characteristics

α, β = parameters to be estimated

ε_{ij} = error term representing unmeasured household and ward characteristics that can influence outcomes

The characteristics of participants, i.e. borrowers and savers, are examined separately for Round 1 and for Round 2 using a logit model (Table 8). A treatment dummy denoting microfinance participation was created containing borrowers and savers to represent participants, i.e. C_{ij} as expressed in equation (26). This dummy is used as a dependent variable and assumes a value of 1 if an individual has self-selected into microfinance and a value of 0 if otherwise.

Table 8: Logit regression of probability of microfinance participation, without sampling weights[26]

Independent variables	**Round 1**	**Round 2**	**Round 3**
Age household head (years)	-0.040*** *0.000*	-0.002 *0.823*	0.041 *0.337*
Age respondent (years)	0.019* *0.057*	0.010 *0.341*	-0.078* *0.071*
Highest grade completed male	0.083*** *0.003*	0.001 *0.956*	-0.016 *0.724*
Respondent married (yes=1)	0.789*** *0.004*	0.760*** *0.008*	-2.712* *0.068*
Muslim (yes=1)	0.510** *0.013*	0.434** *0.030*	0.279 *0.468*
Upper Caste (yes=1)	-0.666*** *0.003*	-0.682*** *0.001*	-0.384 *0.441*
Household size	0.011 *0.920*	-0.074 *0.422*	0.083 *0.780*
Nuclear household (yes=1)	-0.453** *0.042*	-0.293 *0.174*	0.468 *0.353*
Non-SEWA savings	-0.000 *0.190*	-0.000 *0.695*	0.000 *0.788*
Constant	-0.360 *0.713*	-0.413 *0.679*	2.287 *0.459*
Number of observations	768	785	205
Pseudo R-squared	0.059	0.032	0.075

Source: Author's calculations.

Notes: p-values in italics. *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%. Please also note that the following control variables were included in the logit model: age of all individuals, age squared, highest grade completed household head, highest grade completed respondent, sex household head (male=1), number of adult male in household, number of household members aged 0-14, subnuclear household (yes=1), support network in place (Round 3 only); all insignificant.

Table 8 presents the logit regression estimating the probability of microfinance participation. The logit model is required to predict the propensity scores so that the matching procedure can be implemented – section 4.8. outlines this in more depth. The

[26] Since SEWA Bank members are more likely to be savers than borrowers, i.e. as mentioned earlier there were ten savers for every borrower in rounds 1 and 2; and seven savers for every borrower in Round 3, there is a case for using appropriate weights. A separate set of logit regressions across rounds 1-3 were computed adjusted for sampling weights. However, the use of sampling weights led to very minor changes, i.e. slightly lower pseudo R-squared values and lower significance levels for a few coefficients when sampling weights were applied. The results reported in this table do not consider sampling weights.

results presented in Table 8 show that the main variables associated with membership that are statistically significant in Round 1 and in Round 2 are being married, Muslim and upper caste. In addition, age of household head, age of respondent, highest grade completed male and nuclear household are significant in Round 1. Age of respondent and married are the only significant variables in Round 3 but surprisingly their coefficients are negative when compared to those in Round 1 and in Round 2. In fact, most of the coefficients in Round 3 have reversed signs when compared to the corresponding coefficients in Round 1 and in Round 2, i.e. they are positive when negative in Round 1 and in Round 2 and vice versa but mostly insignificant; hence it can be concluded that the Round 3 model does not yield any meaningful information which is not surprising keeping in mind both the smaller sample size and the problems encountered in conducting the survey, as described above. As a result, the Round 3 findings are not further discussed and only referred to in footnotes. Overall, few covariates are statistically significant and the values for the pseudo R-squared across all rounds are rather low which indicates that the model has limited explanatory power[27].

Finally, social capital type variables were included in the logit model for Round 3 in order to help illuminating the role of the unobservables but most of those variables were insignificant and hence were dropped as described in the notes section of Table 8. Quantifying social capital does not appear to be particularly straightforward. The qualitative evidence in the form of direct observations and case study interviews indicate a strong presence of vertical social capital. However, the survey data has failed to confirm this evidence; in fact, most of the social capital variables were insignificant and were characterised by missing data.

4.8. Results

The findings with respect to selected hypotheses are presented in this section. The individual and enterprise level hypotheses of the USAID study led to mixed and rather insignificant results. As argued earlier, measuring impact at these levels is

[27] I experimented with the logit model and tried various other control variables with the objective to enhance the explanatory power of the model but to no avail. The low explanatory power of the model has implications for the reliability of my PSM results; this is further investigated in section 4.8.3. where sensitivity analysis is introduced.

unsatisfactory, and the household level is the most promising way to obtain meaningful impact estimates. Moreover, Round 3 focused on collecting household level data and disregarded gathering data on the individual and enterprise level. Hence, this section focuses on selected household level results only, i.e. on income, housing expenditure and children's education (Table 9). The detailed and re-analysed results of all household level as well as individual and enterprise level hypotheses of the USAID study as well as the Round 3 data can be found in Appendix 2. Furthermore, detailed descriptive statistics outlining the mean values for all outcome variables across rounds 1 - 3 by type of study participant can be found in Appendix 1.

Firstly, the selected household level results of the USAID study are replicated. Replication is an important step in validating results (Hamermesh, 2007). Hence, the USAID data[28] were subjected to ANOVA and ANCOVA. My replication closely reproduced the USAID study results and is thus not discussed further.

Next, PSM is employed on the USAID data to gauge whether more advanced econometric techniques than ANOVA and ANCOVA, which claim to account for selection bias (Chen and Snodgrass, 2001), would produce different results. The original household level results of the USAID study are compared with the results obtained when PSM was applied using 5-nearest neighbour matching and kernel matching with a bandwidth of 0.01 (Table 9). The results for participants, i.e. borrowers and savers together, versus controls are presented first. Further sub-group comparisons are presented later in this chapter. The drop-out sample of 10 households is too small to be meaningful and has not been analysed further. It is suggested that further research on the drop-out issue might be useful to better understand microfinance realities and to improve the products and services of the respective MFI.

Before discussing the results presented in Table 9 and Table 12 in more detail, a few remarks with regard to the implementation of PSM are required. As mentioned in chapter 3, the basic idea of matching is to compare a participant with one or more non-participants who are similar in terms of a set of observed covariates X (Caliendo and Kopeinig, 2005 and 2008; Rosenbaum and Silber, 2001). This requires predicting

[28] The data sets of all three USAID studies can be downloaded here: http://www.microlinks.org/ev_en.php?ID=4678_201&ID2=DO_TOPIC

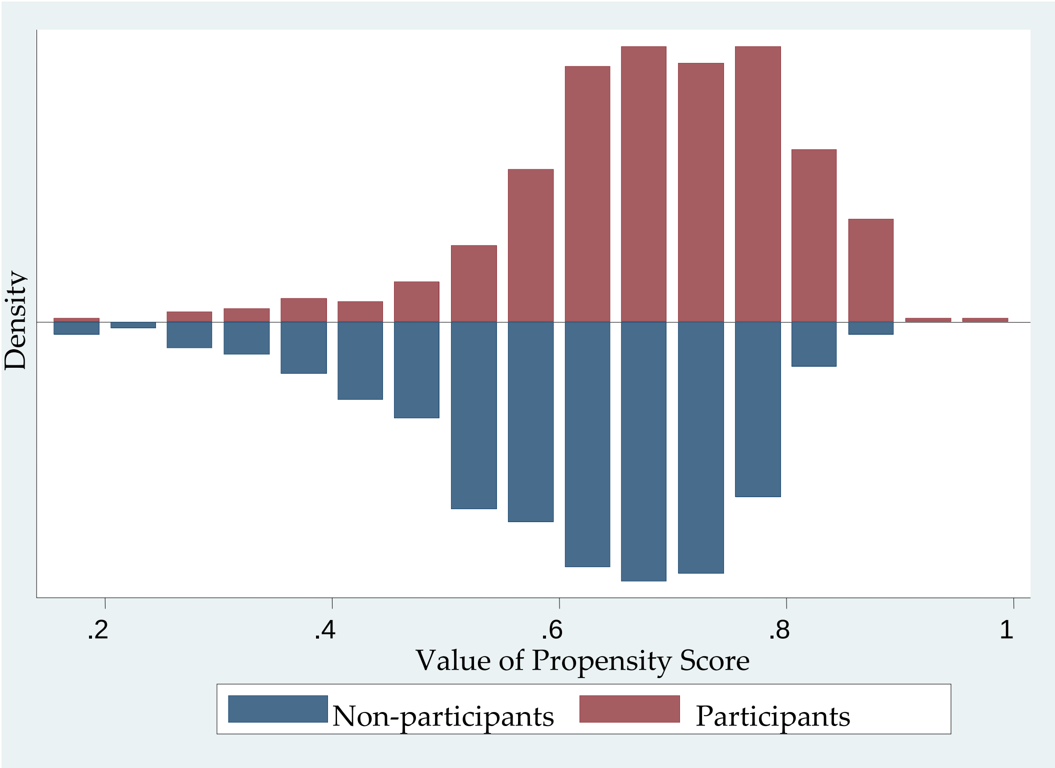
propensity scores for each individual, i.e. participants as well as non-participants using a logit or a probit model. I used the logit model presented in Table 8 to predict those propensity scores. Then, before implementing the actual matching process, I examined whether the propensity scores I had obtained for participants and non-participants fulfil the common support assumption. To recap, Caliendo and Hujer (2005) express the common support assumption as follows:

$$0 < \Pr\,(D = 1 \mid X) < 1 \qquad (27)$$

This assumption indicates whether treatment and control groups provide equal support of X (Caliendo, 2006). This can be investigated graphically. Figure 5 presents the distribution of the propensity score for participants as well as non-participants; it shows that each participant with a certain propensity score has a corresponding non-participant. In other words, if the propensity scores for participants and non-participants overlap reasonably well, then the common support assumption is satisfied and it is recommended comparing those two groups. Next, the differences in the outcome variables for participants and their matched non-participants are calculated (Morgan and Harding, 2006) as presented in Table 9 and Table 12. Furthermore, I used t-tests[29] before and after matching for all results presented in Table 9 and Table 12 to examine the differences of the mean values for each covariate X across treatment and control groups. In addition, those t-tests calculated a 'bias' defined as the mean value of the treatment group and the (matched/unmatched) control group divided by the square root of the average sample variance in the treatment group and the (matched/unmatched) control group. The t-tests I employed indicate that the differences between treatment and control groups as well as the 'bias' were reduced considerably in most cases, hence the matching process was successful in generating a control group that was reasonably similar to the treatment group. Therefore, the use of PSM is justified in this case. As a consequence, I conclude that the balancing properties of the propensity scores were satisfied in all cases.

[29] The STATA command pstest was used.

Figure 5: Distribution of propensity scores for participants and non-participants



Source: Author's calculations.

Table 9: PSM impact estimates of selected household level results – microfinance participants versus controls; without sampling weights[30]

	Round 1	Round 2	Round 3
Total household income per annum in Rupees			
USAID	10,090***	15,302***	N/A
PSM - 5 nearest neighbour matching	8,944***	14,635***	7,030
PSM - kernel matching, bandwidth 0.01	8,638***	13,786***	9,355*
Total household income per annum per capita in Rupees			
USAID	2,063***	2,685***	N/A
PSM - 5 nearest neighbour matching	2,019***	2,486***	1,805
PSM - kernel matching, bandwidth 0.01	1,913***	2,537***	2,222
Expenditure for housing improvements in Rupees			
USAID	3,748***	5,871	N/A
PSM - 5 nearest neighbour matching	3,701***	6,546***	1,150
PSM - kernel matching, bandwidth 0.01	3,484***	6,504***	1,191
School enrolment for girls aged 5 to 10 years			
USAID	-0.020	-0.005	N/A
PSM - 5 nearest neighbour matching	0.011	0.052	0.029
PSM - kernel matching, bandwidth 0.01	0.010	0.028	-0.012
School enrolment for boys aged 5 to 10 years			
USAID	0.065	0.005	N/A
PSM - 5 nearest neighbour matching	-0.027	0.021	-0.057
PSM - kernel matching, bandwidth 0.01	-0.007	-0.004	-0.042
School enrolment for girls aged 11 to 17 years			
USAID	0.015	-0.015	N/A
PSM - 5 nearest neighbour matching	0.028	0.012	0.014
PSM - kernel matching, bandwidth 0.01	0.006	0.009	-0.031
School enrolment for boys aged 11 to 17 years			
USAID	-0.075	-0.020***	N/A
PSM - 5 nearest neighbour matching	-0.025	-0.012	-0.000
PSM - kernel matching, bandwidth 0.01	-0.045	-0.019	0.000

Source: Author's calculations.

Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%. The results in this table refer to the differences in the mean values between matched

[30] With reference to footnote 26, the literature is unclear with regard to accommodating sampling weights in the context of matching. 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, as with the logit regressions presented in Table 8, the analysis was re-run taking sampling weights into account and the balancing properties of the independent variables were investigated. The balancing properties were satisfied in both cases, i.e. with and without sampling weights. Therefore, the analysis across rounds 1-3 was conducted with and without sampling weights and the results obtained in both cases were consistent with each other. This suggests a certain degree of robustness of the results presented in this table. A decision was then made to present the results without sampling weights in Table 9.

samples; they were obtained using the STATA command psmatch2. I also ran the STATA command pscore with the objective to cross-check the psmatch2 results across the various matching algorithms. The results I obtained from the different STATA routines displayed minor differences in terms of the size of coefficient and the level of significance. As mentioned in chapter 3 and as argued by Morgan and Winship (2007), matching results can vary depending on the matching algorithm and PSM routine applied. Results are bootstrapped.

4.8.1. Income[31]

The results for income per annum[32], income per annum per capita[33] and expenditure for housing improvements are positive and statistically significant across Round 1 and Round 2. Those results reflect the differences between participants and non-participants. For example, according to the USAID Round 1 result, income per annum was higher by 10,090 Rupees for microfinance participants than for non-participants whereas the PSM results applying 5-nearest neighbour matching show that income per annum increased by 8,944 Rupees for participants. Similarly, the PSM results applying kernel matching with a bandwidth of 0.01 display an increase in income per annum by 8,638 Rupees for microfinance participants. It can be seen that the degree of impact depends on the econometric technique applied but even when the same technique is applied, i.e. PSM, the impact estimates still vary - though not significantly so - because of the different matching algorithms applied. For example, kernel matching estimates

---

[31] All income figures throughout this study have been deflated to January 1998 prices by using a deflator of 1.156 – as mentioned in the USAID study. This was the value of the Consumer Price Index (CPI-IW) for Ahmedabad in January 2000, expressed on a base of January 1998.

[32] Income per annum is confounded with household size and hence an unreliable measure of outcome. However, since USAID assesses the impact on income per annum and since I am comparing their results with mine, I will continue to report income per annum results throughout the chapter. Nonetheless, the reader should treat those results with caution.

[33] An additional calculation not reported here was completed for total household income per annum per capita across Round 1 and Round 2 as presented in Table 9; I made adjustments using an equivalence scale. Equivalence scales commonly allow the comparison of per capita income of households of various sizes and compositions on an equal basis. A range of equivalence scales exist and choosing one is a rather arbitrary process. The following equivalence scale adjusting for the various household sizes and compositions is used here: $(A + PK)^F$; where A = number of adults ≥ 18, K = number of children < 18, P = 0.7 which is the recommended percent value indicating how much each child contributes to the households consumption relative to the adults, and F = 0.65 – 0.75, a factor that accounts for economies of scales (Source: http://www.irp.wisc.edu/research/method/oakvos.htm). The application of this formula led to a minor increase in the size of the coefficient of per capita income per annum but the significance level remained the same. Since there is no clear recommendation as to which one of the equivalence scales to use and their application is debated, I decided to report total household income per annum per capita without making any adjustments.

with a bandwidth of 0.01 for income per annum and for income per annum per capita are lower than the respective 5-nearest neighbour matching estimates for Round 1 and for Round 2. Also, when the bandwidth increases in the case of kernel matching, the impact estimate tends to increase as well. However, these observations cannot be applied to all outcome variables or rounds as Table 9 clearly demonstrates.

Overall, in the case of income per annum, income per capita and expenditure for housing improvements across Round 1 and Round 2, the general trend and the statistical significance are similar across USAID and PSM results. In other words, the results obtained by applying PSM appear to support the original findings of USAID with the exception that the USAID results for expenditure for housing improvements in Round 2 were not statistically significant while the PSM results are statistically significant[34].

4.8.2. School enrolment

The school enrolment figures are rather inconclusive; the PSM results for school enrolment for girls aged 5 to 10 across Round 1 and Round 2 display a positive trend; i.e. participants do better than non-participants but none of the results are statistically significant. The respective USAID results are negative but also insignificant. The results for school enrolment for girls aged 11 to 17 across Round 1 and Round 2 are equally meaningless. USAID argues that there is an increase in school enrolment in Round 1 but the impact estimate is suddenly negative in Round 2. The PSM results are all positive but insignificant[35].

The picture does not change dramatically when looking at school enrolment figures for boys aged 5 to 10 and 11 to 17. According to the PSM results across all data collection rounds, microfinance has negative impacts on the school enrolment of boys aged 5 to 10 with the exception of one value in Round 2 obtained by applying 5-nearest neighbour matching which implies a negligible positive impact. The USAID results, on the other hand, indicate a positive impact but none of the results are statistically significant. Chen and Snodgrass (2001) argue that most boys in the age group 5 to 10

[34] With regard to Round 3 data, the results for the first three outcome variables are positive but only one of them is statistically significant.

[35] As expected, Round 3 data is not particularly helpful; the size of the impact varies with the matching algorithm applied and the results are statistically not significant.

are in fact already enrolled in school irrespective of microfinance participation. All enrolment figures for boys aged 11 to 17 are negative across USAID and PSM results for Round 1 and for Round 2 with one figure being statistically significant. Hence, it can be concluded that microfinance participation does not seem to have any significant impact on children's education[36].

Overall, the most notable result is that there seems to be a positive impact on total income per annum and per capita as well as on expenditure for housing improvements. Before applying panel methods, the quality of the matching results are assessed using sensitivity analysis as explained in chapter 3.

4.8.3. Sensitivity analysis

The impact evaluation of SEWA Bank needs to answer the question whether the apparent effect of membership compared to the control group is due to the saving and borrowing enabled by membership of SEWA Bank or to some unobserved characteristic of members compared to the control group, such as entrepreneurial abilities, access to social networks, etc. PSM allows to control for observable characteristics included in the propensity score on which members and controls are matched, but it cannot control for unobservables (Caliendo and Kopeinig, 2005 and 2008; Rosenbaum and Silber, 2001). Further to the explanation in chapter 3, this section provides additional information on sensitivity analysis.

Rosenbaum (2002) developed the "conceptual advance" (ibid, p. 106) of Cornfield et al (1959) that the robustness of the estimate of the difference in outcome between treatment and control groups (the impact estimate) could be assessed by asking what magnitude of selection on unobservables (hidden bias) one would need in order to explain away the observed impact, thus:

> "[I]f the association [*Author's note: between treatment and outcome*] is strong, the hidden bias needed to explain it is large" (Rosenbaum, 2002, p. 106).

In the context of death from lung cancer for smokers and non-smokers Cornfield et al (1959) suggested that if the ratio of the likelihood of death from lung cancer for smokers to the likelihood of death from lung cancer for non-smokers was high then a

[36] Again, Round 3 data neither confirms nor contradicts any of the hypotheses tested here and has little explanatory power.

similar high ratio for the unobserved characteristic(s) would be required to make this unobserved characteristic the true cause of the higher prevalence of death from lung cancer by smokers.

Further to the discussion in chapter 3, Rosenbaum (2010) explains that

> "a sensitivity analysis in an observational study asks how the conclusions of the study might change if people who looked comparable were actually somewhat different…" (p. 367).

In other words, the objective of sensitivity analysis is to explore whether the matching estimates are robust to selection on unobservables (Rosenbaum, 2002). Ichino, Mealli and Nannicini (2006) argue that "sensitivity analysis should always accompany the presentation of matching estimates" (p. 19).

Rosenbaum (2002) invites us to imagine a number Γ (gamma) (≥ 1) which captures the required degree of association, of an unobserved characteristic with the treatment, for it (the unobserved characteristic) to explain the observed impact. Γ is the ratio of the odds[37] that the treated have this unobserved characteristic to the odds that the controls have this characteristic[38].

This approach can be implemented using the rbounds procedure in STATA[39]; this procedure uses the data to calculate the confidence intervals (for a given level of

---

[37] Odds, which are widely used in assessing probabilistic outcomes, are derived from probabilities ($0 \leq \pi_i \leq 1$) by the following formula: $\pi_i/(1-\pi_i)$.

[38] Suppose two individuals j & k who are closely matched on observables so that $x_j = x_k$, but for whom p_j not equal to p_k - i.e. probability of being selected into SEWA Bank is not the same despite being equivalent on observables. The probability of being selected can be expressed as an odds ratio (the odds of probability of j/k (p_j/p_k) being selected $p_j/(1-p_j)$ or $p_k/(1-p_k)$). Then imagine there is a number Γ (gamma) such that $1/\Gamma \leq \{p_j(1-p_k)\}/\{p_k(1-p_j)\} \leq \Gamma$, then if Γ =1 $p_j=p_k$ (i.e. there is no difference in the odds of being selected). Γ =2 means that individual j is twice as likely to be selected into SEWA Bank as individual k. This might be considered not unlikely based on my observations of the selection process operated by SEWA Saathi, and by my understanding of the requirements of households to be able to save, and for other to qualify for borrowing.

[39] Rosenbaum's (2002) bounding approach can also be implemented by using the STATA routines mhbounds developed by Becker and Caliendo (2007) applicable for binary outcome variables. rbounds written by DiPrete and Gangl (2004) is commonly used for continuous outcome variables. Both routines are based on the STATA command psmatch2 which was developed by Leuven and Sianesi (2003). Most of the outcome variables of the SEWA Bank data are continuous; hence the STATA command rbounds is employed. Furthermore, Ichino, Mealli and Nannicini (2006) suggest a simulation-based approach which builds on Rosenbaum and Rubin (1983) and Rosenbaum (1987) to yield information on the robustness of the matching

confidence – e.g. 95%) of the outcome variable for different values of Γ. A value of Γ that produces a confidence interval that encompasses zero is one that would make the estimated impact not statistically significant at the relevant level of confidence. If Γ is relatively small (say < 2) then one may assert that the likelihood of such an unobserved characteristic is relatively high and therefore that the estimated impact is rather sensitive to the existence of unobservables (DiPrete and Gangl, 2004). If there is other evidence that there may be unobservables, such as my qualitative observations of the SEWA Bank selection processes, we cannot be confident that the estimated impact is not due to unobservables.

We can illustrate this approach by calculating Γ at which the estimated impact of SEWA Bank membership on household income per capita for Round 1 is no longer statistically significant. Table 9 shows that the 5-nearest neighbour matching estimate for total household income per annum per capita in Round 1 is 2,019 Rupees which is significant at 1%. This suggests that households participating in microfinance earn significantly more income per annum per capita than control households; however, this may not be due to membership *per se* but unobserved characteristics that account for membership (and or its impact). Sensitivity analysis explores the robustness of this impact estimate and demonstrates how it changes in the presence of selection on unobservables. The STATA procedure rbounds reports the estimates[40] and their 95% (or other) confidence intervals for matched pairs of SEWA Bank members and controls (see Table 10).

When Γ = 1 there is no selection on unobservables. If Γ increases to 1.2, then matched individuals differ in their odds of exposure to microfinance by a factor of 1.2 due to selection on unobservables. Table 10 shows that when Γ = 1.2 the statistical significance level ranges from < 0.0001 to < 0.0046. This implies that in this case selection on unobservables is not likely to explain the observed association between exposure to

---

estimates. The STATA routine sensatt developed by Nannicini (2007) and which is based on the STATA command pscore written by Becker and Ichino (2002), is commonly used to implement the approach advocated by Ichino, Mealli and Nannicini (2006). Other approaches to sensitivity analysis exist, e.g. the 'Lechner bounds' developed by Lechner (2000) which assesses the sensitivity of the matching estimates with regard to the common support assumption.

[40] 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 9 which provides the average treatment effects.

microfinance and higher income levels. However, when $\Gamma = 1.3$ or more, a relatively small difference in the odds of exposure implying that it is quite likely that such an unobserved confounding variable exists, the 95% confidence interval of the point estimates encompasses zero. Consequently, we can argue that the observed impact of SEWA Bank membership on household income per capita is not significantly different from zero, and the association between microfinance exposure and higher income levels may well be due to unobservables.

Table 10: Sensitivity analysis for household income per annum per capita in Rupees for microfinance participants for Round 1

	Significance levels		**Hodges-Lehmann point estimates**		**95% Confidence intervals**	
Gamma (Γ)	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
1	< 0.0001	< 0.0001	953	953	520	1,403
1.2	< 0.0001	< 0.0046	559	1,357	141	1,834
1.3	< 0.0001	< 0.0330	391	1,545	-27	2,033
1.4	< 0.0001	< 0.1292	241	1,717	-174	2,220
1.5	< 0.0001	< 0.3181	98	1,876	-311	2,394
1.6	< 0.0001	< 0.5561	-31	2,036	-436	2,562
1.7	< 0.0001	< 0.7637	-148	2,188	-556	2,732
1.8	< 0.0001	< 0.8967	-261	2,328	-669	2,891

Source: Author's calculations.
Notes: see footnote 40. The table shows magnitude of selection on unobservables, range of significance levels, Hodges-Lehmann point estimates and confidence intervals.

Sensitivity analysis was conducted on all the outcome variables presented in Table 9 enabling to test the sensitivity of all impact estimates on the household, enterprise and individual level across Round 1 and Round 2. The evidence provided by those tests are in agreement with the above description, namely that the impact estimates presented in Table 9 are sensitive to selection on unobservables and should be treated with caution[41]. In other words, sensitivity analysis suggests that the impact estimates presented here might be overstated due to the presence of selection on unobservables. However, this is considered a worst case scenario and does not prove that there is no

[41] The detailed results from those sensitivity tests are not presented here but the relevant STATA do-files can be made available upon request.

effect since it assumes both that the unobserved variable has the specific effect on the odds ratio of treatment and that it has a strong effect on the outcome variable (DiPrete and Gangl, 2004, p. 291). Nevertheless, PSM and related tests allow the quantification of selection on unobservables which is helpful. These results lead me to concur with Ichino, Mealli and Nannicini (2006), namely that sensitivity testing should always complement the presentation of matching estimates. In this case, caution in concluding that SEWA Bank membership has a causal effect on per capita income is warranted.

4.8.4. Panel analysis

The panel data analysis with or without PSM reveals nothing new and broadly confirms the results obtained from the cross-section analysis. As illustrated in Table 11, PSM using nearest neighbour matching on Round 1 data caused some households which did not match on observable characteristics to be dropped, and only matched households were merged with Round 2 data. Using the treatment and matched households a regression-adjusted DID model was run on all outcome variables as set out by the following equation which is a fixed effects linear regression model; i stands for household in ward j at period t:

$$y_{ijt} = \alpha_i + \delta_t + \beta \mathrm{C}_{it} + \theta \mathrm{X}_{it} + V_j + \varepsilon_{ijt} \qquad (28)$$

Where:

y_{ijt} = outcome on which impact is measured at period t

C_{it} = level of participation in microfinance, i.e. a membership dummy variable, in period t

X_{it} = vector of household level characteristics in period t

V_j = vector of ward level characteristics

α_i = fixed effects unique to household i

δ_t = period effect common to all households in period t

β, θ = parameters to be estimated

ε_{ijt} = error term representing unmeasured household and ward characteristics at period t

Some evidence was found that there are positive and significant impacts at the household, enterprise and individual level as outlined by Table 11.

Table 11: PSM and DID results - impact of microfinance participation; without sampling weights[42]

Household level hypotheses	
Total household income per annum in Rupees	11,287***
Total household income per annum per capita in Rupees	2,181***
Inverse Simpson index	0.121
Expenditure for housing improvements in Rupees	5,628***
Expenditure on household assets in Rupees	734**
School enrolment for girls aged 5 to 10 years	-0.019
School enrolment for boys aged 5 to 10 years	0.019
School enrolment for girls aged 11 to 17 years	0.025
School enrolment for boys aged 11 to 17 years	-0.011
Food expenditure per day per capita in Rupees	0.93
Enterprise level hypotheses	
Informal sector income of whole household - per month in Rupees	3,091**
Informal sector income of respondent only - per month in Rupees	1,876**
Microenterprise revenues of all enterprises in household - per month in Rupees	3,050**
Microenterprise revenues of microenterprises for which respondent is primarily responsible - per month in Rupees	1,559**
Current value of fixed assets of all microenterprises in household in Rupees	211
Current value of fixed assets of microenterprises for which respondent is primarily responsible in Rupees	482
Hours worked in previous week in all microenterprises in household	13.78***
Days worked in previous month in all microenterprises in household	10.17***
Main types of suppliers - inferior[43] suppliers? Yes=1, No=0	0.060*
Main types of customers - inferior[44] customers? Yes=1, No=0	0.083**
Individual level hypotheses	
Respect by other household members? Yes=1, No=0	0.015
Prepared to deal with future? Yes=1, No=0	0.019

Source: Author's calculations.

Note: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%. The STATA procedure xtreg was applied to implement DID.

[42] As discussed in footnote 26 and 30.

[43] Individuals/households and retailers are inferior sources of supply as defined by Chen and Snodgrass (2001).

[44] Individual consumers are considered to be inferior customers as defined by Chen and Snodgrass (2001).

These findings are broadly in agreement with the results that USAID presented. However, caution is required when interpreting the panel data findings. One would have expected some differences between the cross-section and panel data results but since there are no differences between the panel and the cross-section, one possible interpretation is that unobservables have not affected impact estimates. However, one could argue that the USAID data set is not a real panel with a 'true' baseline because Round 1 respondents were already microfinance clients when the baseline data set was collected. The same clients and control households were then re-surveyed two years later. Strictly speaking, the baseline should have collected data on households that were not participating in microfinance at the time of the baseline data collection but became microfinance clients between survey rounds. This would have allowed a before and after comparison which would have been better suited to the analysis because it would have been possible to compare the treatment and control households in terms of all observables including outcomes; this would have allowed one to assess whether these two samples were broadly similar in these terms, although it would still not be possible to control for unobservables which affect response to the treatment.

4.8.5. Summary

Most of the PSM results confirm the findings of the USAID study if one ignores unobservables. What does this outcome mean for the issue of selection bias and the utility of PSM? Chen and Snodgrass (2001) argue that ANCOVA using a suitable control group accounts for selection bias to a certain degree. Based on the PSM results presented in Table 9, the first impression is that this assessment is indeed accurate. However, doubts remain as there are strong qualitative and theoretical reasons to think that unobservables have not been fully controlled for. This notion is confirmed by the sensitivity analysis which shows that the matching estimates are quite sensitive to selection on unobservables.

Moreover, there is a selection or screening process at work as indicated by the quantitative and qualitative evidence presented in this chapter which is driven by the unobservables, i.e. social capital, which determines who becomes a participant in microfinance and who remains a non-participant. Ito (2003) has provided credible evidence that horizontal as well as vertical social capital does play a role in this screening process and my own qualitative evidence presented in Box 1 supports this view. As a result, a social capital section was added in Round 3 which has made an attempt to capture the extent of social capital in microfinance but failed to provide convincing evidence due to the problems of measuring social capital and with survey execution.

The doubts about the reliability of the matching estimates as well as the discrepancies between the quantitative and the qualitative results indicate that the unobservables may not have been controlled for by any of the techniques applied. Thus, neither ANCOVA nor PSM have succeeded in accounting for selection bias. This is not too surprising at least in the case of PSM since its drawbacks are well-known although still debated (see the debate between Dehejia and Wahba, 1999 and Smith and Todd, 2005 as discussed in chapter 3).

Also, the quality of the matches is doubtful, considering PSM requires rich and large data sets in order to function properly (Heckman, Ichimura and Todd, 1997; Heckman et al, 1998; Smith and Todd, 2005). Moreover, the panel does not resolve the issue because it is not a 'true' panel, and, even if it were, might not control for the effects of unobservables. Microfinance clients might have been better off than non-clients even before participating in microfinance, i.e. in terms of access to social networks, wealth, skills or motivations. This may in turn have led them to self-select or to be selected into microfinance either by their peers or the staff of the microfinance organisation, and to be able to benefit more from membership than otherwise observationally similar households.

4.8.6. Sub-group comparisons

Having reached these preliminary conclusions with regard to membership of SEWA Bank, whether as saver or (saver and) borrower, sub-group comparisons were conducted to understand the impact of savings compared to saving and borrowing. As argued in section 4.3. and 4.4., Ghate (2007), as well as many microfinance practitioners, believe that the poor need savings more than credit. The following comparisons were investigated: borrowers versus controls, savers versus controls, borrowers versus savers, one-time borrowers versus savers, repeat borrowers versus savers, one-time borrowers versus controls and repeat borrowers versus controls.

Again, only the key findings are presented. The results of the comparisons of the various borrower groups with savers are similar to the various borrower group comparisons with controls in terms of absolute numbers and level of significance, hence only the latter comparisons are discussed (Table 12) (details of all sub-group comparison results are available in Appendix 3).

Table 12: Selected household level PSM results – sub-group comparisons; without sampling weights[45]

	Total household income per annum		**Total household income per annum per capita**		**Expenditure for housing improvements**	
	5-nearest neighbour	Kernel matching, bandwidth 0.01	5-nearest neighbour	Kernel matching, bandwidth 0.01	5-nearest neighbour	Kernel matching, bandwidth 0.01
Borrower versus control						
Round 1	12,323***	12,323***	2,364***	2,347***	5,046***	5,069***
Round 2	17,915***	18,256***	3,222***	3,378***	8,137**	8,160**
Round 3	1,084	3,806	885	1,459	2,652	2,615
Borrower versus saver						
Round 1	9,152***	9,020***	1,567**	1,405*	3,700**	3,349**
Round 2	11,014***	10,141***	1,634**	1,634**	4,547	4,115
Round 3	-9,594	-1,164	-1,823	-192	3,045*	3,045*
Saver versus control						
Round 1	7,236**	6,472**	1,545**	1,431**	2,212**	1,858*
Round 2	10,162***	10,085***	1,899***	1,909***	5,044***	4,508*
Round 3	3,767	4,299	311	758	-171	-171
One-time borrower versus control						
Round 1	11,196**	12,212***	1,998**	2,186**	5,125***	5,107***
Round 2	30,099***	27,700***	5,500***	5,669***	18,619*	18,468*
Round 3	171	5,038	642	1,574	2	-2
Repeat borrower versus control						
Round 1	17,556***	15,738***	3,410***	3,203***	6,059***	6,027***
Round 2#	2,319		1,112		-825	
Round 3#	23,800		3,067		29,866*	

Source: Author's calculations.

Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%. All figures in Indian Rupees. The results in this table refer to the differences in the mean values between matched samples; they were obtained using the STATA command psmatch2. I also ran the STATA command pscore with the objective to cross-check the psmatch2 results across the various matching algorithms. The results I obtained from the different STATA routines displayed minor differences in terms of the size of coefficient and the level of

[45] As discussed in footnotes 26, 30 and 42, the literature is unclear with regard to accommodating sampling weights in the context of matching. Hence, as before, the analysis was re-run taking sampling weights into account and the balancing properties of the independent variables were investigated. The balancing properties were satisfied in all cases, i.e. with and without sampling weights. Therefore, the analysis across all sub-group comparisons across rounds 1-3 was conducted with and without sampling weights and the results obtained in both cases were conclusive with each other, slight variations in terms of size of coefficient and level of significance could be observed. This suggests a certain degree of robustness of the results presented in this table.

significance. As mentioned in chapter 3 and as argued by Morgan and Winship (2007), matching results can vary depending on the matching algorithm and PSM routine applied. # No values for kernel matching in Round 2 and in Round 3, the sample was too small with propensity scores outside the common support region, no adequate matches were found. Results are bootstrapped.

The discussion of sub-group comparisons focuses on selected household level hypotheses, namely income per annum, income per annum per capita[46] and expenditure for housing improvements. The outcome variables with regard to children's education were dropped because their results were mostly insignificant across all rounds and across all sub-group comparisons, hence confirming the earlier findings of the cross-section and panel data analysis. Again, 5-nearest neighbour matching as well as kernel matching with a bandwidth of 0.01 were the matching algorithms of choice.

The PSM results in Table 12 indicate that borrowers do significantly better than controls across all three outcome variables. In detail, in the borrower versus control comparison the results of the outcome variables income per annum and income per annum per capita are consistent across Round 1 and Round 2 in terms of size of impact and level of significance but with slightly higher absolute impact figures in Round 2. This suggests that impact strengthens over time; i.e. the longer a client is participating in microfinance the more likely he or she is to reap the benefits[47], but these additional advantages are minor once a saver has become a borrower. Similar trends can be observed in the borrower versus saver comparison where the size of the coefficients is slightly lower than in the borrower versus control comparison.

Similarities can also be observed in the savers versus control comparison where the outcome variables income per annum and income per annum per capita are consistent across Round 1 and Round 2 in terms of size of impact and level of significance but their absolute impact figures are slightly lower than the ones reported in the borrower versus control comparison. More detailed results can be found in Appendix 3 and the

[46] Total household income per annum per capita was re-calculated using the formula for the equivalence scale described in footnote 33. As in the case of the results presented in Table 9, the equivalence scale adjusted results led to a minor increase in the size of the coefficient of per capita income per annum but the significance level remained the same, hence following the earlier procedure the results that are not adjusted are reported in Table 12.

[47] The Round 3 figures are all insignificant and in absolute terms show less positive impact than round 1 and 2 figures.

figures presented here merely serve as an illustration demonstrating that savers do better than controls but are slightly worse off than borrowers, which is to be expected. As mentioned earlier, the SEWA Bank model is built around mobilising savings. As of FY 2007, SEWA Bank had on average seven times more savers than borrowers. Hence, it appears that focusing on a savings approach is indeed a desirable strategy since savers have significantly higher impact estimates than control group members. This is in contrast to Berg's (2010) finding who finds no significant impact on household income when comparing savers versus controls. Without having access to her re-constructed data set and code I can only speculate as to why this is the case, e.g. differences in the variable construction might play a role as well as differing estimation strategies.

However, no clear picture seems to emerge when comparing one-time borrowers versus controls with repeat borrowers versus controls; in part this is because sample sizes are small. Comparing Round 1 figures only, it appears that repeat borrowers do significantly better than one-time borrowers who do better than controls, and both do better than savers. When using Round 2 figures the reverse appears to be true, with repeat borrowers worse off than one-time borrowers; and savers[48].

The evidence provided in Table 12 and Appendix 3 was subjected to sensitivity analysis (only significant matching estimates were tested). The results concur with those presented earlier in this chapter, namely that the matching estimates are sensitive to selection on unobservables[49]; i.e. the results of the sub-group comparisons potentially overstate the impact of microfinance participation.

4.9. Conclusion

This study contributes to the impact evaluation literature by providing new insights from re-analysing the existing USAID panel data with PSM and DID; it contributes to the microfinance literature by throwing doubt on the claims of impact of a well known

---

[48] A word of caution, Round 1 and Round 3 define repeat borrowers as borrowers who have taken out more than two loans. In Round 2, repeat borrowers refer to borrowers who have repaid their earlier loan and have taken out a new loan between survey rounds. There are only 56 repeat borrowers between the two survey rounds. The definition of repeat borrowers differs across rounds which would explain the inconsistency of the results. The sample size of 56 is simply too small to provide any meaningful PSM results.

[49] The detailed results from those sensitivity tests are not presented here but the relevant STATA do-files can be made available upon request.

microfinance project. The basic PSM results presented in this chapter approximate those obtained by USAID, i.e. borrowers do better than savers who in turn do better than controls. Presented in this way, these findings broadly support the existing belief that savings by themselves are desirable and that savings tools are complementary to a credit approach. The evidence is inconclusive whether repeat borrowers do better than one-time borrowers, and it appears that this is not always the case. The estimates obtained from the repeat borrower versus control comparison are unreliable due to small sample sizes which did not allow implementation of an adequate matching procedure.

However, sensitivity analysis of the PSM estimates shows that the matching estimates do not appear to be particularly robust as they indicate high sensitivity to selection on unobservables. This supports qualitative evidence from this study and the literature (Ito, 2003; Fernando, 1997) of the presence of strong selection on unobservables. Sensitivity analysis suggests that the more likely true impact estimates would be significantly lower, and possibly not significantly different from zero, for all outcome variables across all sub-groups and data collection rounds because these results are not robust to unobservables.

Further questions are raised about the ability of these methods to control for unobservables with these data because the USAID panel data set is not a 'true' panel. It does not allow a before and after comparison. What is being compared is the change in the outcome variable between a group that was already a member of SEWA Bank in Round 1 and a control group surveyed at the same time, with both groups at a later date. While compared to a proper before and after comparison this may underestimate the total impact assuming the two groups are indeed comparable. At the same time, it reduces the possibility of controlling for unobservables because any differences between the participants and controls in the absence (before) SEWA Bank cannot be empirically observed in these data. It cannot be shown that the treatment group before treatment was indistinguishable in terms of outcome variables, or, of course, unobservables, from the control group because there are no data from before treatment. Further doubts are raised by the way in which treatment and controls were sampled. This failed to explicitly rule out bias because the method of sampling of controls is not reported sufficiently. Indeed, the description of the procedure lays open a strong possibility that control households may have less ability to benefit from SEWA Bank than participating households because they had that opportunity but either chose not to participate or were selected out by self, peers or SEWA Bank staff.

The collection of the new cross-section data has been of limited help since their results were rather inconclusive and yielded little explanatory power mainly due to the shortcomings of the data pointed out earlier in this chapter. However, the effort to collect a new wave of data was highly instructive, including giving insights into current selection processes, which were likely to have been operative to some degree in the past.

Based on the findings in this chapter, it can be argued that a selection or screening process could be at work which is driven by the unobservables, e.g. entrepreneurial drive, business skills, possibly social capital, which together affect microfinance participation and cross section and panel differences between the treatment and control groups. The qualitative results presented here indicate a strong presence of social capital influencing (as outlined in Box 1) participation but the quantitative results cannot confirm this view due to a lack of adequate data. This leads to the question of how well social capital can be measured in the first place; this is a topic that the World Bank extensively dealt with from the mid 1990s onwards and measurement tools such as SOCAT, which was included in the questionnaire of the new cross-section data (discussed in section 4.5.), were developed (Grootaert and Bastelaer, 2002). Those tools, however, are mostly too general to yield any useful data, as found in this study. As discussed earlier, this is perhaps not very surprising since the concept of social capital is rather fuzzy after all (Harriss and de Renzio, 1997; Molyneux, 2002), and hence difficult to measure.

The debate on the appropriateness of the evaluation methods currently used to account for selection bias is far from over; but it is clear there is no miracle cure. The discussion in this chapter demonstrates that the evaluation techniques currently available have drawbacks in one way or another. PSM is not the wondrous tool as advocated by many and the impact estimates presented in this chapter should be taken with appropriate qualifications. There is qualitative evidence that there are strong unobservable effects, and that the unobservables have not been accounted for by any of the econometric techniques employed. Thus, controlling for biases due to unobservable characteristics remains a major challenge and a clear-cut solution to this issue has not yet been found. One point is clear, however, it is recommended to complement strictly quantitative approaches with qualitative ones.

Finally, not only do these data and methods not provide support for the idea that microfinance is highly beneficial to the poor, rather than perhaps benefitting a slightly better off group, but it leaves open whether microfinance is of any real benefit at all, since much of the apparent difference between microfinance participants and controls is likely due to differences in their characteristics rather than the intervention per se, not withstanding "inspiring stories" (Armendáriz de Aghion and Morduch, 2005, p. 199). This raises the question of under what circumstances, and for whom microfinance has been, and could be of real rather than imagined benefit to the poor.

This is further taken up in the next chapter with the most prominent data set, i.e. the study conducted by PnK which is based on strong design and methods but not free from debate. The re-investigation of PnK provides evidence that reduces the credibility of the quantitative support for microfinance and for lending to women in general. Furthermore, qualitative evidence (Fernando, 1997) strongly suggests other less beneficent interpretations leading to an unraveling of the microfinance narrative.

5. High noon for microcredit impact evaluations: re-investigating the evidence from Bangladesh[50]

5.1. Introduction

In this chapter I critically examine the most substantive evidence (RnM) on the impact of microfinance on poverty, and on the superior performance and effects of women as borrowers, provided by Mark Pitt and Shahidur Khandker. I first reprise the debate introduced in chapters 1 and 3 on the value of RCTs versus traditional observational methods. I showed that RnM, Deaton (2009), Imbens (2009) and others argue that the present drive towards RCTs also suggests renewed calls for taking a closer look at the value of observational studies which collect data through non-random processes. As outlined in chapter 3, observational studies are not uncontested either; there are threats to both internal and external validity that arise in observational data as well (Shadish, Cook and Campbell, 2002). Furthermore, observational data typically require the application of more complex econometric techniques, i.e. PSM, IV and DID estimations. However, these econometric techniques cannot usually deal adequately with selection bias due to unobservable characteristics. Given this context, this chapter looks at the evidence provided by one of the most authoritative microfinance impact evaluations conducted by PnK in Bangladesh which suggests that microcredit has a positive impact, in particular when women are involved. A number of studies, e.g. Morduch (1998, henceforth Morduch), RnM have made an attempt to replicate the findings of the original PnK study, and Chemin (2008, henceforth Chemin) has applied PSM, but with rather contradictory results. Morduch found hardly any impact, Pitt (1999, henceforth Pitt) defended the original claims, but Chemin and RnM found rather negligible impacts of microcredit.

The aim of this chapter is to re-investigate these studies by carefully re-constructing their analyses and to point to unresolved issues such as the claim that microcredit is more beneficial when targeted on women than on men, and on variables overlooked in

[50] I wish to thank Richard Palmer-Jones for his encouragement, persistence and invaluable help during the data set re-construction in which he invested a large amount of time. Many thanks also to David Roodman for supporting the data set re-construction and to Matthieu Chemin for sharing some of his STATA do-files to allow me to comprehend his data analysis. Many thanks also to the World Bank for providing additional data files that were not available online but essential for my analysis.

the data including borrowings from non-microcredit sources. The objective is to contribute to the methodological debate on the use of quantitative techniques and investigate their ability to account for selection bias due to unobservable characteristics.

5.2. The debate

As mentioned earlier, PnK's results have been challenged and their substantive claims that microcredit has positive impacts in particular when women are involved in borrowing have been subjected to criticism. Understanding this dispute is important for comprehending the problems of controlling for selection bias and the role of the unobservables in evaluations based on observational data. This section begins by briefly outlining the studies (PnK; Morduch; Pitt; Khandker, 2005; Chemin; RnM) involved in this debate before moving on to discussing the challenges of replicating and evaluating them.

The study conducted by PnK uses cross-sectional data from a World Bank funded study which conducted a survey in 1991-1992 on three leading microfinance group-lending programmes in Bangladesh, namely GB, BRAC and BRDB (PnK, p. 959). According to Morduch, at the time these three programmes catered to more than four million microfinance clients in Bangladesh (p. 2). A quasi-experimental design was used which sampled target (having a choice to participate/eligible) and non-target households (having no choice to participate/not eligible) from villages with microfinance programme (treatment villages) and non-programme villages (control villages).

The survey was conducted in 87 villages in rural Bangladesh; 1,798 households were selected out of which 1,538 were target households (eligible[51]) and 260 were non-target households (not eligible). According to PnK, out of those 1,538 households, 905 effectively participated in microfinance (59%). Data were collected three times in the 1991-1992 period in order to account for seasonal variations, i.e. various rice harvest seasons exist, namely Aman (November - February) which is the peak season, Boro (March - June) and Aus (July - October) which is the lean season (Khandker, 2005 –

[51] Eligibility criteria are subject to debate as discussed in depth in section 5.3.1. PnK deem any household with landholdings of less than 0.5 acres eligible.

henceforth Khandker, p. 271). The study focuses on measuring the impact of microfinance participation by gender on indicators such as labour supply, school enrolment, expenditure per capita and non-land assets. PnK find that microcredit has significant positive impacts on many of those indicators and find larger positive impacts when women are involved in borrowing.

Further to this, Khandker investigated the long-term impact of microcredit and re-surveyed the same households as in the original PnK study in 1998-1999. In addition, the follow-up survey

> "also added new households from the original villages, new villages in the original thanas, and three new thanas, raising the number of sample households to 2,599" (Khandker, p. 271).

Khandker argues that cross-sectional data only allows the measurement of short-term impacts of microcredit and that this is short-lived. Hence, he further argues that a panel data set is needed to gauge long-term impacts of microcredit programmes, because it allows control of unobservables (as discussed in chapter 3). Based on the panel data analysis Khandker finds that microfinance benefits the poorest and has sustainable impacts on poverty reduction among programme participants. In addition, positive spill-over effects are observed such as a reduction in poverty at the village level.

The debate, which is the central topic of this chapter, is focused on the cross-section data set (henceforth called R1-3) that was collected in 1991-1992 and which has been re-examined by Morduch, Pitt, Chemin and RnM – as mentioned above[52]. Morduch re-analysed R1-3 applying naïve methods. After a decade, RnM then re-visited both Morduch and the PnK cross-section analysis, and were the first to re-visit the panel data set and replicate Khandker.

The debate about the PnK findings started with Morduch who re-examined PnK and applied a simple DID approach to estimate the impact of microcredit. Morduch focuses on the problems of adequately enforcing the eligibility criterion of landownership,

[52] The data were also used by Khandker (1996, 2000); Pitt et al (1999); Pitt (2000); McKernan (2002); Pitt and Khandker (2002); Pitt et al (2003); Menon (2006); and Pitt, Khandker and Cartwright (2006).

which is discussed in more detail in section 5.3.1. In summary, Morduch reaches the conclusion that microcredit impact is overestimated due to the fact that the eligibility criterion was not always strictly enforced. Morduch finds that microcredit access leads to a decrease in labour supply and consumption variation across R1-3. It appears that households participating in microfinance have the ability to smooth consumption more so than their counterparts in non-programme villages and hence succeed in decreasing their risk and vulnerability. Morduch further argues that there is no evidence, however, that microfinance increases actual per capita expenditure or educational enrolment which is in contrast to the findings of PnK. Overall, Morduch finds that the impact of all three microcredit programmes is small or non-existent.

The concerns Morduch raised were reviewed by Pitt in a study which re-examined the original analysis of PnK taking Morduch's concerns into account. Pitt's study confirmed the findings of the original study. Pitt concluded that Morduch misinterpreted PnK's study, applied inappropriate methods, and thus found that programme effects were non-existent or very small.

There was no further response to Pitt's paper until recently when RnM re-visited the debate and replicated the studies of PnK, Morduch and Khandker and at the same time indirectly refuted Pitt's claims. RnM applied an advanced econometric package called cmp[53] developed by Roodman (2009) to replicate all of the PnK related studies. RnM find that their replication exercise does not provide any strong evidence that microfinance is indeed an effective strategy to alleviate poverty, especially when women are involved and it remains to be seen whether the poorest of the poor benefit as argued in the literature. Moreover, they refer to the current methodological debate centred around the use of RCTs versus observational designs and argue that in the absence of any high quality data, obtaining accurate microfinance impact estimates remains a challenge. In addition, challenges with regard to re-constructing the original

[53] cmp developed by Roodman (2009) is a STATA routine that contains a range of STATA commands that calculate recursive mixed process estimators, i.e. multi-equation models with different types of dependent variables which can also appear on the right hand side of any of the other equations in the model. RnM's cmp results were replicated with my re-constructed data set; I could approximate their results reasonably well despite minor differences in the calculation of some of the underlying variables.

PnK data set collected by the World Bank make it particularly difficult to exactly replicate PnK's findings – more on this in section 5.3.

Somewhat related to this debate is the study by Chemin which applies PSM to the PnK data. Chemin's impact estimates for all outcome variables except male labour supply are lower than the ones reported in PnK but higher than Morduch's results. Chemin does not report the impact of microcredit by gender, which would have been helpful since PnK stress that microcredit has a more positive impact when women are involved in borrowing than men.

This outline briefly introduced the main players involved in the debate, their methodological approaches and findings. None of the authors that re-visited the original PnK study could replicate and confirm the original findings (Morduch, Chemin and RnM). This gives rise to concern. With access to the original data and adequate documentation, it should be possible to fully replicate the findings of the original study. The fact that this is not the case might hold lessons for the general conduct of research, i.e. the need for full disclosure of original data as well as thorough documentation of data analysis procedures, and the proper application of quantitative techniques. The next section begins by briefly discussing the merits of replicating existing studies and then focuses on the unresolved issues that have emerged during the PnK replication exercise.

5.3. Replication and other challenges

Before discussing the main themes of the debate and suggesting ways forward, the rationale for replicating the existing PnK studies is outlined together with the challenges that occurred during this process. This replication exercise had the objective to comprehend the intricacies and limitations of the various studies as well as clarify the various methodological approaches.

The objective of replicating an author's work is to understand how the original results were derived and verify its findings (Hamermesh, 2007). A pure replication, according to Hamermesh (2007, p. 716), involves the availability of the original data in order to be able to re-investigate the original research questions and to apply the same models as in the original study. The objective is to allow other researchers to assess the robustness of the findings. To allow for replication, a certain degree of good documentation is required, i.e. authors should be prepared to share data and details of their variable construction and analysis in the form of computer codes[54].

In the case of PnK, most of the data including questionnaires and codes are (at the time the replication was undertaken and at the time of writing this chapter) available on the World Bank website[55] but certain data necessary for replication were (and are) missing, including consumer price indices, sampling weights and landholding details. Some of these data could be obtained after contacting the authors. PnK did not share the computer codes of their original data analysis as they can apparently no longer be recovered, though Pitt made some of his computer codes available when he responded to Morduch allowing to re-run his simulations (see STATA do-files available at http://www.pstc.brown.edu/~mp/sim.do). However, Pitt's code was of marginal interest to my replication efforts since it was exclusively written as a response to Morduch's claims and unrelated to PnK's original data analysis.

The replication exercise reported here was greatly facilitated by RnM who acted as a model of transparency and documentation by making all the data and codes available online[56] for the purpose of encouraging other researchers to re-visit their results and form their own opinion. RnM provide a Microsoft SQL Server database to manage the data files, correct data errors, and compile variables in data files which are imported into STATA. Subsequent data manipulation and analysis is in STATA making use of two .ado files written by Roodman (xtabond2 and cmp; see Roodman 2006 and 2009). My replication is entirely in STATA and hence does not require access to SQL.

Replicating the results presented in the various studies dealing with the PnK data has been a major challenge (this is also acknowledged by RnM) mainly due to gaps in the documentation and the general complexity of the study such as unwieldy questionnaires, imprecise definitions of variables, missing codes and inconsistencies and inaccuracies in data entry. To be fair, at the time the PnK study was conducted

[54] The American Economic Review (AER), for example, requires its authors to make their data sets available which are then uploaded onto a website maintained by the AER especially for this purpose. Authors have been compliant with this policy so far but can opt out in case their data are proprietary and/or confidential (Hamermesh, 2007, p. 717).

[55]http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:21470820~pagePK:64214825~piPK:64214943~theSitePK:469382,00.html

[56] http://www.cgdev.org/content/publications/detail/1422302

replication was not a norm nor was the publication of computer codes which may partly explain the lack of rigor with regard to documentation practices.

Since RnM replicated the key studies dealing with the PnK data (excluding Chemin), my replication exercise essentially does the same but in addition replicates Chemin's study as well. To begin with, the focus was on re-constructing the data set RnM used to replicate PnK, Morduch and Khandker in order to re-run their analysis to determine the robustness of their results. I first checked my data re-construction against RnM and managed to re-construct most of RnM's data set with the exception of a few variables related to landed assets, non-landed assets, labour supply and cumulative loans. Thanks to communication with Roodman, I resolved most of those discrepancies and made changes as I deemed necessary. My data set now approximates that of RnM apart from some minor differences where I deliberately chose a different interpretation, e.g. I included savings-in-kind when calculating non-landed asset variables and worked with slightly different assumptions when calculating landed asset variables. Nonetheless, re-running RnM's STATA do-files which replicate PnK, Morduch and Khandker using my data set closely approximates RnM's results and those of the other studies as already mentioned in footnote 53. In addition, I ran cmp on my re-constructed data set with model specifications that were different from the ones used by RnM; the results of this analysis are presented in section 5.4.3. which also explains the mechanisms of cmp in more detail.

This is the first time to my knowledge that Chemin's study has been replicated. Chemin does not engage in the PnK debate; he applies PSM to his variable constructions from the original data. Chemin's computer code is not publicly available, but in correspondence he sent an incomplete selection which was of limited assistance in re-creating the variables he uses. My replication of Chemin has displayed certain shortcomings which are discussed in section 5.3.3. and 5.3.4. Next, some unresolved issues related to the identification strategy and multiple sources of borrowing are dealt with before returning to Chemin's study and my suggestions for obtaining better impact estimates.

5.3.1. Identification strategy - eligibility criterion

PnK estimate the impact of microfinance participation on a set of outcome variables which involves comparing treated and non-treated households in treatment villages to non-treated households in control villages. However, as discussed in earlier chapters, non-random programme placement will hamper this comparison. Applying village-level fixed-effects may provide a solution to this problem by controlling for unobservable differences between treatment and control villages. However, applying village-level fixed effects to deal with the confounding effects of village placement are not without problems; control villages might have been affected by spill-over effects which adversely affect the accuracy of impact estimates (Ravallion, 2008). Furthermore, and unnoticed in other reports of these data, households in control villages had access to other formal or informal loan sources (access to and utilisation of multiple sources of borrowing is a microfinance reality as mentioned in chapter 2) provided by lending institutions other than GB, BRAC or BRDB and/or may have altered their behaviour in other ways due to these spill-over effects. As a result of access to other loan sources and spill-over effects, control villages would have been contaminated and are hence not unproblematic as control villages (chapter 3 discussed the issue of contaminated control groups). In fact, spill-over effects are very likely in this case considering the typical village set-up in rural Bangladesh and its networks. Bangladeshi villages are organised by residence and kinship groups (Hartmann and Boyce, 1983). Kinship practices can be described in a simplified way using the concepts of *gushti* and *bangsha*. According to Mannan (2002), *bangsha* refers to a form of lineage which may have different social origins, i.e. ethnic, occupational or religious without having a clear genealogical linkage. *Gushti* is a sub-segment of *bangsha* and refers to ancestors of the near past that are no more than three generations away and live in a single homestead (Mannan, 2002). There is considerable intermarrying among those kinship groups across villages suggesting substantial cross-village communication and highly developed trade networks across villages. As a consequence, considerable exchange of information across villages can be expected, contrary to common opinion, villages in Bangladesh are not (totally) isolated, and hence using village-level fixed-effects may not necessarily be useful in this case.

Moreover, comparing households in treatment and control villages is not sufficient for obtaining impact estimates for microfinance programme participation because households commonly self-select into microfinance, are selected by their peers and/or by microfinance loan officers giving rise to selection bias. Hence, using PnK's terminology, a comparison of target (eligible) versus non-target (not eligible) households is required. In order to place households in either target (eligible) or non-target (not eligible) groups, PnK use a discontinuity rule (Chemin) as a basis for their identification strategy. The discontinuity or eligibility rule for placing households in target (eligible) or non-target (not eligible) groups in this study is landownership which is assumed to be exogenous. In their original study PnK (p. 971) imply that households owning more than 0.5 acres of land will be excluded from joining any of the three microfinance programmes under investigation. Households owning less than 0.5 acres of land are on the other hand eligible to join. The eligibility rule, however, was not always strictly enforced. According to Morduch, PnK use membership as a sign of eligibility for participating households but the landownership criteria of 0.5 acres for non-participating households in treatment and control villages, which leads to inconsistencies (this is discussed in more detail later in this section).

In their theoretical exposition, PnK consider two villages, i.e. village 1 represents a control village and village 2 represents a treatment village; both villages contain landed and landless households. The authors express their identification strategy by the following equations (p. 968); *i* stands for household in village *j*:

$$(29) \qquad y_{ij} = \mathrm{C}_{ij}\delta + \mathbf{X}_{ij}\boldsymbol{\beta}_y + \mu_j^y + \varepsilon_{ij}^y$$

<u>Where:</u>

y_{ij} = outcome between households with and without choice

C_{ij} = level of participation in microfinance

δ = effect of the credit programme

X_{ij} = landownership

β_y = parameter to be estimated

μ_j^y = unmeasured determinant of *Cij* fixed within a village

ε_{ij}^{y} = non-systematic error term

$$(30a) \qquad E(y_{ij} \mid j = 1,\ \mathbf{X}_{ij} = 0) = \mu_1^y$$

$$(30b) \qquad E(y_{ij} \mid j = 1,\ \mathbf{X}_{ij} = 1) = \boldsymbol{\beta}_y + \mu_1^y$$

$$(30c) \qquad E(y_{ij} \mid j = 2,\ \mathbf{X}_{ij} = 1) = \boldsymbol{\beta}_j + \mu_2^y$$

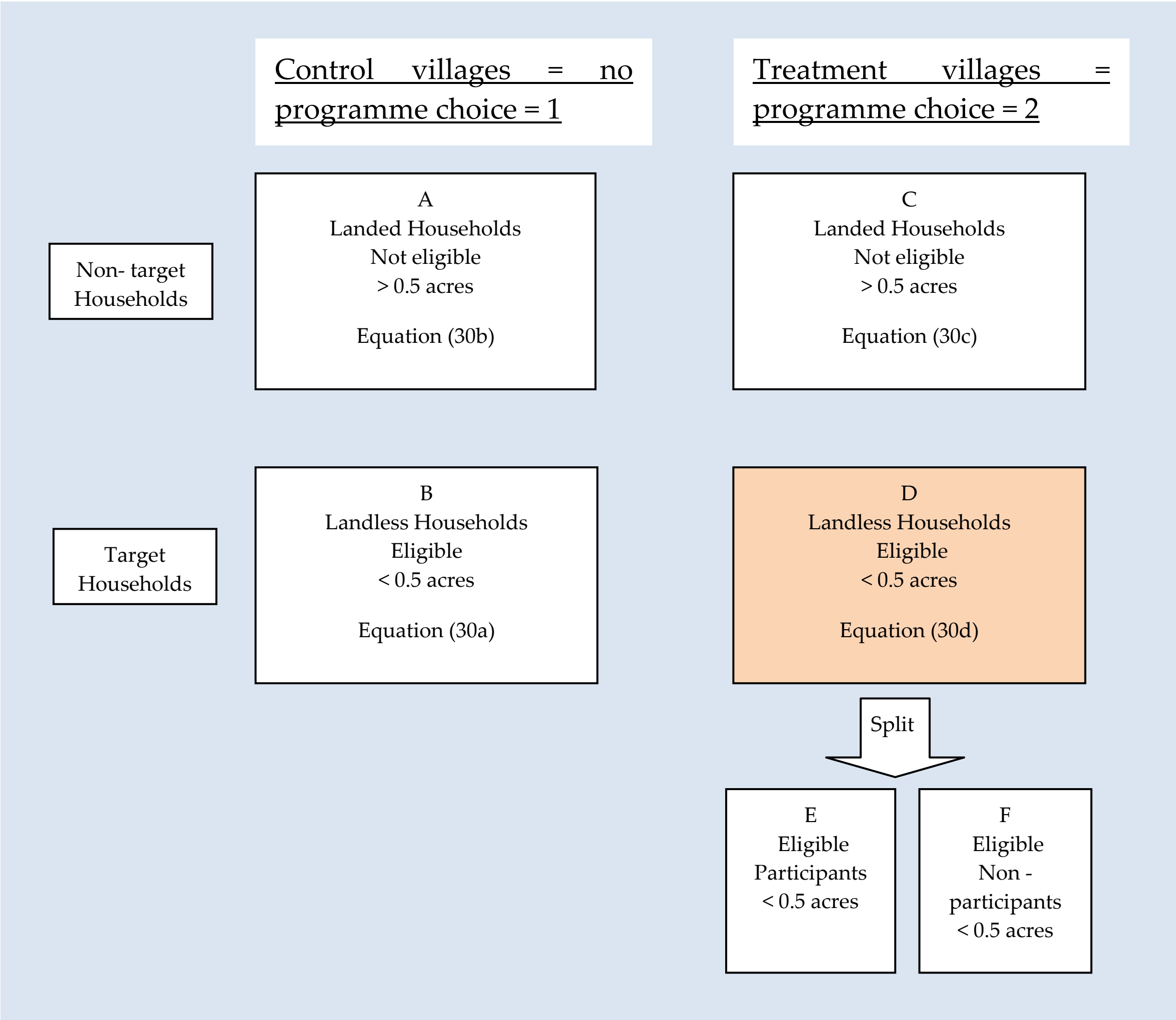
$$(30d) \qquad E(y_{ij} \mid j = 2,\ \mathbf{X}_{ij} = 0) = Þ\delta + \mu_2^y$$

Where:

Þ = proportion of landless household in treatment village that choose to participate

PnK’s identification strategy can be understood graphically by looking at Figure 6.

Figure 6: Identification strategy corresponding to equations (30a) to (30d)



Source: Author's illustration based on Morduch and Chemin.

As mentioned earlier, PnK suggest comparing target (eligible) versus non-target (not eligible) households, i.e. group B to A and group D to C in Figure 6. The difference of these two comparisons is compared between treatment and control villages applying village-level fixed-effects to account for unobserved differences between treatment and control villages. This is essentially a DID design, though PnK do not explicitly state that they are following such an approach. Ravallion (2008) supports this interpretation arguing that

> "…Pitt and Khandker (1998) do not mention the DD *[Author's note: double-difference]* interpretation of their design. However, it is readily verified that the impact estimator implied by solving Eqs. (4)(a)-(d) in their paper *[Author's note: equations 30a-30d here]* is the DD estimator…" (Footnote 41, p. 3818).

The application of an eligibility criterion as an identification strategy is a sensible approach provided it is implemented correctly and adhered to strictly. This, however, is not clear in the study conducted by PnK. Morduch claims that the PnK strategy did not succeed because the eligibility criterion they applied was not strictly enforced. Morduch further argues that the eligibility criterion was strictly observed in control villages (groups A and B), but not strictly observed in the treatment villages where membership was used. This mismatch creates a problem which hampers evaluation of the differences between villages and leads to misleading impact results. Ravallion (2008, p. 3818) and Chemin (p. 465) support Morduch's view and voice concern about the strict enforcement of the eligibility criterion. Morduch points out that PnK label any participating households in the programme villages (group D) as eligible, even households that should have been excluded according to the eligibility criterion. As a result of this, mistargeting occurred. Group D in fact contains participants which own more than 0.5 acres of land.

Chemin and Morduch argue that simply comparing groups E to F or groups E to B (see Figure 6) is misleading due to selection bias. As discussed earlier, participants commonly self-select into a microfinance programme, or are selected by their peers or by loan officers. Hence, comparing groups E to F or groups E to B introduces selection bias. As a result, Morduch proposes to compare the outcomes of groups E + F to those in group B which would provide bias-free impact estimates.

> "The focus is thus on measuring the impact of eligibility rather than participation…" (Morduch, p. 7).

Assuming that there are no spill-over effects between participants and non-participants, Morduch further states that by measuring the impact of eligibility, estimates of the average impact of participation can then be recovered

> "…by dividing the impact per eligible household by the proportion of eligible households that participate" (p. 7).

However, this comparison assumes that landholdings are exogenous, i.e. that membership in groups E, F or B is not influenced by self-selection (Morduch, p. 7). Furthermore, the comparison Morduch proposes does not "…reflect general differences across villages" (p. 8). Therefore, assuming that there are minor spill-over effects from group E to C or A, he suggests to employ a simple DID estimation that compares the outcomes of groups E + F to C. Similarly, he recommends conducting a comparison for group A relative to group B (Morduch, p. 8). Morduch further argues that

> "these within-village differences can then be compared across the two villages [*Author's note: treatment and control villages*], yielding a refined estimate of the average impact of eligibility" (1998, p. 8).

After employing those comparisons, Morduch finds no statistically significant impacts of exposure to microfinance.

Pitt disputes those findings and responds to Morduch's claims, in particular to those that maintain that mistargeting occurred. He reworks the original analysis taking Morduch's concerns into account and finds that the earlier findings of PnK were accurate. Pitt further argues that Morduch has inappropriately applied the eligibility rule, i.e. he has based its calculations on total land owned instead of cultivable land owned. One needs to note, however, that nowhere in the PnK oeuvre does one find an explicit statement as to which of the land variables in the data set is used to construct the eligibility criterion. Overall, Pitt states "that Morduch mischaracterizes the approach of PK [*Author's note: PnK here*]" (p. 8) and that the "eligibility rule [*Author's note: Morduch proposes*] is unjustified and will likely result in a biased estimate of program effects" (p. 11). Pitt further claims that the

> "failure to adequately deal with the mistargeting issue is not the only mistake to bedevil Morduch's new evidence. He fails to set out a clear framework justifying his difference-in-differences estimate…" (p. 11).

In his response, Pitt repeatedly criticises Morduch's findings and stresses that Morduch simply misinterpreted the PnK study and thus found that programme effects were non-existent or very small.

Table 13: Degree of mistargeting, total land owned (< 0.5 acres), prior to joining the microcredit programme

Participating households in treatment villages			
Programme	**Not eligible**	**Eligible**	**Total**
BRAC	60	225	285
BRDB	57	251	308
GB	84	228	312
Total	201	704	905

Source: Author's calculations based on PnK data across R1-3 downloaded from the World Bank website.

Table 14: Degree of mistargeting, cultivable land owned (< 0.5 acres), prior to joining the microcredit programme

Participating households in treatment villages			
Programme	**Not eligible**	**Eligible**	**Total**
BRAC	48	237	285
BRDB	48	260	308
GB	69	243	312
Total	165	740	905

Source: Author's calculations based on PnK data across R1-3 downloaded from the World Bank website.

Table 13 and Table 14 are my partial and simplified replication of Pitt's table 1 (p. 21) illustrating the degree of mistargeting among participating households by using different landownership variables. I could not exactly re-construct Pitt's table 1 but my figures are good approximations. Table 13 displays the results obtained when using total land owned as the main variable, while Table 14 uses cultivable land as the main variable. Both tables illustrate that there is substantial mistargeting among participating households, i.e. many households have more than 0.5 acres of land. As illustrated by Table 13, 22.2% of participating households are mistargeted according to a straight forward land area criterion; this confirms Morduch's statement that 20-30% (p. 12) of participating households are mistargeted. When applying cultivable land as the main variable (Table 14), still 18.2% of participating households are mistargeted which broadly matches McKernan's (2002) estimates. She argued that 17% of participating households were mistargeted, i.e. they owned more than 0.5 acres of cultivable land (McKernan, 2002, p. 102).

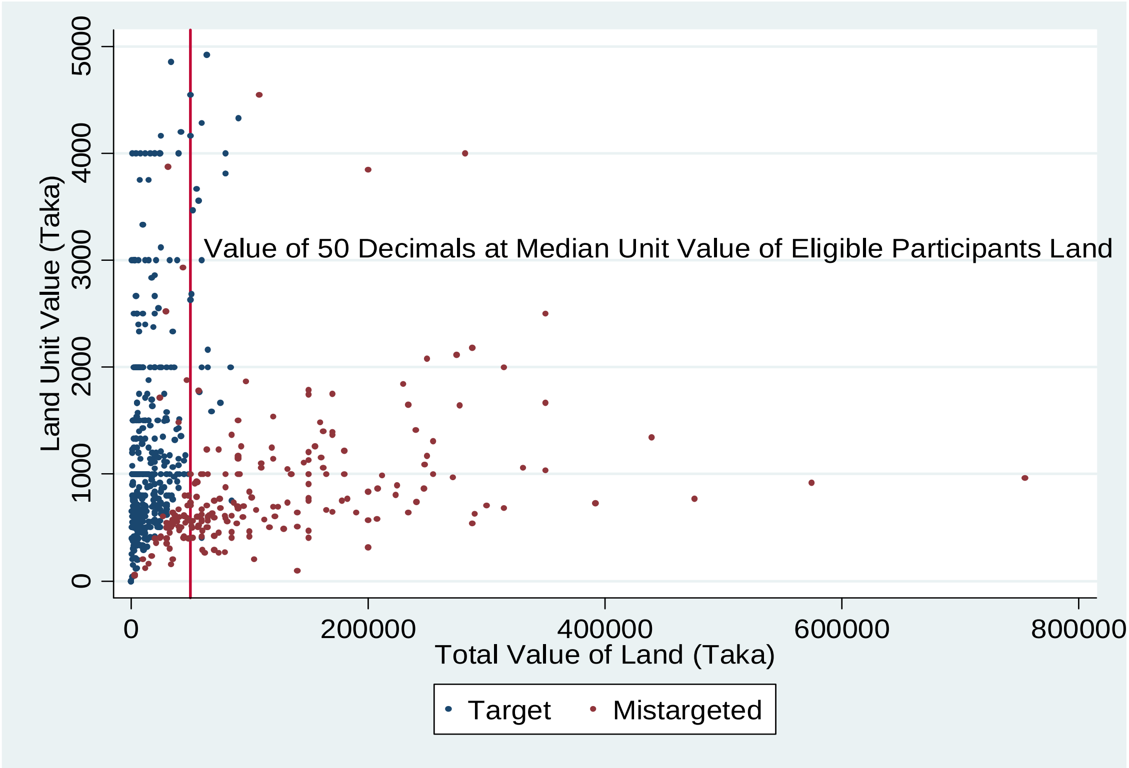
Pitt asserts that ownership of cultivable land is in fact rather ambiguous for the purpose of establishing eligibility because of variations in land productivity (and its value/price). He argues that the three microfinance programmes are

> "aware of the differential cultivability of plots of land, and might be expected to make adjustments for land quality in judging the eligibility of a household" (ibid, 1999, p.3).

If this is indeed true and the three microfinance programmes do take land quality into account when establishing programme eligibility, then the underlying assumption is that the mistargeted households that participate would have total land values that would be no more than the median unit value of land of the correctly identified households that participate (i.e. less than 0.5 acres).

Looking at the data provides the following insights: the median unit value of land of participating households having less than 0.5 acres of land (i.e. eligible participants) equals 1000 Taka per decimal (50 decimals equal 0.5 acres). Thus, one might suggest the cut-off point for establishing programme eligibility is 50,000 Taka, i.e. mistargeted households that participate should have land valued at less than 50,000 Taka following Pitt's argument. However, 50% of the mistargeted households that participate have total land values of greater or equal to 85,000 Taka, and 72% of those mistargeted households have total land values of greater or equal to 50,000 Taka. Hence, Pitt's argument does not convince (see Figure 7).

Figure 7: Land unit values by total land value and targeting



Source: Author's calculations based on PnK data R1 downloaded from the World Bank website.

In other words, if Pitt's approach is correct then the total value of land of mistargeted households should be no more than the value of 50 decimals of land. Figure 7 shows the unit values by the total value of land and a vertical line showing the approximate value of 50 decimals at the median value of participants' land (1000 Taka/decimal). While quite a number of mistargeted households have lower total values of land than this cut-off point, the majority have significantly larger total value of land than the upper limit, suggesting that they are indeed mistargeted.

As mentioned earlier, Pitt asserts that using land ownership as an eligibility criterion can be rather ambiguous. Either way, mistargeting appears to remain a significant problem and the overall question that arises from this discussion is whether the research design PnK applied in this context was indeed the right choice. The effectiveness of their design heavily relies on the strict enforcement and clear definition of an eligibility criterion which in this case could not be warranted.

PnK support the use of landownership as an eligibility criterion by arguing that the virtual absence of an active land market justifies its application (p. 970). Morduch, however, provides evidence to the contrary; he argues that there is substantial evidence on an active land market in South Asia (p. 4). He argues that close to one eighth of participants in fact bought substantial amounts of land a few years before the survey was conducted. Further to this, estimating actual landownership is rather difficult and a sensitive topic. Many land owners are worried about disclosing the true size of their landholdings for fear of losing them since they might have obtained their land by using fictitious documents or other means (Ahmed, 2002). Land disputes are a common occurrence in the country. Lewis (1991) argues that in principle landholdings should not have changed since Partition and independence due to the Bengal Tenancy Act of 1950 which recognised the rights of landlord and tenants. In reality, this Act was hardly enforced, to the disadvantage of the poorer strata of society (Lewis, 1991). Bangladesh's land market is characterised by scarcity of land and concentrated landholdings. However, rental practices are widespread and access to operational land is common (Lewis, 1991). Ideally, this access to operational land should in fact be taken into account when establishing eligibility for microcredit participation but this is difficult to estimate.

5.3.2. Multiple sources of borrowing

None of the studies mentioned here investigated the origin of the loan portfolio of the participating household, i.e. whether the main loan source of a household in a particular treatment village is from the microfinance organisation that dominates that particular village. Surely, for the purpose of the PnK study, a household that has a BRAC loan should reside in a BRAC village and should in fact borrow only from BRAC; similarly, if residing in a BRDB village, the households should only have a BRDB loan, and so on. The data provides evidence that BRAC villages also include households that borrow from BRDB or GB (see Table 15). This shows that the classification of participants and treatment villages is not without error. One would have expected a clear distinction of villages by microcredit programme.

Table 15: Degree of multiple programme memberships, household level data, n = 1,798

Programme	BRAC villages	BRDB villages	GB villages	Control villages	Total
BRAC households	285	0	0	0	285
BRDB households	5	303	0	0	308
GB households	10	3	299	0	312
Control/non-participating households	197	198	198	300	893
Total	497	504	497	300	1,798

Source: Author's calculations based on PnK data across R1-3 downloaded from the World Bank website.

Furthermore, some participating households obtained additional loans from borrowing sources other than microcredit, e.g. loans from formal sources such as government controlled banks like the Krishi Bank or from informal sources such as relatives, friends, landlords, and so on. Table 16 illustrates the extent of multiple borrowing among microcredit participants and non-participants across treatment and control villages.

Table 16: Multiple sources of borrowing among microfinance participants and non-participants, person level data

Programme	Microfinance participants	Non-participants	Total
BRAC	8	32	40
BRDB	30	24	54
GB	8	9	17
Eligible Non-member	0	192	192
Control households	0	114	114
Total	46	371	417

Source: Author's calculations based on PnK data across R1-3 downloaded from the World Bank website.

Table 16 shows that 46 microcredit participants have other sources of borrowing – formal as well as informal[57] – in addition to the microcredit loan they have taken on. Interestingly, the highest number of multiple borrowing can be found among BRDB

[57] Table 16 shows that 417 individuals had multiple sources of borrowing, out of which 107 had formal sources of borrowing, 284 had informal sources of borrowing and 26 had both.

members. It is not clear why this is the case. I am purely speculating but the size of the BRDB loans could possibly be too small driving clients towards finding other sources of borrowing in order to satisfy their financing needs. Furthermore, among the non-participants across treatment and control villages a substantial number of individuals have other formal or informal sources of credit which are not microcredit. In particular among the households in the control villages, 114 individuals have in fact some kind of borrowing which should disqualify them from being in the control group since they would contaminate it as discussed in section 5.3.1.

This leads us to the question of what we are trying to measure. Ultimately, all impact studies are trying to find out how the lives of the poor would have turned out if an intervention such as microfinance for example had not been introduced (Blundell and Costa Dias, 2008; Heckman and Vytlacil, 2007a). In other words, what would have happened to participating individuals had the microfinance programme not been available? This is the challenge of measuring the counterfactual, a process which commonly introduces biases that can adversely affect impact evaluation results, as shown in chapters 1 and 3. Counterfactuals cannot be observed, thus every programme evaluation can only make an attempt at creating an estimate of such a counterfactual. Those estimates are then used to pinpoint the effect of the programme (Bryson, Dorsett and Purdon, 2002). The challenge in the case of PnK is to single out the impact of microcredit, when there is substantial borrowing from a range of other non-microcredit sources across participants and non-participants in treatment and control villages. Assuming that in the absence of MFIs no other credit sources other than traditional informal sources would have been available makes it difficult to identify an appropriate control group that is not contaminated by borrowing from other sources. Consequently, is it really possible to measure the impact of participating in microcredit versus the hypothetical case of not participating? Or would it be more appropriate to measure the impact of microcredit participation versus participating in the next best alternative source of credit? Section 5.4. deals with this issue and suggests ways forward.

What have we learnt so far? The most authoritative microfinance impact evaluation (RnM) was not contested for more than 10 years mainly because of data complexities and a sophisticated econometric model that was challenging to replicate. Morduch, Pitt, Chemin and RnM have made an effort to re-visit the PnK study with the aim to replicate or defend (in the case of Pitt) its findings but with rather mixed results. The latest effort was made by RnM with the intention to clean up the debate but some unresolved issues remain. I could replicate RnM's analysis closely thanks to a transparent analytical process and their analysis convinces. However, RnM were only concerned with the replication, and did not explore the possibility of applying other econometric techniques such as PSM to which I now turn.

5.3.3. Chemin replication

Chemin re-investigates the PnK data set and in contrast to the studies by PnK, Morduch, Pitt and RnM, he employs a matching technique with the aim to compare the outcomes of microfinance participants in treatment villages to non-participants in treatment and control villages. Chemin bypasses the PnK debate arguing that the issue of eligibility can be avoided by PSM (p. 465). To recap, PSM is performed by matching participants to non-participants based on the predicted probability of programme participation or the "propensity score" (Ravallion, 2001) (as outlined in chapter 3 and 4). 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" (p. 465).

However, this chapter argues that it is not clear whether PSM can achieve all this, and whether, indeed, it is an appropriate technique for solving the particular problems in the PnK data set. Chemin rightly states that "matching can cover selection into the programme based on observables but not unobservables" (p. 465). As discussed in chapter 3, it appears that PSM is not the wondrous tool as advocated by many as it commonly fails to solve the evaluation problem (Smith and Todd, 2005). Bryson, Dorsett and Purdon (2002) support this view and emphasize that the quality of the underlying data is crucial to obtaining reliable impact estimates.

This chapter uses Chemin's study to demonstrate the challenges of replication as well as the drawbacks of PSM. I attempt to replicate Chemin's paper with the objective to understand the details of his study and verify his findings. Section 5.4. extends the

PSM approach to examine issues not dealt with by Chemin including investigating the impact of microcredit by gender and comparing various treatment groups.

Table 17 provides the details of Chemin's logit model specifications. He 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. According to Chemin, specification 1 replicates Pitt; specification 2 contains the same variables as specification 1 as well as additional control variables which Chemin argues are derived from economic theory and might be of use for predicting microfinance participation. Finally, specification 3 is the logistic regression model Chemin uses in his PSM.

Chemin's original results could not be fully replicated (as illustrated in Table 18) mainly because of challenges with regard to accurately re-constructing his data set which was not an easy task. First of all, he does not state the origin of the data set he used. I simply assume that he downloaded the data from the same World Bank website[58] I also accessed and that we both downloaded the same data files. I further assume that he did not obtain the additional data files related to consumer price indices, sampling weights and landholdings (as mentioned in section 5.3.) which were not publicly available on the World Bank website but which I had acquired by contacting the World Bank directly. Those data files are important for the analysis and not including them can partially explain some of the discrepancies between Chemin's analysis and my replication[59].

Specifically, only some of the mean values of many of Chemin's variables could be replicated. For example, my values for age, number of adult males in the household, education and savings are quite different to Chemin's results (see Table 17, column 2 for Chemin's original analysis and Table 18, column 2 for my replication). Appendix 4 provides the descriptive statistics reported by PnK and RnM for R1 to allow a

[58]http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:21470820~pagePK:64214825~piPK:64214943~theSitePK:469382,00.html – same web link as presented in footnote 55.

[59] All PnK related data files are now publicly available on David Roodman's website (http://www.cgdev.org/content/publications/detail/1422302) but they were uploaded after Chemin had published his study – same web link as presented in footnote 56.

comparison. The mean values reported by Chemin and by me are averages across R1-3 which can partially explain some of the differences to PnK and RnM but not all. The main reasons for those discrepancies are differences in the calculations of the underlying variables. I contacted Chemin and obtained some of his STATA do-files to understand his re-construction of PnK's data set. However, his STATA do-files were incomplete and did not provide a full account of his analysis. It was not always clear which variables he used for his calculations, steps in the analysis were missing and hence exact replication of his original data set was not possible due to lack of information and computer code.

The discrepancies in the mean values follow through to the logit coefficients presented in Table 17 and Table 18. As displayed in Table 18, the logit coefficients for sex, age and age of household head in specification 1 are similar to Chemin's results in Table 17 in terms of size and significance. The remaining logit coefficients, however, differ. A similar pattern can be found in specification 2 which includes 23 additional control variables. According to Chemin those control variables were all insignificant (Chemin, p. 471). However, I found that 5 out of those insignificant control variables were in fact significant – Table 18 provides the details. Moreover, the pseudo R-squared in my replication across all logit specifications is clearly lower than the figures provided by Chemin; this could be due to problems of exactly re-constructing Chemin's original data set and the remaining underlying differences of the data sets used by Chemin and by me. The number of observations differs as well.

Table 17: Chemin's logit specifications predicting the probability of microfinance participation

Independent variables	Means	Spec. 1	Spec. 2	Spec. 3
Highest Grade completed	2.255 *3.173*	0.041 *0.03*	0.024 *0.036*	
Sex (male=1)	0.513 *0.499*	-0.886*** *0.123*	-1.515*** *0.182*	-1.136*** *0.128*
Age (years)	22.327 *17.422*	0.051*** *0.004*	1.224*** *0.269*	1.065*** *0.159*
Age household head (years)	42.313 *12.383*	-0.046*** *0.006*	-0.035*** *0.009*	-0.014** *0.006*
Number adult male in household	0.024 *0.153*	1.951 *1.268*	2.854* *1.562*	0.832*** *0.308*
Landholdings HH head parents	0.246 *0.56*	0.137 *0.14*	0.094 *0.147*	
Landholdings HH head brothers	0.714 *1.224*	0.019 *0.065*	-0.023 *0.068*	
Education	0.551 *0.497*			0.336*** *0.113*
Savings	1128.9 *4201.37*		0.0002*** *0.0004*	0.0002*** *0.00003*
Have non-farm enterprise (yes=1)	0.468 *0.499*		0.763*** *0.173*	0.630*** *0.111*
Livestock value	3273.15 *5533.9*		0.0000397 *0.00003*	0.00005*** *0.00002*
Household size	6.232 *2.632*		-0.117*** *0.041*	-0.147*** *0.028*
Non-agricultural wage (in Taka)	4.023 *16.303*		-0.002 *0.004*	-0.006* *0.003*
Agricultural wage (in Taka)	2.987 *9.755*		0.013** *0.007*	0.010** *0.005*
Age squared	802 *1109.7*		-0.033*** *0.01*	-0.028*** *0.006*
Age power of 4	1874542 *5029988*		-1.73E-6* *0.000000944*	-1.16E-6*** *0.000000501*
Village dummies	Yes	Yes	Yes	Yes
Number of observations		4215	4205	5037
Pseudo R-squared		0.1502	0.3561	0.3313

Source: Chemin, table 1, p. 471.

Notes: p-values in italics. * significant at 10%, ** significant at 5%, *** significant at 1%. PnK data across R1-3 downloaded from the World Bank website are used. According to Chemin, specification 1 replicates Pitt, specification 2 includes other control variables such as landed assets, equipment assets, transport assets, injuries, change of residence in the last 2 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.

Table 18: Replication of Chemin's logit specifications predicting the probability of microfinance participation

Independent variables	Means	Spec. 1	Spec. 2	Spec. 3
Highest grade completed	1.824 *3.104*	-0.007 *-0.589*	0.017 *-0.377*	
Sex (male=1)	0.513 *0.500*	-0.561*** *0.000*	-0.760*** *0.000*	-0.698*** *0.000*
Age (years)	33.464 *14.549*	0.011*** *0.000*	0.463*** *0.000*	0.492*** *0.000*
Age household head (years)	43.028 *13.311*	-0.011*** *-0.003*	-0.001 *-0.878*	-0.001 *-0.897*
Number adult male in household	1.660 *1.117*	-0.390*** *0.000*	-0.034 *-0.646*	-0.103 *-0.129*
Landholdings HH head parents	0.176 *0.381*	-0.157 *-0.174*	-0.008 *-0.946*	
Landholdings HH head brothers	0.361 *0.480*	-0.218** *-0.017*	-0.026 *-0.795*	
Education	4.006 *3.953*			-0.041*** *-0.001*
Savings	384.491 *1525.339*		0.000*** *-0.003*	0.000 *-0.084*
Have non-farm enterprise (yes=1)	0.446 *0.497*		0.413*** *0.000*	0.476*** *0.000*
Livestock value	3753.324 *6044.382*		0.000 *-0.710*	0.000 *-0.241*
Household size	6.031 *2.772*		-0.051* *-0.074*	-0.109*** *0.000*
Non-agricultural wage (in Taka)	6.634 *10.496*		0.000 *-0.984*	0.001 *-0.743*
Agricultural wage (in Taka)	4.520 *6.220*		-0.005 *-0.491*	0.001 *-0.881*
Age squared	1331.469 *1191.801*		-0.007*** *0.000*	-0.008*** *0.000*
Age power of 4	3193038 *6159567*		0.000*** *0.000*	0.000*** *0.000*
Number adult female in household	1.452 *0.832*		-0.243*** *-0.003*	
Agricultural income (in Taka)	31.101 *4772.773*		0.000** *-0.010*	
Household land (in decimals)	58.269 *104.119*		-0.004*** *-0.004*	
Marital status (married=1)	0.699 *0.459*		0.316** *-0.011*	
Other assets (in Taka)	6368.578 *15663.870*		-0.000*** *0.000*	
Village dummies		Yes	Yes	Yes
Number of observations		5357	5357	5450
Pseudo R-squared		0.100	0.191	0.172

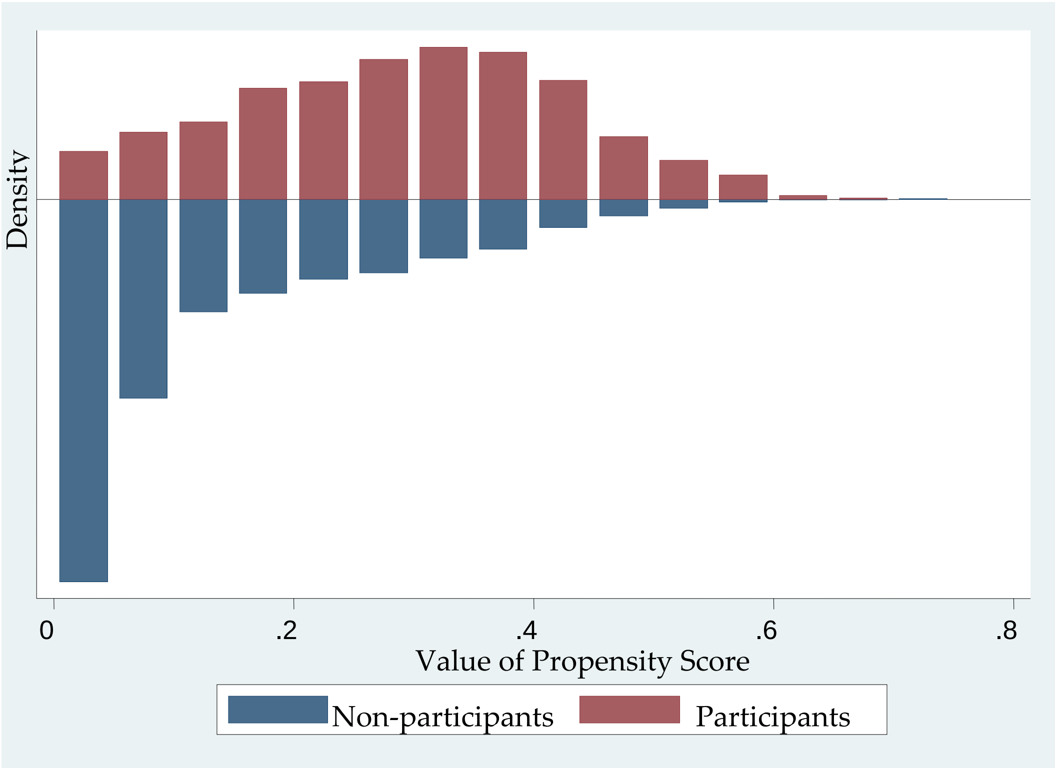
Source: Author's calculations.

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

To summarise the argument so far, Chemin's results could not be fully replicated probably due to differences in the calculation of the underlying variables. His STATA do-files were incomplete and did not allow me to fully understand the steps in his data analysis. This speaks to the desirability of requiring authors to deposit computer code and data along with their published papers, as is now more common practice (Hamermesh, 2007).

In his paper, Chemin first presents the results that compare participants in treatment villages with non-participants in treatment villages as well as participants in treatment villages with individuals in control villages by examining log of per capita expenditure and hence I begin by examining those results (see Table 19 which presents Chemin's original PSM results as well as my replication results). However, before discussing the PSM results in more detail, I first investigate the distribution of propensity scores for participants and non-participants across treatment and control villages with the objective to examine whether comparing those groups is sensible (as described in chapter 4).

Figure 8: Distribution of propensity scores for participants and non-participants in treatment and control villages



Source: Author's calculations.

Figure 8 presents the distribution of propensity scores and indicates 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 matches between treatment and control groups are available. Not surprisingly, slightly more overlap can be observed when the same figure is plotted for participants and non-participants in treatment villages only. Generally speaking, the lack of overlap implies that the common support assumption is not fully satisfied, and subsequently the question must be asked whether PSM is a suitable technique in the specific context of PnK given those facts. The robustness of my matching estimates is further investigated through sensitivity analysis in section 5.3.5.

Chemin's PSM results indicate that microcredit has a significant negative impact on participants' log of per capita expenditure when compared to non-participants in treatment villages (Table 19, row 3). Participants spend 3.5% to 4.6% less per capita than non-participants. This is surprising and contrary to the expectation that microcredit participation has positive impacts. It appears that matched non-participants in treatment villages are not necessarily worse off than participants in treatment villages. An explanation could be that the matched non-participants have access to and make use of other sources of borrowing which has a positive impact on their log of per capita expenditure. The replication results presented in Table 19, row 6 confirm the pattern of Chemin's results and provide negative impact estimates across both matching algorithms. According to my replication results, the impact estimates are less dramatic than the ones calculated by Chemin; participants spend 0.7% to 2.3% less than non-participants. However, when stratification matching with 10 strata is applied, then participants appear to spend 0.3% more than non-participants – significantly so. This is puzzling and cannot be explained. There are also discrepancies in the kernel matching results, Table 19, row 6 provides negative and insignificant estimates while Chemin's kernel estimates in Table 19, row 3 are all significant.

Table 19: Chemin's impact estimates and their replication for log of per capita expenditure (Taka)

Control group	Stratification 20	Stratification 10	Stratification 5	Bandwidth of kernel matching 0.05	Bandwidth of kernel matching 0.02	Bandwidth of kernel matching 0.01
Chemin's reported results[1, 2]						
Non-participants in treatment villages	-0.035*	-0.044*	-0.044*	-0.039*	-0.044*	-0.046*
Individuals in control villages	0.028	0.028***	0.028*	0.028***	0.028***	0.028***
Replication of Chemin[3]						
Non-participants in treatment villages	-0.007***	0.003***	-0.006***	-0.022	-0.021	-0.023
Individuals in control villages	0.003***	-0.009***	-0.002	-0.078***	-0.081***	-0.081***

Notes: * significant at 10%, ** significant at 5%, *** significant at 1%. 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. As in chapter 4, 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.
1. Source: Chemin, table 2, p. 476.
2. 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.
3. 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. Source: Author's calculations.

In Table 19, row 4 participants in treatment villages are compared with individuals in control villages. Chemin finds that microcredit has a significantly positive impact on participants in treatment villages across all matching algorithms. Microfinance participation increases the log of per capita expenditure and participants spend 2.8% more than individuals in control villages. The replication results presented in Table 19, row 7 differ substantially from Chemin's. I find that participants are worse off than individuals in control villages and in fact spend 0.9% to 8.1% less than individuals in control villages – significantly so. However, stratification matching (as explained in chapter 3) using 20 strata provides a significant and positive result, which is unexpected. My replication results suggest that participants are spending less than individuals in control villages with even lower impact estimates when compared to individuals in treatment villages. Table 19 presented impact estimates for two different comparisons but for only one outcome variable. Table 20 and Table 21 present the impact estimates for all six outcome variables comparing microfinance participants in treatment villages with non-participants across treatment and control villages. Table 20 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.

Table 20: Chemin's impact estimates for all 6 outcome variables matching participants to non-participants across treatment and control villages.

Outcome variables	**Bandwidth of kernel matching**		
	0.05	**0.02**	**0.01**
Variation of log per capita expenditure (Taka)	-0.008	-0.008	-0.008
Log women non-landed assets (Taka)	0.037	0.037	0.038
Female labour supply, aged 16-59 years, hours per month	9.503	9.507	9.521
Male labour supply, aged 16-59 years, hours per month	17.001***	16.996***	16.974***
Girl school enrolment, aged 5-17 years	0.051***	0.051***	0.052***
Boy school enrolment, aged 5-17 years	0.035*	0.035	0.036

Source: Chemin, table 3, p. 477.
Notes: * significant at 10%, ** significant at 5%, *** significant at 1%. 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.

My replication results presented in Table 21 cannot confirm Chemin's results and instead suggest that microcredit has a significant and positive impact on the log of women's non-landed assets and boys' school enrolment with all other results being insignificant.

Table 21: Replication of Chemin's impact estimates for all 6 outcome variables plus log per capita expenditure matching participants to non-participants across treatment and control villages.

Outcome variables	Bandwidth of kernel matching		
	0.05	0.02	0.01
Variation of log per capita expenditure (Taka)	-0.001	-0.002	-0.002
Log per capita expenditure (Taka)	-0.023	-0.022	-0.023
Log women non-landed assets (Taka)	1.313***	1.304***	1.306***
Female labour supply, aged 16-59 years, hours per month	1.248	0.489	0.314
Male labour supply, aged 16-59 years, hours per month	-0.333	4.077	4.723
Girl school enrolment, aged 5-17 years	0.028	0.033	0.033
Boy school enrolment, aged 5-17 years	0.066**	0.075**	0.079**

Source: Author's calculations.

Notes: * significant at 10%, ** significant at 5%, *** significant at 1%. 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. As in chapter 4, 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.

To summarise the main points of the comparison between Chemin's findings and my replication results: Chemin finds no impact on the log of women's non-landed assets while my replication supports the view that microfinance has a significantly positive impact on the log of women's non-landed assets. In addition, Chemin argues that there is a positive and significant impact on male labour supply and on girls' school enrolment which cannot be supported by my analysis. Finally, I find that microfinance has a significant and positive impact on boys' school enrolment which Chemin's results do not agree with.

5.3.4. Limitations

As noted above, Chemin's study does not provided sufficient information on the calculation of the underlying variables he uses for his models and hence it is not surprising that his results could not be fully replicated. More details would be required in order to be able to re-construct his findings more accurately. Furthermore, Chemin does not assess the impact by gender separately which is a drawback since PnK emphasize that microcredit has more of a positive impact when women are involved in borrowing. In addition, my logit model with the specification 3 has a rather low pseudo R-squared which not only has implications for the explanatory power of the model and the quality of the matches, but also raises questions about the construction of variables in Chemin – without further details of these constructions the difference seems unresolvable.

The low explanatory power of my logit model explaining microfinance participation leads to the question of the usefulness of matching in this context. Does PSM provide convincing impact estimates that account for selection bias due to observable as well as unobservable characteristics? The next section examines these issues and assesses the sensitivity of the matching estimates I presented in Table 21.

5.3.5. Sensitivity analysis

The inability of PSM to control for unobservable characteristics, led Rosenbaum (2002) to develop a sensitivity analysis approach assessing the robustness of PSM results. Chapters 3 and 4 discussed and applied sensitivity analysis in depth and hence it is not further explained in this chapter but simply applied to the matching estimates I presented in Table 21.

As discussed in chapter 4, sensitivity analysis can be implemented using the rbounds procedure in STATA (Becker and Caliendo, 2007); this procedure uses the data to calculate the confidence intervals (for a given level of confidence – e.g. 95%) of the outcome variable for different values of Γ (gamma), a number which captures the required degree of association, of an unobserved characteristic with the treatment, for it (the unobserved characteristic) to explain the observed impact.

A value of Γ that produces a confidence interval that encompasses zero is one that would make the estimated impact not statistically significant at the relevant level of confidence. If Γ is relatively small (say < 2) then one may assert that the likelihood of such an unobserved characteristic is relatively high, because its effect does not need to be large to counteract the estimated treatment effect, and therefore that the estimated impact is rather sensitive to the existence of unobservables (DiPrete and Gangl, 2004) (as discussed in chapter 4).

For example, sensitivity analysis can be illustrated by calculating Γ at which the estimated impact of microfinance participation on boys' school enrolment across R1-3 is no longer statistically significant. Table 21 shows that the kernel matching estimate with a bandwidth of 0.05 for boys' school enrolment is 0.066 and statistically significant at 5%. This suggests that boys' school enrolment is significantly higher for participating households than for control households. However, this may not be due to membership *per se* but to unobserved characteristics that account for membership (and or its impact). Sensitivity analysis explores the robustness of this impact estimate and demonstrates how it changes in the presence of selection on unobservables. The STATA procedure rbounds reports the estimates[60] and their 95% (or other) confidence intervals for matched pairs of microfinance members and controls (see Table 22).

When $\Gamma = 1$ there is no selection on unobservables. If Γ increases to 1.2, then matched individuals differ in their odds of exposure to microfinance by a factor of 1.2 due to selection on unobservables. Table 22 shows that when $\Gamma = 1.2$, a relatively small difference in the odds of exposure, the statistical significance level ranges from < 0.0001 to < 0.1045. Hence, when $\Gamma = 1.2$ or more, the 95% confidence interval of the point estimate does straddle zero and this implies that for this variable selection on unobservables is likely to explain the observed association between exposure to microfinance and higher boys' school enrolment. Consequently, I argue that the observed impact of microfinance membership on boys' school enrolment is not significantly different from zero, because the odds of exposure to an unobservable that produces a confidence interval for the outcome variable that encompasses zero, is low, which leads to the conclusion that the association between microfinance exposure and higher boys' school enrolment may well be due to unobservables.

[60] 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 21 which provide the average treatment effects.

Table 22: Sensitivity analysis for boys' school enrolment for microfinance participants across R1-3

	Significance levels		Hodges-Lehmann point estimates		95% Confidence intervals	
Gamma (Γ)	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
1	< 0.0001	< 0.0001	0.022	0.022	0.017	0.025
1.2	< 0.0001	< 0.1045	0.015	0.026	-0.015	0.030
1.3	< 0.0001	< 0.4120	0.003	0.028	-0.039	0.335
1.4	< 0.0001	< 0.7687	-0.016	0.030	-0.055	0.037
1.5	< 0.0001	< 0.9480	-0.037	0.033	-0.066	0.041
1.6	< 0.0001	< 0.9931	-0.051	0.036	-0.133	0.044
1.7	< 0.0001	< 0.9994	-0.060	0.039	-0.451	0.047
1.8	< 0.0001	< 0.9999	-0.073	0.042	-0.455	0.053
1.9	< 0.0001	< 1	-0.437	0.044	-0.458	0.415
2.0	< 0.0001	< 1	-0.452	0.472	-0.461	0.440
2.1	< 0.0001	< 1	-0.455	0.051	-0.464	0.452
2.2	< 0.0001	< 1	-0.458	0.071	-0.466	0.466

Source: Author's calculations.
Notes: see footnote 60. The table shows magnitude of selection on unobservables, range of significance levels, Hodges-Lehmann point estimates and confidence intervals.

Sensitivity analysis was conducted on all the outcome variables I presented in Table 21, testing the sensitivity of all impact estimates across R1-3. The evidence provided by those tests are in agreement with the description above, namely that the impact estimates presented in Table 21 are sensitive to selection on unobservables[61]; these findings are in agreement with the observations made in the case of SEWA Bank (as discussed in chapter 4).

The next section re-analyses the PnK data set using PSM expanding Chemin's analysis by using a data set with transparently calculated variables which were obtained during the re-construction of RnM's data set. In addition, a different logit model is suggested as well as new comparison groups with the objective to obtain more informative impact estimates and to determine whether using PSM is a viable strategy in this case.

[61] The detailed results from those sensitivity tests are not presented here but the relevant STATA do-files can be made available upon request.

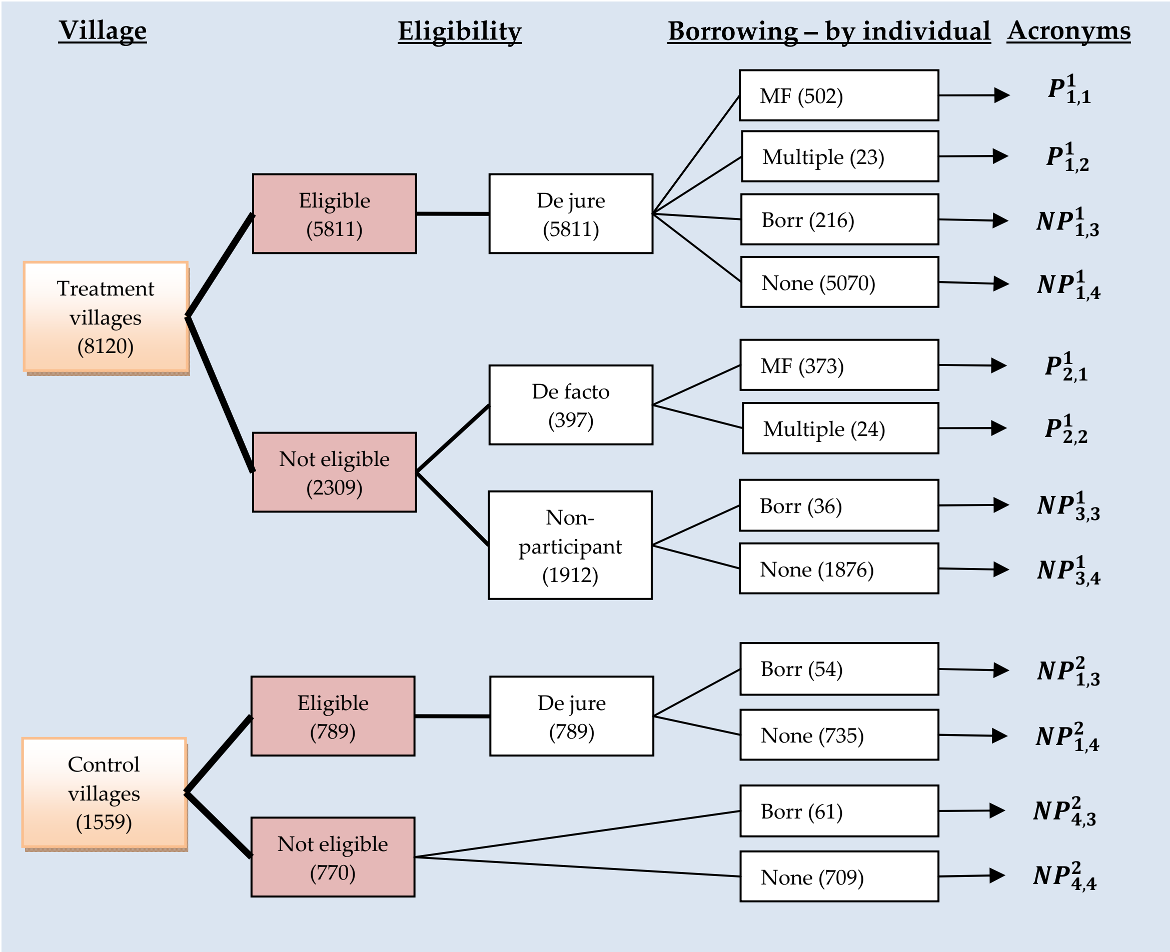
In addition, Khandker's panel data set has not yet been fully explored, only RnM have made an effort in this regard but they have merely focused on replicating the panel. Hence, I contribute to the debate by applying PSM more rigorously and by re-investigating the panel data set applying PSM and DID as well as other panel data techniques.

5.4. PSM – ship of fools?

As discussed in previous chapters, identifying and quantifying the counterfactual represents a challenge, which is exacerbated in the particular context of microfinance by the provision of loan products through other formal or informal credit organisations as well as competing MFIs which might have entered an area where MFIs are already present offering similar products and services. Most microfinance impact evaluations are designed on the assumption that other formal and informal credit organisations would not have entered. However, as illustrated by Figure 9, realities are often different and it is not possible to accurately assess impact by simply comparing microcredit participants with non-participants because those non-participants make use of other sources of borrowing which effectively disqualifies them from being potential control group members. This problem is rarely acknowledged in microcredit impact evaluations and adversely affects impact estimates as discussed in sections 5.3.1. and 5.3.2. Looking at the case of PnK's research design, Figure 9 suggests that many treatment options exist across target and non-target households and across treatment and control villages. Households have access to and participate in other finance schemes, i.e. they obtain loans from other formal or informal sources of borrowing. Hence, simply comparing microcredit participants with non-participants will lead to unreliable results unless realities are taken into account and microcredit participation versus participation in the next best alternative, i.e. other non-microcredit sources of borrowing, is assessed. The studies by PnK, Morduch, Pitt, Chemin, RnM and others who used the PnK data paid no attention to those alternative sources of borrowing and thus doubts are raised about the reliability of their impact estimates.

To begin with, I approach the issue of multiple borrowing from an empirical viewpoint. The different sources of borrowing can be split into formal borrowing such as government banks and informal borrowing such as friends, relatives and neighbours (Table 16 illustrated the multiple sources of borrowing among PnK study participants). These different sources of borrowing are all different in terms of their objectives and target groups and might attract individuals who are very different to the typical microcredit participant in terms of observable and unobservable characteristics.

Figure 9: Availability of treatment options in PnK study



Source: Author's illustration.

Notes:

1. MF = Participant in microfinance only; Multiple = Participant in microfinance and other non-microfinance (formal/informal) borrowing; Borr = Participant in other non-microfinance (formal/informal) borrowing; None = No borrowing at all
2. The number of individuals is given in brackets.
3. Eligibility is < 0.5 acres of land.
4. Explanation of acronyms:

$$Y^a_{b,c}$$

Where:

Y = treatment status

a = village type (1=treatment village, 2=control village)

b = eligibility (1=eligible de jure, 2=not eligible de facto, 3=not eligible non-participant, 4=not eligible)

c = treatment option (1=MF, 2=Multiple, 3=Borr, 4=None)

Figure 9 illustrates some sample characteristics of the PnK data. A noticeable feature is the difference in eligibility between treatment and control villages. An individual is much more likely to be eligible for microfinance programmes in the treatment than control villages reflecting the higher proportion of landless and marginal farmer households (owning less than 0.5 acres of land). In control villages about 50% of all individuals are eligible; while in treatment villages more than 70% fulfil the eligibility criterion (see Table 23).

Table 23: Eligibility by village category in %

Eligibility	Village category		
	Control	**Treatment**	**All**
Not eligible	49.39	28.44	31.81
Eligible	50.61	71.56	68.19
Total	100.00	100.00	100.00

Source: Author's calculations.
Notes: PnK data across R1-3 downloaded from the World Bank website are used.

Figure 9 shows that out of 922 de facto microfinance participants 47 had sources of borrowing other than microcredit. Among the eligible individuals in treatment villages, 216 who did not participate in microfinance had borrowing from other formal or informal sources, but 5,070 (87%) did not report borrowing. 397 (17%) not eligible individuals in treatment villages (out of 2,309 not eligible ones) participated in microfinance – a significant proportion. In all the treatment villages 299 individuals had borrowings from other sources. In the control villages, there were a lower proportion of eligible individuals, but the borrowing from non-microfinance sources in R1-3 was much greater than among treatment villages (8% versus 3.5%). This suggests that microfinance may have partly crowded out other formal or informal sources of borrowing (as investigated by Khandker, 2000). Coleman (1999), Fernando (1997) and others point out that microfinance borrowers might make use of other sources to repay microcredit loans, and that microcredit borrowing does not exhaust their credit requirements. Participants from the control group used in fact other formal or informal sources of borrowing. This clearly shows that the empirical strategy envisaged by PnK cannot work since a comparison between treatment and control group members is confounded by the difference between the villages in important characteristics. Therefore, an alternative strategy using different comparisons may be more appropriate and provide more soundly based and relevant impact estimates.

Given the theory of impact evaluation and what I suggest are the characteristics of microfinance markets in rural Bangladesh (as described in chapter 2), these characteristics suggest a number of possible groups from which the counterfactual sample can be drawn. This strategy is supported by the application of PSM which matches participants and non-participants from within these possible groups on the basis of observable characteristics. Firstly, there is the question of whether the participants are a homogeneous group, and specifically what characterises them. A significant proportion of microfinance borrowers are not formally eligible. This was discussed earlier in this chapter when I refuted Pitt's claims, and suggested that it is very unlikely that microfinance participants who cultivate more than 0.5 acres of land have land of much lower value than eligible participants. Secondly, there is the question of whether microfinance participants who borrow from other sources should be considered similar to those who borrow from MFIs alone. It is possible that they have different characteristics, either lower ability to pay, causing them to borrow to repay their microfinance loans, or having greater access due to unobserved (or observed) characteristics, i.e. using the acronyms introduced in Figure 9 I can compare $P^1_{1,1} + P^1_{2,1}$ versus $P^1_{1,2} + P^1_{2,2}$. However, directly comparing $P^1_{1,1} + P^1_{2,1}$ with $P^1_{1,2} + P^1_{2,2}$ will not yield any meaningful results since the sample sizes of $P^1_{1,2} + P^1_{2,2}$ are too small and hence too few individuals are available for matching. Thirdly, since not eligible non-participants are observably different to eligible participants they are not a suitable control group. Fourthly, there is the question of whether the population of control villages can be considered appropriate counterfactuals at all since the village economies differ in ways which mean that the eligible participants (owning less than 0.5 acres of land) are significantly different from eligible non-participants in the treatment villages.

Nonetheless, it seems likely that the eligible individuals in the control villages are the most suitable control group there is (with or without those who borrow from non-microfinance sources), i.e. $NP^2_{1,3} + NP^2_{1,4}$ may be a suitable control group, or at least the most suitable available. The next most appropriate control group is the eligible non-participants in treatment villages (with or without those who borrow from non-microfinance sources), i.e. $NP^1_{1,3} + NP^1_{1,4}$ can be a control group. The matching procedure in STATA allows me to pool or combine these control groups according to the type of comparison that is being made and depending on whether impact is assessed across treatment and control villages or separately within villages.

The treatment options that are available in the treatment villages are: 'Participant in microcredit only', 'Participant in microcredit and other non-microcredit borrowing', 'Participant in non-microcredit borrowing' and 'No borrowing at all'. Those four treatment options are represented by dummy variables (see Table 24) Y^{MF}, $Y^{Multiple}$, Y^{Borr} and Y^{None} each corresponding to the respective treatment. The effectiveness of microcredit can now be assessed by comparing each treatment option with the others and with a control group which contains non-participants that have no reported sources of borrowing whatsoever.

Table 24: Available treatment options in treatment villages

Treatment	$\boldsymbol{Y^{MF}}$	$\boldsymbol{Y^{Multiple}}$	$\boldsymbol{Y^{Borr}}$	$\boldsymbol{Y^{None}}$
Participates microfinance (MF)	Observable as Y	Counterfactual	Counterfactual	Counterfactual
Participates microfinance and other borrowing (Multiple)	Counterfactual	Observable as Y	Counterfactual	Counterfactual
Participates other borrowing (Borr)	Counterfactual	Counterfactual	Observable as Y	Counterfactual
Participates no borrowing (None)	Counterfactual	Counterfactual	Counterfactual	Observable as Y

Source: Table adjusted from Morgan and Winship (2007), p. 55.
Notes: MF = Participant in microfinance only; Multiple = Participant in microfinance and other non-microfinance (formal/informal) borrowing; Borr = Participant in other non-microfinance (formal/informal) borrowing; None = No borrowing at all.

In the PnK data set, some individuals borrow from more than one source irrespective of their eligibility (as outlined in Figure 9) and are spread across treatment and control villages thereby making the analysis more complex. As a result, a number of comparisons are possible; e.g. assessing either the impact of microfinance participation or the impact of eligibility (as discussed by Morduch (p. 7)) amongst other comparisons. Morduch suggests focusing on assessing the impact of eligibility (p. 7). However, I am essentially interested in assessing whether microcredit participants do better than comparable participants in other borrowing schemes or non-participants irrespective of the eligibility criterion. I thus propose the following comparisons, again expressed using the acronyms introduced in Figure 9. The explanation below expands on the discussion above where the suitability of control groups was discussed. All comparisons include the spouses of microcredit participants as potential matches:

1. $P_{1,1}^{1} + P_{2,1}^{1}\ versus\ NP_{1,4}^{1} +\ NP_{1,4}^{2}$

This comparison looks at de jure and de facto microfinance participants versus all other eligible individuals across treatment and control villages that do not have any other borrowing at all, groups Y^{MF} and eligible Y^{None} are compared. Since all individuals in this comparison fulfil the eligibility criterion, I assume a certain degree of homogeneity within these groups which makes them suitable for comparison.

2. $P_{1,1}^{1} + P_{2,1}^{1}\ versus\ NP_{1,4}^{1} +\ NP_{3,4}^{1} +\ NP_{1,4}^{2} +\ NP_{4,4}^{2}$

This comparison is analogous to comparison 1.; it compares de jure and de facto microfinance participants versus all other individuals but irrespective of eligibility across treatment and control villages that do not have any other borrowing at all, i.e. groups Y^{MF} and all eligible and not eligible Y^{None} are compared.

3. $P_{1,1}^{1} +\ P_{1,2}^{1} +\ NP_{1,3}^{1} +\ P_{2,1}^{1} + P_{2,2}^{1} + NP_{3,3}^{1} + NP_{1,3}^{2} + NP_{4,3}^{2}\ versus\ NP_{1,4}^{1} +\ NP_{3,4}^{1} +$ $NP_{1,4}^{2} +\ NP_{4,4}^{2}$

In this comparison all individuals that participate in either microfinance or other non-microfinance borrowing across treatment and control villages and across eligibility criteria are pooled. In other words, $Y^{Multiple} +\ Y^{Borr}$ versus all eligible and not eligible Y^{None} are assessed.

4. $NP_{1,3}^{1} +\ NP_{3,3}^{1} + NP_{1,3}^{2} + NP_{4,3}^{2}\ versus\ NP_{1,4}^{1} +\ NP_{3,4}^{1} +\ NP_{1,4}^{2} +\ NP_{4,4}^{2}$

Finally, comparison 4. examines individuals that have other non-microfinance borrowing, i.e. Y^{Borr} and compares those with all individuals that have no borrowing at all, i.e. all eligible and not eligible Y^{None} across treatment and control villages irrespective of eligibility. This comparison excludes all microfinance participants.

As mentioned earlier, PnK find microcredit is more effective when women are involved, a claim that Chemin fails to investigate. Hence, I rectify this shortcoming by providing separate impact estimates not only for all individuals but also for women and men separately (discussed in section 5.4.2.).

In addition, I would ideally split other non-microcredit sources of borrowing into formal and informal sources but this is not recommended with the PnK data since the comparison groups become too small to provide any meaningful results. As it will be seen later, the number and quality of matches PSM provides are already rather small raising questions about the robustness of the results. I had encountered similar problems when the sub-group comparisons on the SEWA Bank data were run, as discussed in chapter 4.

5.4.1. Determinants of microfinance participation

Before undertaking PSM, descriptive statistics for the main logit and outcome variables are provided for individuals belonging to the respective treatment groups (see Table 25).

Table 25: Descriptive statistics of individuals belonging to any of the four treatment groups across treatment and control villages and across eligibility criteria

Independent variables	Y^{MF}	$Y^{Multiple}$	Y^{Borr}	Y^{None}
Sex HH head (male=1)	1.055 *0.228*	1.053 *0.224*	1.016 *0.126*	1.036 *0.187*
Age >=15 years	34.616 *10.551*	34.672 *10.529*	40.687 *12.752*	32.618 *15.244*
Age household head (years)	41.044 *11.944*	40.867 *11.881*	43.589 *13.334*	43.178 *12.407*
Number adult male in household	1.344 *0.794*	1.347 *0.796*	1.639 *1.073*	1.612 *1.085*
Marital status (yes=1)	0.873 *0.333*	0.873 *0.333*	0.938 *0.241*	0.337 *0.473*
Landholdings HH head parents	0.209a *0.533*	0.222b *0.549*	0.261c *0.565*	0.248d *0.561*
Landholdings HH head brothers	0.557a *1.071*	0.567b *1.083*	0.766c *1.414*	0.720d *1.223*
Highest education any HH member	3.619 *3.429*	3.649 *3.424*	5.350 *3.944*	4.455 *4.022*
Highest education female HH member	1.178 *2.341*	1.183 *2.346*	2.248 *3.378*	1.788 *3.118*
Savings	3543.534 *5168.575*	3651.482 *5533.265*	4418.86 *20083.07*	4091.61 *17911.66*
Livestock value	2603.311 *3843.594*	2654.342 *3908.822*	3935.737 *5926.48*	3678.958 *6014.571*
Own non-farm enterprise (yes=1)	0.555 *0.4972*	0.556 *0.497*	0.442 *0.497*	0.467 *0.499*
Household size	5.456 *2.063*	5.454 *2.081*	6.191 *2.633*	6.514 *2.735*
Outcome variables				
Total HH expenditure per capita per week (Taka)	76.872 *33.196*	77.805 *34.639*	97.231 *62.918*	81.035 *48.065*
Women non-landed assets (Taka)	2476.51 *6736.685*	2434.943 *6634.52*	2968.477 *13068.11*	2741.315 *9006.549*
Female labour supply, hours per month, aged 16-59 years	101.409 *166.251*	98.449 *165.597*	13.350 *62.106*	18.481 *74.266*
Male labour supply, hours per month, aged 16-59 years	225.607 *332.272*	237.157 *334.151*	456.793 *303.905*	121.542 *257.661*
Girl school enrolment, aged 5-17 years (yes=1)	0.638e *0.481*	0.644f *0.479*	0.681g *0.467*	0.616h *0.487*
Boy school enrolment, aged 5-17 years (yes=1)	0.652i *0.477*	0.656j *0.475*	0.758k *0.429*	0.665l *0.472*
Number of observations	875	922	371	8387

Source: Author's calculations.

Notes: Standard deviation in italics. PnK data across R1-3 downloaded from the World Bank website are used. MF = Participant in microfinance only; Multiple = Participant in microfinance and other non-microfinance (formal/informal) borrowing; Borr = Participant in other non-microfinance (formal/informal) borrowing; None = No borrowing at all.

- a: n = 861; b: n = 908; c: n = 368; d: n = 8278
- e: n = 516; f: n = 542; g: n = 232; h: n = 5621
- i: n = 554; j: n = 582; k: n = 248; l: n = 5769

The mean values in Table 25 differ from the mean values presented by PnK and RnM as illustrated in Appendix 4. ANOVA has been applied examining all possible pairwise comparisons to assess whether the differences in the mean values between the various comparison groups are statistically significant. The ANOVA results show that for most variables differences are not significant at conventional levels of significance, with few exceptions. Mean values of Y^{MF} versus Y^{None} significantly differ for age of household head, landholdings of household head's parents, total household expenditure per capita per week, log of female non-landed assets, female and male labour supply.

Having identified the relevant groups to compare, I now describe the matching process. Chapters 3 and 4 had already explained the workings of PSM and hence I begin by deriving a model that predicts microfinance participation and then run PSM on the various comparison groups.

Armendáriz de Aghion and Morduch (2005) argue that there are certain attributes that influence the decision of participants to join microfinance programmes such as village attributes, observable and unobservable attributes of individuals and broad economic changes. Village attributes consist of particulars of where a person lives, e.g. access to markets. Observable attributes of individuals can be age, education and experience, whereas unobservable attributes relevant to selection into microfinance are often considered for example entrepreneurial skills, organisational abilities, willingness to take risks, etc. (Armendáriz de Aghion and Morduch, 2005). All these factors play a role when assessing the impact of microfinance. For example, according to Armendáriz de Aghion and Morduch (2005, p. 203), there is a high correlation between entrepreneurial skills, age and microfinance participation. Also, if microfinance participants are wealthier than their non-participating peers before joining the programme, as this is suggested by studies conducted by Coleman (2006) and Alexander (2001), then they might have more potential for income growth. In an earlier paper, Coleman (1999) is supportive of this view and lists further unobservable characteristics such as access to social networks and business skills that tend to increase the likelihood of individuals participating in microfinance.

Moreover, potential borrowers may prefer to form groups with individuals that are less risky and more likely to repay (Chowdhury, 2010; Ghatak, 1999); this notion is referred to as assortative matching as explained in chapter 2. Also, it is argued that groups are homogeneous in their set-up and formed on the basis of the above mentioned observable and unobservable characteristics as well as other criteria such as similar ethnic background, neighbourhood, occupation, marital status, trust between group members, etc. (Cassar, Crowley and Wydick, 2007). However, the PnK data do not have variables representing or allowing me to proxy for these unobservable characteristics.

Given the variables that the PnK data provide, the following model is used building on work published by Armendáriz de Aghion and Morduch (2005), Coleman (2006), Alexander (2001) and Coleman (1999):

$$(31) \qquad \mathrm{y}_{ij} = \mathrm{C}_{ij} + \mathrm{G}_{ij} + Z_{ij} + \varepsilon_{ij}$$

<u>Where:</u>

y_{ij} = participating household

C_{ij} = vector of individual-specific variables

G_{ij} = vector of household-specific variables

Z_{ij} = village-level fixed-effects

ε_{ij} = non-systematic error term

The dependent variable (y_{ij}) in the model presented in equation (31) represents eligible participants (*i*) in village (*j*); a value of 1 is assumed when an individual participates and a value of 0 if not. C_{ij} is a vector or individual-specific variables such as age and marital status, G_{ij} is a vector of household-specific variables representing variables such as education and wealth.

Before discussing the logit model results in Table 26, a brief note is needed with regard to the choice of the model. Propensity scores can be predicted by either applying a logit or a probit model. Logit and probit models are rather similar, although, logit models are mathematically simpler than probit models and hence preferred by many researchers (Gujarati, 2003). As a result, a logit model is chosen for the above mentioned reasons but comparative results are reported for a probit model.

The logit model in Table 26 builds on the model specifications used by Chemin and corresponds to the advice from Armendáriz de Aghion and Morduch (2005), Coleman (2006), Alexander (2001) and Coleman (1999). However, new variables have been added, specifically a dummy variable for multiple borrowing; I suspect that individuals borrowing from other non-microfinance sources are less likely to participate in microcredit programmes since their demand for credit is already - at least partially - satisfied by the use of the alternative source of borrowing. Moreover, the data suggest that more men (n = 323) access other non-microcredit sources of borrowing than women (n = 48). And more women than men participate in microcredit rather than other sources of borrowing (641 women versus 281 men).

Table 26: Logistic regression model predicting the probability of microfinance participation using eligible individuals

Independent variables	Means	Logit
Multiple Borrowing (yes=1)	0.244 *0.429*	-1.019*** *0.000*
Sex HH head (male=1)	1.041 *0.198*	0.700*** *0.001*
Age >=15 years	33.503 *14.548*	0.008** *0.011*
Age household head (years)	42.841 *12.441*	-0.010** *0.012*
Number adult male in household	1.565 *1.052*	-0.246*** *0.000*
Marital status (yes=1)	0.411 *0.492*	1.214*** *0.000*
Landholdings HH head parents	0.241 *0.556*	-0.052 *0.516*
Landholdings HH head brothers	0.692 *1.210*	-0.052 *0.181*
Highest education any HH member	4.342 *3.974*	0.022 *0.143*
Highest education any female HH member	1.724 *3.052*	-0.067*** *0.001*
Savings	3925.60 *16800.93*	0.000** *0.039*
Livestock value	3414.713 *5744.381*	-0.000 *0.107*
Own non-farm enterprise (yes=1)	0.464 *0.499*	0.320*** *0.000*
Household size	6.309 *2.695*	-0.044* *0.056*
Village dummies		Yes
Number of observations		5436
Pseudo R-squared		0.145

Source: Author's calculations.

Notes: p-values in italics. * significant at 10%, ** significant at 5%, *** significant at 1%. PnK data across R1-3 downloaded from the World Bank website are used. Also note that age excludes everybody below the age of 15 since it is assumed that microcredit participants are 15 or above, hence including all ages in the logit can be misleading. The variable 'own non-farm enterprise' could be a proxy for business/entrepreneurial skills.

Other variables such as sex of household head, highest education of any female household member and marital status were also newly included with the intention to improve the predictive power of the model. The pseudo R-squared, however, is still rather low at 0.145 and thus lower than the pseudo R-squared of Chemin's specification 3 which has a value of 0.3313 (see Table 17). Again, reasons for this discrepancy can be the underlying differences in the data sets used. A low pseudo R-squared will have implications for the quality of the matches and thus the robustness of the impact estimates, and consequently may have implications for the conclusions I draw, but even after considerable re-examination of the data and variable constructions I could not reproduce the level of pseudo R-square reported by Chemin.

Table 26 shows that the main variables associated with microfinance participation that are statistically significant at 1% are multiple borrowing, sex of household head, number of adult males in the household, marital status, highest education of any female household member and ownership of a non-farm enterprise. Furthermore, age, age of household head and the amount of savings are significant at 5%. Household size is the only variable that is significant at a 10% significance level. Noteworthy is the coefficient for multiple borrowing which is negative; this indicates that individuals are less likely to participate in microcredit when utilising other sources of borrowing as hypothesized earlier. Next, the propensity scores are predicted using the logit model outlined in Table 26 and the various treatment groups are compared by employing PSM.

5.4.2. Treatment group results

I argued earlier that PnK's empirical strategy did not work for various reasons including mistargeting, contaminated control groups, and so on (see section 5.3.1.). I further argued that additional comparisons are required to obtain impact estimates of microcredit participation that were more relevant to the desired evaluation because they compared more homogeneous target groups. An additional question is raised, namely is microfinance indeed the key to poverty alleviation or is the use of other formal or informal sources of finance equally effective in improving the socio-economic well-being of the poor? This section addresses this question and presents the PSM results for the proposed comparisons of the four treatment groups (see Table 27) in simplified form showing the sign and statistical significance of the estimated impact. Two different matching algorithms were applied, i.e. 1-nearest neighbour matching with replacement and kernel matching using three different bandwidths (0.01, 0.02 and 0.05) with the objective to assess the degree of variability of the different matching results across algorithms. The decision for using those algorithms was made in an arbitrary way since the literature in this area is not yet very developed as briefly discussed in chapter 3. Morgan and Winship (2007, p. 109) argue that kernel matching which was first introduced by Heckman et al, (1998) and Heckman, Ichimura and Todd (1998) appears to be the most efficient and preferred algorithm. In addition, 1-nearest neighbour matching was chosen for its popularity which is probably due to its being easy to understand and comparatively easy to implement. Furthermore, as described in section 5.3.3., the distribution of the propensity score was examined for the various treatment groups before the matching procedure was implemented[62]. The results suggest some overlap for most treatment group comparisons. However, as before, the common support region is often rather narrow, i.e. very few good matches were found across the various treatment groups. The robustness of the matching estimates is further examined later in this section using sensitivity analysis.

First, I discuss the four main comparisons identified above before comparing my impact estimates with the ones presented by PnK, Morduch, Chemin and RnM. Table 27 and Table 28 provide the impact estimates for 1. Y^{MF} versus eligible Y^{None}, 2. Y^{MF} versus Y^{None}, 3. $Y^{Multiple}$ + Y^{Borr} versus Y^{None} and 4. Y^{Borr} versus Y^{None}. Table 27 illustrates the impact estimates for microcredit participation for all participants (male and female) while Table 28 provides separate impact estimates for male and female participants separately. I opted for a simplistic representation of the impact estimates to facilitate understanding the direction and significance of the impact estimates and to illustrate the broader trends more easily without diverting attention from the main argument. The number of stars (*) in a cell indicates statistical significance or strength of association while the sign has conventional interpretation. The detailed results in conventional (number and sign) format are available upon request.

[62] The different graphs presenting the distribution of the propensity scores for all treatment group comparisons are not further presented here because there are simply too many of them all confirming the same point namely that the common support region was often too narrow. Those graphs can me made available upon request.

Table 27: Simple matching estimates across gender using 1-nearest neighbour matching with replacement and kernel matching bandwidth 0.05 for all four comparison groups

Outcome variables	Y^{MF} **vs eligible** Y^{None}	Y^{MF} **vs** Y^{None}	$Y^{Multiple}$ + Y^{Borr} **vs** Y^{None}	Y^{Borr} **vs** Y^{None}	Y^{MF} **vs eligible** Y^{None}	Y^{MF} **vs** Y^{None}	$Y^{Multiple}$ + Y^{Borr} **vs** Y^{None}	Y^{Borr} **vs** Y^{None}
	Nearest neighbour matching				**Kernel matching, 0.05**[63]			
Variation of log per capita expenditure (Taka)	+	-	-	+*	-	-	-	+**
Log per capita expenditure (Taka)	+	+	+	+**	-	-	+	+***
Log women non-landed assets (Taka)	+***	+**	-	+	+***	+***	+	+
Female labour supply, aged 16-59, hours per month	+***	+***	+***	-***	+***	+***	+***	-***
Male labour supply, aged 16-59, hours per month	-	-***	+***	+***	-	-***	+***	+***
Girl school enrolment, aged 5-17 years	+	+***	+	+*	+**	+*	+**	+
Boy school enrolment, aged 5-17 years	+	+*	+*	+*	+	+	+**	+***

Source: Author's calculations.
Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%. PnK data across R1-3 downloaded from the World Bank website are used, STATA routine psmatch2 is applied. The logit model outlined in Table 26 is used. The results in this

[63] Kernel matching with bandwidth 0.01 and 0.02 were applied in addition to 0.05 but the various bandwidths results did not differ much and thus only the results using a bandwidth of 0.05 are shown here.

table refer to the differences in the mean values between matched samples. As in chapter 4, 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.

The objective of Table 27 is to illustrate whether microfinance is indeed the answer to poverty alleviation or whether other formal and/or informal sources of finance can be equally effective in achieving the same. The results are rather mixed with different comparisons showing different levels of significance for different outcome variables. It appears that individuals participating in other non-microcredit sources of borrowing do significantly better than microcredit participants, which could of course be due to selection on unobservables (the vulnerability of these results to unobservables is discussed in section 5.4.2.1.) - this is comparison 4., Y^{Borr} versus all eligible and not eligible Y^{None} - in terms of increasing their total expenditure per capita, male labour supply and boys' school enrolment. However, when comparing Y^{MF} versus all eligible and not eligible Y^{None} (comparison 2.), microcredit participation appears to significantly improve women's non-landed assets, female labour supply and girls' school enrolment, more so than participation in other non-microcredit sources of borrowing. Similarly comparison 1., Y^{MF} versus eligible Y^{None} confirms that microcredit participation significantly improves women's non-landed assets and female labour supply. However, most other outcome variables remain insignificant within this comparison. It appears that a comparison among strictly eligible individuals only leads to more mixed and less significant results.

The results above are further confirmed when women are involved in borrowing (see Table 28). It seems that microfinance participation has an apparently significantly positive impact on female related outcome variables such as women's non-landed assets, female labour supply and partially on girls' school enrolment. However, there is little effect on the remaining variables. It should be noted that the results vary across matching algorithms, mainly in terms of their level of significance; thus participation in microcredit or non-microcredit borrowing can either have a strongly significant impact by one algorithm but low significance by another, although the sign is generally the same. Furthermore, in the case of Y^{MF} versus all eligible and not eligible Y^{None} de facto and de jure eligibility criteria were applied but little differences across results were found. Occasionally the level of significance varied but the underlying trend remained the same, and hence it can be concluded that the eligibility criterion does not have much of an effect when assessing the impact of microfinance.

Table 28: Simple matching estimates split by gender using 1-nearest neighbour matching with replacement and kernel matching bandwidth 0.05 for all four comparison groups

Outcome variables		Y^{MF} vs eligible Y^{None}	Y^{MF} vs Y^{None}	$Y^{Multipl}$ + Y^{Borr} vs Y^{None}	Y^{Borr} vs Y^{None}	Y^{MF} vs eligible Y^{None}	Y^{MF} vs Y^{None}	Y^{Multip} + Y^{Borr} vs Y^{None}	Y^{Borr} vs Y^{None}
		Nearest neighbour matching				**Kernel matching, 0.05**			
Variation of log per capita expenditure (Taka)	Women	-	-	-	+	-	-	-	+
	Men	-	-	-	+***	-	-*	-	+***
Log per capita expenditure (Taka)	Women	-	+	+	-	-	-	+	+
	Men	+	+	+	+***	-	-	+	+***
Log women non-landed assets (Taka)	Women	+**	+***	+*	+*	+***	+***	+**	+
	Men	+**	+	+**	+	+**	+	+	-
Female labour supply, aged 16-59, hours per month	Women	+***	+***	+***	+*	+***	+***	+***	+***
	Men	-***	-***	+***	-***	-***	-***	+***	-***
Male labour supply, aged 16-59, hours per month	Women	-***	-***	-	-***	-***	-***	-***	-***
	Men	+***	+***	+***	+***	+***	+***	+***	+***
Girl school enrolment, aged 5-17 years	Women	+	+	+**	+	+	+*	+	+*
	Men	+**	+	+**	+	+	+	+**	+
Boy school enrolment, aged 5-17 years	Women	+	+*	+	+*	+	+	+	+
	Men	+	+	+**	+*	-	+	+*	+***

Source: Author's calculations.

Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%. PnK data across R1-3 downloaded from the World Bank website are used, STATA routine psmatch2 is applied. The logit model outlined in Table 26 is used. The results in this table refer to the differences in the mean values between matched samples. As in chapter 4, 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.

To conclude, the findings presented in Table 27 and Table 28 are mixed and it is not obvious that microcredit participation is associated with more significant impacts than participation in other non-microcredit sources of borrowing. Comparison 3. which looks at $Y^{Multiple} + Y^{Borr}$ versus all eligible and not eligible Y^{None} suggests that finance in general appears to make a difference. The results in Table 27 across all comparisons provide evidence that participation in either microcredit or other sources of borrowing is associated with significantly positive impacts on literally all outcome variables. It appears that any form of finance – microcredit, formal or informal borrowing - is effective in improving the well-being of participating households. This is interesting and relates to the findings of the studies I mentioned in the introduction which cast doubts on the microfinance phenomenon (Banerjee et al, 2009; RnM; Karlan and Zinman, 2009). Studies investigating the impact of the various financing tools such as credit, savings, insurance and business services provided by other formal and/or informal financial institutions separately should be further encouraged to assess whether the hype surrounding microfinance is indeed justified or whether alternative forms of finance are as effective and thus deserve to be equally promoted.

Furthermore, when examining the results by gender (see Table 28), I find that impacts for male labour supply are more beneficial for men in the case of male borrowing. Similarly the impact of female labour supply is more beneficial for women in the case of female borrowing. Again, the results differ in terms of significance across matching algorithms but the underlying trend remains the same. However, how robust are those findings? The various STATA routines such as psmatch2 and pscore[64] were applied across the various matching algorithms. The results obtained from those routines did not vary dramatically which suggests some degree of robustness. However, few good matches were found during the matching procedure, variables were dropped and the balancing requirements were not always met. This raises the question of the suitability of PSM in the context of PnK. The PnK data set has very few control households which is a major drawback since PSM works best when more control than treatment

[64] The STATA routine psmatch2 was developed by Leuven and Sianesi (2003) and pscore was developed by Becker and Ichino (2002).

households are available. Moreover, a rich and high quality data set is required to optimise results (Smith and Todd, 2005). Sensitivity analysis, as argued in sections 3.6.1.2.6., 4.8.3. and 5.3.5. is a more appropriate method of assessing the robustness of these results.

5.4.2.1. Sensitivity analysis on treatment group comparisons

Sensitivity analysis was carried out on all outcome variables for all four treatment group comparisons. As before, the results concur with those presented earlier in this chapter as well as with those presented in chapter 4, i.e. the impact estimates are sensitive to selection on unobservables. In the majority of cases, however, the value of Γ was very small; usually < 2 and often close to 1 which indicates that only a rather small effect of unobservables is required to render the treatment effects statistically insignificantly different from zero. Further to the discussion presented in chapter 4, the case of PnK provides additional evidence that PSM estimates of treatment effects are likely to be vulnerable to selection on unobservables. However, in combination with sensitivity analysis I can at least quantify the likelihood of the unobservables required (to offset the treatment effect), and hence sensitivity analysis should be a prerequisite every time PSM is implemented (Ichino, Mealli and Nannicini, 2006).

5.4.3. cmp with new model specifications

In addition to the replication of Chemin and the application of PSM to the various treatment group comparisons, I apply cmp to my re-constructed data set using model specifications that are different to RnM's with the objective to (a) see whether my re-constructed data set makes any difference to the findings of RnM (small discrepancies are expected due to differences in the underlying data sets), and to (b) investigate if the use of a new model that includes additional variables, e.g. a variable representing multiple sources of borrowing among others, makes any difference.

As briefly mentioned earlier (see footnote 53), cmp is an econometric package that was developed by Roodman (2009) and which contains a wide range of official and user-written STATA estimation commands - as such it depends on

> "a common approach to modeling such limited dependent variables is to assume that the data-generating process is classically linear and unbounded at its heart, with a normally distributed error term. Link functions of chosen form

translate these latent variables into the observed ones. Examples include the probit, ordered probit, rank-ordered probit, multinomial probit, and Tobit models, as well as those for interval data and truncated distributions. Also common are situations in which it is desirable to model or instrument several such variables at once..." (Roodman, 2009, p. 1).

Roodman (2009) further explains that

> "cmp is the first general Stata tool for this class of models, and even it could be extended much further. At this writing, cmp implements an estimator for all the model types above except rank-ordered probit; and it allows mixing of these models in multi-equation systems" (p. 1).

Hence, the key is to set up an appropriate model which is then run by cmp. I re-ran RnM's complete analysis with new model specifications. I am presenting the findings for the log of per capita expenditure to illustrate my approach to cmp with these new specifications. I replicate the first column of RnM's table 4 (p. 25) which presents the 2-stage least-squares (2SLS) estimates of the impact of cumulative borrowing on the log of per capita expenditure across R1-3. These estimates are analogous to the limited-information maximum likelihood (LIML) fixed effects estimates presented by PnK (RnM, p. 23-24).

After successfully replicating RnM's results (as presented in their table 4 (p. 25)) with their original data as well as my re-constructed data set, I modified RnM's STATA do-file to reflect the changes I made to the model, i.e. I used the model of my earlier analysis as described in section 5.4.1. The original results of RnM as well as my results of cmp with the new model specifications for the log of per capita expenditure are presented in Table 29.

Table 29: 2SLS estimates of RnM replication and with new model specifications

Log per capita expenditure	RnM estimates	New model specifications
Log female borrowing from BRAC	-0.121	0.025
Log male borrowing from BRAC	0.212*	0.106
Log female borrowing from BRDB	-0.304*	-0.161
Log male borrowing from BRDB	-0.136	-0.213**
Log female borrowing from GB	-0.056	-0.082***
Log male borrowing from GB	-0.063	-0.006

Source: RnM, column 1, table 4, p. 25 and author's calculations.
Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%. PnK data across R1-3 downloaded from the World Bank website are used. The impact of cumulative borrowing on the log of per capita expenditure is assessed.

The original RnM estimates as presented in Table 29, column 2 could be closely matched by my replication (see Table 29, column 3) using a new model which includes a dummy for multiple sources of borrowing. Like RnM, I find little or no impact of cumulative borrowing across R1-3 on the log of per capita expenditure. While RnM find significantly positive effects of male borrowing from BRAC and significantly negative effects of female borrowing from BRDB, I find significantly negative effects of male borrowing from BRDB and of female borrowing from GB. Nonetheless, I agree with RnM, in terms of the size and the sign of most of the coefficients (except for female borrowing from BRAC), that microcredit borrowing has no or little effects on the log of per capita expenditure. Furthermore, I ran cmp with the new model specifications not only on the log of per capita expenditure but on all other outcome variables as well. Next, I discuss my results and compare them with those of the other studies that dealt with the PnK data.

5.4.4. Comparison of results

Having implemented the cmp replication of PnK and RnM with my data constructions and new model specifications as well as the various treatment group comparisons, this section compares the results of my analysis with those presented by PnK, Morduch, Pitt, Chemin and RnM. Discrepancies across the findings of the various studies dealing with the PnK data are expected due to differences in the underlying data sets (the reasons for this were discussed earlier in this chapter). Furthermore, my results only allow a crude comparison to the ones provided by PnK, Morduch, Pitt, Chemin and

RnM since different estimation strategies and methods across studies are used. All those studies also neglect to account for other non-microcredit sources of borrowing hampering a direct and accurate comparison with my results. Appendix 5 provides a very simplified overview of the headline results of the various studies with the purpose to present a quick summary of the methods used, the particularities of each study and their key findings. The reader is encouraged to revisit the original studies in their entirety for further details.

In brief, PnK and Pitt find that microcredit participation has a significantly positive impact on the log of total per capita expenditure which Morduch and RnM cannot confirm. In fact, Morduch and RnM find a significantly negative impact on the log of total per capita expenditure. Chemin focuses on investigating the impact on the variation of the log of total per capita expenditure and finds that there is no effect of microcredit on expenditure variation. My results are consistent with those presented by Morduch, RnM and Chemin; most of the results I present in Table 27 and Table 28 across treatment group comparisons and across gender are insignificant. Consequently, I conclude that microcredit has no impact on either the log of per capita expenditure or the variation thereof. My cmp results which were obtained by using a different model to RnM are also consistent with those results. All studies (other than PnK) agree that there is no evidence to support PnK's original claims regarding the impact of microcredit on expenditure. With regard to the log of women's non-landed assets, PnK and Pitt find significantly positive impacts which are confirmed by my PSM (see Table 27 and Table 28) and cmp results while Chemin's results are positive but insignificant. RnM also find that the impact of microcredit on the log of women's non-landed assets is significantly positive in particular when women borrow, as confirmed by my results (see Table 28). Furthermore, PnK and Pitt find that female labour supply increases significantly with microcredit participation across males and females. My PSM and cmp results are supportive of this finding but only when women are involved in borrowing; significantly negative impacts are found on female labour supply when men borrow (see Table 28), and hardly any impact is found when male and female borrowing are combined. Chemin's findings are positive but not significant and hence it can be concluded that there is no effect of microcredit borrowing. RnM and Morduch claim that female labour supply significantly decreases especially when men are involved in borrowing (RnM, p.27); this is confirmed by my PSM (see Table 28) and cmp results. The findings for male labour supply are rather mixed. PnK and Pitt find positive as well as negative but mostly insignificant effects; similarly Morduch, RnM and my cmp findings. My PSM results, however, suggest a significantly negative impact across treatment group comparisons, particularly when women are involved in borrowing. Chemin on the other hand argues that male labour supply significantly increases with microcredit borrowing. With regard to school enrolment, PnK, Pitt, Chemin and my PSM findings (see Table 27 and Table 28) demonstrate that girls' and boys' school enrolment increases significantly with microcredit participation while Morduch, RnM and my cmp results find little or no effects. However, in the case of boys' school enrolment, my cmp findings suggest significantly positive impacts when women are involved in borrowing with mixed results when men borrow.

I agree with Morduch, Chemin and RnM in arguing that PnK and Pitt overestimated the impact of microcredit participation. All non-PnK studies concur that in particular the impact estimates for total expenditure per capita were grossly overstated. The evidence suggests that PnK's headline result, which states that total expenditure per household increases especially when women are involved in borrowing, cannot be confirmed. The results of the remaining outcome variables presented by the various studies differ, sometimes substantially, with regard to the direction and significance of the coefficients providing a rather mixed but generally inconclusive picture of microcredit impact.

As mentioned earlier, the reason for those mixed results can partly be explained by the use of different methods and estimation strategies across studies as well as their lax enforcement, e.g. the eligibility criterion was not strictly applied in the case of PnK. Reconstructing PnK's data set posed additional challenges as discussed earlier in this chapter. Moreover, despite re-applying Chemin's estimation strategy, reproducing his original findings was still not possible. The approach I implemented investigated the various treatment group comparisons across female and male borrowing which allows examining the impact of microcredit in a more differentiated way. In addition, sensitivity analysis of my matching estimates suggests that it is not unlikely that unobservables could account for the observed treatment effects, which none of the other studies that re-analysed the PnK data, had investigated. Furthermore, I implemented cmp with new model specifications and the results confirmed my PSM findings. However, despite all those efforts, the challenges of accurately measuring impact remain and the issue of selection bias due to unobservable characteristics persists. Hence, the next section investigates the PnK panel data set with the objective to provide new insights.

5.5. Panel data

Khandker and RnM suggest that longitudinal studies can be informative (RnM, p. 41) and much more convincing than ordinary cross-section studies. In addition, Armendáriz de Aghion and Morduch (2005) and Khandker argue that longitudinal studies can possibly resolve the problem of the unobservables (as discussed in chapter 3). Hence, I now take a closer look at the PnK panel data set which has so far only been explored by Khandker and RnM.

The panel data study conducted by Khandker was briefly introduced in section 5.2.; the objective of Khandker's study is to assess the long-term impact of microcredit participation on poverty reduction. It does not investigate all of the original six outcome variables that were introduced by PnK but instead focuses only on household per capita food expenditure, household per capita non-food expenditure and household per capita total expenditure. Khandker concludes that microcredit has positive impacts on the poorest and reduces poverty among programme participants, especially when women are involved in borrowing, and thus confirms PnK's main headline result. In addition, microfinance contributes to reducing poverty at the village level and thus helps the local economy (Khandker).

A follow-up data set (henceforth R4) was collected in 1998-1999 with the purpose to re-survey the same households that were already interviewed in R1-3. In addition to the original households new households were sampled from the original villages as well as new villages in original and new thanas increasing the overall sample size to 2,599 households (Khandker, p. 271). Khandker explains that

> "because this study relies on panel data to assess the impact of program participation, the study sample was restricted to the 1,638 households that were interviewed in both periods. Of the original group of 1,769 households, 237 households had split into 546 households in 1998/99, resulting in 1,947

> households. To maintain a one-to-one correspondence among matching households, the split households were treated as a single household in the resurvey data" (p. 271).

The rate of attrition between survey rounds was 7.4 percent (Khandker, footnote 10, p. 271). The issue of attrition and the handling of dissolved households posed a challenge for the re-construction of Khandker's R4 data set, e.g. attrition bias is potentially a concern. However, after formal testing, Khandker concludes that attrition bias can largely be ignored (footnote 10, p. 271). I decided to drop all newly sampled households and keep only those from R1-3. As for the treatment of the dissolved households, all members of R1-3 households were manually matched to members from split households in R4. This time-consuming process was the only viable option to obtain a data set that contained only those respondents that were in R1-3 as well as in R4 since matching household members by name, age or sex was not possible due to inconsistencies in the data[65]. At the end of this process a data set that closely resembled RnM's data set for R4 had been compiled. As with the replication of PnK, I re-ran RnM's STATA do-files which replicate Khandker and I could approximate RnM's results; this was expected as RnM's data set is reasonably similar to that compiled for this study.

Before investigating the panel data in more detail, an important point should be raised with regard to re-surveying the original control group members from R1-3. As discussed earlier, the control group of the original PnK study was already rather small but its size was even further diminished between survey rounds. This was due to the rapid developments in Bangladesh's microfinance sector leading to an influx of MFIs expanding into new areas which used to be virgin territory during the earlier survey in 1991-1992. The saturation of the market for microfinance has profound consequences for future studies evaluating the impact of microfinance in Bangladesh since finding suitable control groups, i.e. households that do not participate in microfinance or any other form of finance but are otherwise similar to participating households, has become increasingly difficult.

[65] The problem was that household members in R1-3 had identifiers which did not tally with their identifiers in R4 in quite a number of cases – they were shown as having been in R1-3 but the identifiers were incorrect; these household members had to be manually matched, and this was accomplished for the majority of such cases.

The panel was re-analysed by a combination of PSM and DID which Khandker, Koolwal and Samad (2010) among others claim is the way forward to controlling for observable as well as unobservable characteristics assuming they are time-invariant as discussed in section 3.6.5.1. The PSM matches of R1-3 were retained and merged with R4. PSM using nearest neighbour matching on R1-3 caused some households which did not match on observable characteristics to be dropped, and only matched households were merged with R4. Using the treatment and matched households a regression-adjusted DID model was run on all outcome variables as set out by the following equation which is a fixed effects linear regression model; i stands for household in village *j* at period *t*:

$$(32) \qquad y_{ijt} = \alpha_i + \delta_t + \beta \mathrm{C}_{it} + \theta \mathrm{X}_{it} + V_j + \varepsilon_{ijt}$$

<u>Where:</u>

y_{ijt} = outcome on which impact is measured at period *t*

C_{it} = level of participation in microfinance, i.e. a membership dummy variable constructed based on eligibility criterion (ownership of < 0.5 acres of land), in period *t*

X_{it} = vector of household level characteristics in period *t*

V_j = vector of village level characteristics

α_i = fixed effects unique to household *i*

δ_t = period effect common to all households in period *t*

β, θ = parameters to be estimated

ε_{ijt} = error term representing unmeasured household and village characteristics at period *t*

In addition, the panel was subjected to a random effects model and the results were compared with the estimates obtained from the PSM/DID model as illustrated in Table 30.

Table 30: Impact of microcredit participation, comparison of random effects model with PSM & DID model

Outcome variables	Random-effects model	PSM and DID
Variation of log per capita expenditure (Taka)	-0.016***	-0.017**
Log per capita expenditure (Taka)	-0.001	0.014
Log women non-landed assets (Taka)	0.314**	-0.118
Female labour supply, aged 16-59, hours per month	49.78***	47.37***
Male labour supply, aged 16-59, hours per month	-39.62***	-70.45***
Girl school enrolment, aged 5-17 years	0.147*	0.299***
Boy school enrolment, aged 5-17 years	0.252***	0.289**

Source: Author's calculations.

Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%. PnK data across R1-3 and R4 downloaded from the World Bank website are used, STATA routine xtreg is applied.

Table 30 shows some evidence that microfinance participation has significantly negative impacts on the variation of the log of per capita expenditure and on male labour supply across both models, i.e. the random effects model and the PSM/DID model, while the values for the log of per capita expenditure are insignificant with values close to zero. In contrast to this, the results for female labour supply, and girls' and boys' school enrolment all indicate a positive and significant impact in both models. However, the level of significance for girls' and boys' school enrolment differs between models. The results for the log of women's non-landed assets across both models are inconsistent. The random effects model indicates significantly positive effects while the PSM/DID model shows a negative but insignificant effect. It is unclear why this is the case. Except for the log of women's non-landed assets, the results for all other outcome variables given by both models are consistent across models; slight discrepancies occur in the size of the coefficient and the level of significance.

The panel data results confirm some of the previous cross-section findings described in sections 5.4.2. and 5.4.3. Firstly, I had concluded that microcredit has no impact on the log of per capita expenditure, a notion confirmed by the results presented in Table 30. In contrast to this are Khandker's panel data results which find a positive impact of microfinance participation on the log of per capita expenditure. With regard to the variation of the log of per capita expenditure, I had argued, along with Chemin, that there is no impact of microcredit on this outcome variable (see section 5.4.4.). However, the results in Table 30 suggest otherwise and indicate significantly negative results. Nonetheless, the overall conclusion remains unchanged, i.e. the panel data analysis conducted here does not support PnK's and Khandker's original claims which suggest that microcredit has significantly positive impacts on the log of per capita expenditure. However, PnK and this analysis find significantly positive impacts on the log of women's non-landed assets, particularly when women are involved in borrowing (as also suggested by RnM). These findings are confirmed by the random effects model but not by the PSM/DID model. Moreover, the panel results for female labour supply are supportive of the earlier findings, i.e. microcredit participation has a significantly positive impact as indicated in Table 27 and by PnK. This contrasts with RnM's findings that female labour supply significantly decreases with participation. As discussed earlier, the cross-section findings for male labour supply in PnK, Morduch and RnM were rather mixed. However, the results found in this study for male labour supply were predominantly negative, significantly so, and the panel in fact confirms these findings (Table 27). With regard to school enrolment, PnK, Chemin and the findings in this study demonstrate that girls' and boys' school enrolment is significantly higher among participants in microcredit and the panel data supports these findings. Overall, the cross-section and panel data analyses I conducted in this study cannot confirm PnK's and Khandker's original findings which, this study, RnM, Chemin and Morduch suggest provide an overly positive picture of the impact of microcredit.

Finally, as in the case of the SEWA Bank panel discussed in chapter 4, the PnK panel results confirm the key findings from the cross-section results presented in Table 27. This is surprising since the panel claims to control for unobservables and the results might have been different from the cross-section findings. In the case of SEWA Bank, I argued that the panel was not a 'true' panel because survey respondents were already microfinance participants at the time of the baseline survey and hence a before and after comparison was not possible. The same argument applies to the PnK panel which sampled respondents that were already participating in microfinance at the time of the first survey round that was collected in 1991-1992, and consequently cannot be shown to have been similar to the control groups at that time.

5.6. Conclusion

The replication of PnK and associated studies posed a challenge due to the complex research design and poor documentation. All studies that dealt with the PnK data, i.e. Morduch, Chemin, RnM and this study, agree that PnK overstate the impacts of microcredit. PnK estimated positive and significant impacts for literally all of the six outcome variables with stronger impacts when women were involved in microcredit (PnK, p. 987-988). Morduch argues that PnK overestimated the impact of microcredit because the eligibility criterion was not strictly enforced, i.e. he cannot support PnK's claims that microcredit increases per capita expenditure, school enrolment for children (Morduch, p. 30) or labour supply. Chemin finds lower impact estimates than PnK, though for half the outcome variables such as male labour supply and children's school enrolment he finds a significantly positive impact which contradicts Morduch's findings. Doubts about both Morduch and Chemin arise because of problems in replicating their data constructions. RnM's findings are mixed and mostly insignificant. The reasons for these discrepancies across studies can be explained by shortcomings in the empirical strategy that PnK put forward, e.g. the application of the eligibility criterion was not strictly enforced, and hence a problem with mistargeting occurred (see section 5.3.1.).

Moreover, as discussed earlier, the studies by PnK, Morduch, Chemin and RnM neglect the role of multiple sources of borrowing which has implications for the nature of the control group (i.e. whether it is appropriate), the accuracy of the impact estimates as well as the appropriate definition of the counterfactual. As a result, this study proposed novel treatment group comparisons to examine the impacts found using these more appropriate, and homogeneous, control groups. This strategy found mixed results when comparing microcredit participation with participation in other non-microcredit schemes, and so there is no clear evidence for or against microcredit as such. However, it appears that the utilisation of finance in general has significantly positive impacts across all outcome variables and indicates that other sources of finance can be as effective as microcredit. Many practitioners agree that individuals essentially need to borrow from multiple sources to obtain sufficient funds that would allow them to engage in more productive activities. Many microcredit loans are often too small to meet the needs of microentrepreneurs (Venkata and Yamini, 2010). In addition, multiple sources of borrowing are often required to smooth income and consumption patterns as well as to cope with emergencies (Venkata and Yamini, 2010). Moreover, Coleman (1999), Fernando (1997) and Venkata and Yamini (2010) find that it is common for individuals to use borrowing from one source to pay off the loans of another on time. Overall, criticisms of the more strident and unqualified claims about microfinance are becoming more common and further investigations as to the impact of microcredit versus other financial tools should be encouraged, i.e. using RCTs or carefully designed observational studies that allow the collection of rich and high quality data sets.

The analysis in this chapter has raised doubts about the appropriateness of PSM in the context of PnK. There are often too few matches of low quality due to a small control group sample size and this adversely affects the reliability of the matching estimates. Rich, high quality and large data sets are needed which ideally contain more control than treatment observations (Smith and Todd, 2005). Moreover, sensitivity analysis indicated that it is not unlikely that unobservables could result in over- or underestimating the impact of microcredit. Similar observations were made in chapter 4 when the case of SEWA Bank was discussed.

As to the panel data analysis, Khandker and RnM argued that longitudinal studies remedy the shortcomings of cross-sectional studies but this appears not always to be the case. The panel data results of the random effects model did not provide any new insights and generally confirmed the findings of the cross-section data analysis. Khandker, Koolwal and Samad (2010) among others, claim that the combination of PSM and DID is the way forward since it allows controlling for observable as well as unobservable characteristics assuming that the latter remain constant over time. However, the results of the PSM/DID model did not offer anything different to what was found by the random effects model. As in the case of SEWA Bank, doubts are raised about the ability of techniques such as PSM and DID to account for selection on unobservables with the PnK data perhaps because it is not a ‘true’ panel which would allow a before and after comparison with a more demonstrably appropriate control group. What is compared is the change in outcomes between a group that was already participating in microfinance in R1-3 and a control group surveyed at the same time, with both groups at a later date. As already discussed in chapter 4, this comparison is not adequate for reliably assessing the impact of microcredit and controlling for unobservables because any differences between the treatment and control groups before microfinance cannot be empirically observed in these data.

Overall, what can be learnt? The results provided by the various studies dealing with the PnK data are rather mixed ranging from significantly positive impacts to significantly negative ones depending on the econometric techniques applied. However, all studies agree that PnK and Khandker most likely overstated their impact estimates and the replication of their original findings is challenging. Furthermore, other sources of borrowing need to be accounted for when assessing the impact of microcredit; most evaluations work with contaminated control groups which adversely affect the robustness of impact estimates. Thus, methodological problems still remain particularly selection bias due to unobservable characteristics, inappropriate counterfactuals, and poor data quality as well as control groups that are contaminated and limited in size. As mentioned earlier for the SEWA Bank case, in PnK, the control group is far too small to provide convincing matches which hampers the usefulness of PSM in this context.

To sum up, poor quality data, poor research design and possibly inappropriately implemented econometric techniques fail to illuminate the role of the unobservables. Sensitivity analysis of the matching results indicated that the unobservables could readily confound impact estimates which are demonstrated to be not robust to unobservables. The next chapter discusses the findings presented in this thesis and possible solutions that may allow a better understanding of the unobservables, and why this is important for the impact evaluation arena as a whole.

6. Conclusion

This thesis was written with the objective to contribute to the wider methodological debate on the use of quantitative impact evaluation techniques with a focus on observational data in the context of microfinance in India and Bangladesh. I provide new insights by replicating and re-analysing the existing USAID and PnK panel data with PSM, DID and cmp to control for selection bias and to point up the role of the unobservables in impact evaluations using observational data. The thesis contributes to the microfinance literature by throwing doubt on the claims of the impact of these two well known microfinance projects. The focus is on the study conducted by PnK which is the most authoritative microfinance impact evaluation to date. PnK's study was uncontested for many years, which is surprising given the poor data and questionable research design. However, poor documentation and the general complexities of their econometric modelling may partially explain why few researchers have engaged with PnK.

This chapter begins with a summary of the main arguments of each of the chapters, focusing on the contributions made to the methodological debate. Chapter 2 began by providing the contextual background of the study by describing the characteristics of rural credit markets and the emergence of microfinance in India and Bangladesh. It further described how rural credit markets regularly failed to financially include the rural poor and provided explanations of the failure of government-run credit programmes. The emergence of microfinance can be understood in this context. Microfinance started as a grass-roots movement in the early 1970s in rural Bangladesh. The sector grew rapidly through the 1980s and transformed into an industry spanning the globe, ultimately in the first decade of the 21st century drawing attention from commercial banks and private investors. Despite the growth of microfinance, however, its impact on the socio-economic well-being of the poor remains unclear (Armendáriz de Aghion and Morduch, 2005). Numerous impact evaluations had tried to pin down the success of microfinance but without being convincing. The evidence of microfinance impact almost 40 years after its emergence is still rather mixed (ibid). This is likely partly because of the challenging nature of measuring impact, i.e. identifying an appropriate counterfactual, selection bias, etc. These challenges are not unique to microfinance. However, microfinance evaluations are plagued by additional problems such as fungibility, drop-outs and the powerful presence of selection and/or screening processes that influence microfinance participation (Hulme, 2000; Sebstad and Chen, 1996; Armendáriz de Aghion and Morduch, 2005; PnK; Coleman, 1999; Alexander-Tedeschi and Karlan, 2007).

To understand the challenges of evaluating impact, chapter 3 set out the theoretical foundations of impact evaluation and linked them to the specific problems that commonly arise when assessing the impact of microfinance. The recent methodological debate on the relative merits of RCTs and observational studies was related to the context of microfinance. The evaluation strategies commonly discussed in the impact evaluation literature were introduced and critically examined. I outlined the drawbacks of the various evaluation techniques with regard to their ability to control for selection bias with a particular focus on accounting for selection on unobservables. Throughout this thesis I have argued that the majority of the econometric techniques fail to control for selection on unobservables; nevertheless a wealth of evaluation studies continue to claim that their impact estimates are robust and provide definite answers to the evaluation problem (e.g. PnK; Pitt). This can be misleading. Heckman, LaLonde and Smith (1999) argue that the results of an impact evaluation heavily depend on the quality of the underlying data. In other words, advanced econometric techniques will not be able to control for poor quality data. This point is reiterated by Caliendo and Hujer (2005, p. 1) who state that many evaluations in the past did not provide particularly meaningful results because of the non-availability of rich and high quality data sets. This makes it important, I suggest, that those who are to analyse or who properly understand the analytical techniques and their data dependence should be involved in the design of an impact evaluation early on to ensure the collection of rich data since this is one way to avoid pitfalls in the subsequent analytical process (Rosenbaum, 2002). Rosenbaum and Silber (2001), for example, suggest using ethnographic or other qualitative tools with the objective to improve data collection procedures and the overall design of an evaluation. Heckman, LaLonde and Smith (1999), Rosenbaum (2002), Rosenbaum and Silber (2001) and Caliendo and Hujer (2005) suggest that it is not necessary to introduce ever more sophisticated econometric techniques, but instead focusing on collecting better quality data can be part of the solution to the evaluation problem. Therefore, not only do the econometric techniques employed require scrutiny when assessing the quality of an impact evaluation but so do the underlying data. Angrist and Pischke (2010) respond to Leamer's (1983) pessimistic view on the credibility of econometric methods by claiming that empirical economics has gone through a "credibility revolution" (Angrist and Pischke, 2010, p. 4) thanks to the availability of better data but more importantly thanks to a focus on enhancing the quality of research designs. Whether these claims are credible can be debated; nevertheless they motivated the discussion in the two empirical chapters (4 and 5) which used primary and secondary data from microfinance programmes in India and Bangladesh to contribute to these methodological debates.

The common belief is that microfinance is pro-poor and pro-women (e.g. Yunus, 1999). The evidence presented in this thesis, however, is mixed and does not provide convincing evidence of either effect. If we maintain that the null hypothesis is that there is no effect, then in this case the evidence provided in this thesis fails to contradict this null hypothesis, and also fails to contradict the alternate hypothesis (that there is a positive effect). Failing to contradict the alternate hypothesis encourages one to believe there is a positive effect and therefore to tend to (continue to) reject the null (no effect) hypothesis even though it (no effect) may be true. This of course depends on the decision procedure (see Neyman and Pearson, 1933, for a detailed discussion on decision rules) and weighing the costs and benefits of an intervention. Maybe microfinance does work/has worked well for the poor and women, in which case it is good that the alternate hypothesis has not been rejected. On the other hand, if it is the case that there is/was no effect then 10 - 15 years have been lost (the PnK results first came out in 1996) when one could have been acting on the hypothesis that microfinance does not benefit the poorest and hence seeking alternatives. It is possible that more convincing evidence would have been found of beneficial microfinance effects, especially if the methodological approach taken by PnK (and to a lesser extent in the SEWA Bank study) had been more critically assessed. But, failing this, it may be that many benefits especially for the poorest have been forgone by the continuing belief in the efficacy of microfinance. In other words, failing to reject (accepting) the null hypothesis may have had greater benefits. Since I find that there is no good evidence to contradict the idea that microfinance has little effect on the well-being of the poorest, one might have done well to have looked elsewhere for interventions to benefit them.

I argue that despite the use of advanced econometric techniques, unobservables that drive selection and/or screening processes that determine microfinance participation have not been controlled for. The analysis in chapters 4 and 5 illustrates that many impact evaluation strategies do not adequately account for selection on unobservables and this can have adverse effects on the reliability of impact estimates. Overall, I provide evidence that reduces the credibility of the quantitative support for microfinance and for lending to women in preference to men. Qualitative evidence (Fernando, 1997) strongly suggests other less beneficent interpretations of microfinance impact leading to an unraveling of the microfinance narrative.

Chapter 4 presented the empirical findings from SEWA Bank in India and, with the support of original qualitative investigations, explored selection processes by loan officers, the role of the unobservables and related topics such as social capital. I re-analysed the existing USAID panel data and employed PSM and DID. SEWA Bank has a particular focus on a savings approach and the sample contained borrower, saver and control households which allowed me to conduct various sub-group comparisons to assess the impact of microfinance across those different groups. My results approximate those obtained by USAID, i.e. microfinance has a positive impact on many outcome variables with borrowers doing better than savers who in turn do better than controls. Those findings confirm the notion that savings tools are complementary to credit approaches. However, sensitivity analysis suggests that the matching estimates are not robust to unobservables. In other words, my PSM results are highly sensitive to selection on unobservables which the qualitative evidence I collected during my fieldwork suggested were present, as does the literature (Ito, 2003; Fernando, 1997). Furthermore, the panel data analysis that combined PSM and DID with the aim to control for the unobservables, did not provide any new insights. To the contrary, the panel results confirmed the conclusions of the cross-section analysis. Furthermore, because the USAID data are not a 'true' panel - there is no pre-project baseline which could demonstrate that participants and non-participants were truly equivalent before the project - any impact found from the panel is vulnerable to the charge that the project works but only for the sorts of people who became members.

Thus, these data do not allow a before and after comparison but compare the change in the outcome variable between a group that was already a member of SEWA Bank at the time of the baseline survey and a control group surveyed at the same time, with both groups at a later date. Compared to a proper before and after comparison this may underestimate the total impact (assuming the two groups are indeed comparable). At the same time it eliminates the possibility of showing there were no differences between the participants and controls in the absence (before) SEWA Bank and therefore cannot conclusively control for unobservables. It cannot be shown that the treatment group before treatment was indistinguishable in terms of outcome variables, or, of course, unobservables, from the control group because there are no data from before treatment. Furthermore, there are doubts about the sampling method used to select the control group as this has not been precisely reported by USAID. The sampling procedure described suggests that control households may have less ability to benefit from SEWA Bank services than participating households because they had had that opportunity but either chose (self-selected out) not to participate or were selected out by peers or SEWA Bank staff. Hence, the comparison between treatment and control groups may have been biased by the sampling procedure. The qualitative information described in chapter 4 together with the reporting lacunae indicate that a selection or screening process driven by the unobservables is indeed likely to have been at play. This affects microfinance participation and cross-section and panel differences between the treatment and control groups.

Chapter 5 supports most of the findings presented in chapter 4. However, the replication of PnK and the various studies dealing with the PnK data posed additional challenges due to complex research designs, lack of documentation, and, possibly, poor quality data. As a result, the various studies, i.e. Morduch, Pitt, Chemin and RnM, dealt with differently re-constructed data sets which had implications for the consistency of the results across all studies. Also, different analytical approaches are followed by different authors. The results presented by Morduch, Chemin, RnM and myself are diverse ranging from significantly positive impacts to significantly negative ones depending on the estimation strategy and the methods that were applied. However, all studies other than those directly associated with PnK agree that the original PnK results most likely overstated impacts and that the headline results cannot be confirmed. These findings contrast with the SEWA Bank study which concludes that microfinance participation has on the whole a positive impact on the well-being of the poor and my results confirm this notion, subject to the unresolved question of the unobservables which sensitivity analysis of the PSM results and qualitative fieldwork observations suggest are not unlikely to have been present and influential. I discuss this point in more depth later in this section.

There are further problems with PnK and related studies such as Morduch, Chemin and RnM as they neglect multiple sources of borrowing. This has implications for the appropriate definition of the counterfactual, the quality of the control group, and ultimately for the accuracy of the impact estimates. It is a microfinance reality that many households have multiple sources of borrowing which are often required to invest in enterprises, smooth income and consumption patterns, and to cope with emergencies (Venkata and Yamini, 2010). Moreover, Coleman (1999), Fernando (1997) and Venkata and Yamini (2010) find that it is common for individuals to use borrowing from one source to pay off the loans of another. Since many impact studies fail to take multiple borrowing sources into account, I conducted various treatment group comparisons, for example comparing microcredit participants with participants in other non-microcredit schemes to provide further insights. My findings are mixed and neither support nor undermine the claim that microcredit has a beneficent impact on the poor, i.e. my impact estimates are all highly vulnerable to selection on unobservables. Nonetheless, the results indicate that the utilisation of any form of finance has significantly positive impacts across all outcome variables indicating that other sources of finance can be as effective as microcredit. As mentioned earlier, criticisms of the more strident and unqualified claims about microfinance are becoming more common (see Banerjee et al, 2009; Bateman and Chang, 2009; Dichter and Harper, 2007; Karlan and Zinman, 2009) and further investigations as to the impact of microcredit versus other financial tools are surely warranted.

As indicated by the findings presented in chapter 5, the re-analysis of PnK raised doubts about the ability of PSM to overcome the difficulties posed by the research design. In the case of PnK, the matches could only come from a small portion of the non-borrowers, and can as a result be hypothesised to be of low quality with adverse effects on the reliability of the matching estimates. It is not clear what low quality in matches means since cases are matched by the propensity score. One way to think about this is that a good match implies a low ability to explain participation (I owe this idea to Robert Lensink). The idea here is that you need units of observation with very similar characteristics to participants but who just do not happen to participate. If it is indeed the case that truly good matches are driving the low pseudo R-squared of the logit used to predict the propensity scores, then removing units which are chosen as matches and re-running the logit model should result in a significantly improved pseudo R-squared. If not, then it is the case that the low pseudo R-squared is due to a lack of covariates which adequately characterise participants. It is beyond the scope of the thesis to further explore this issue but more research on this would certainly be interesting. Thus, as argued above, rich and high quality data sets are needed which have the ability to explain participation and, ideally, contain more control than treatment observations (Smith and Todd, 2005).

With regard to sensitivity analysis and the panel data analysis, the points made in chapter 4 are confirmed by the re-examination of the PnK data. In the SEWA Bank analysis I concluded that unobservables are not unlikely to have been present, playing a role in selection into microfinance and influencing the estimated impact; the panel is not a 'true' panel and does not allow a before and after comparison and so cannot exclude unobservables distinguishing SEWA Bank members from the control group. Khandker, Koolwal and Samad (2010) among others argue that combining PSM with DID controls for the unobservables, assuming they remain constant over time, but my findings from the PSM/DID estimation in the case of SEWA Bank and PnK did not provide any new insights and instead the cross-section results were confirmed. If the PSM/DID estimation does not show any results that are different to the cross-section results which are vulnerable to the unobservables, then it follows that the PSM/DID results will certainly be vulnerable to the unobservables as well and one does not actually need a panel to support this. Even if the panel produces different results they are vulnerable to the argument that it is not a proper panel. Therefore, doubts remain about the ability of techniques such as PSM and DID to account for selection on unobservables in the absence of an appropriate research design and a high quality and rich data set.

As briefly mentioned earlier in this section, the re-analysis of SEWA Bank indicates a positive impact of microfinance while the evidence of PnK's re-examination suggests limited or no impact of microfinance. This discrepancy can be explained by the presence of unobservables. As suggested by the results of sensitivity analysis, the impact estimates of SEWA Bank as well as of PnK are both susceptible to selection on unobservables. In addition, differences in terms of country-context, year of data collection, survey design and/or the scale of the study can influence the study results in positive as well as negative ways. For example, the microfinance context in Bangladesh in the early 1990s was certainly very different to the one in Western India in the late 1990s. Moreover, SEWA Bank supports an individual lending scheme in its urban areas while the microfinance programmes investigated by PnK in Bangladesh pursued group-lending schemes in rural areas. This implies that different selection processes may be at play; i.e. the case of SEWA Bank illustrated that vertical social capital (Ito, 2003) plays a role and that selection and/or screening processes exist that are driven by loan officers who recruit individuals into microfinance. This is different to the case of PnK where horizontal as well as vertical social capital (Ito, 2003) is at work, which implies that group members typically self-select into groups or are selected by their peers rather than loan officers. Moreover, there are further differences with regard to the set-up of the microfinance programmes. For example, the three microfinance programmes in Bangladesh focus on a credit-only approach while SEWA Bank favours a savings approach which is complemented by credit where every individual is required to build up savings first and is then 'upgraded' to a borrower status. This 'upgrade' is, again, driven by SEWA Bank's loan officers and hence different unobservables are at play in both cases. This is all speculation, however, because by definition the unobservables cannot be observed by conventional data production techniques and refined qualitative tools are indeed needed to further illuminate their role.

To conclude, limited research design, poor quality data, lack of ethnographic insights as well as inappropriately employed econometric techniques fail to illuminate the role of the unobservables which continue to confound impact estimates. Tools such as sensitivity analysis can provide, admittedly limited, insights as to the magnitude of the effect of unobservables required to render the estimated impact statistically insignificant (and hence their likelihood), but cannot identify them or quantify their roles.

6.1. Lessons learnt and recommendations

In many cases researchers are not part of the initial stages of a study and will have to adjust their evaluation strategy ex-post according to the design of the programme, the selection processes at play and the availability and quality of data. This is not ideal and researchers should be involved in the evaluation design and the data collection process as early as possible in order to be able to obtain high quality data, select an appropriate evaluation strategy and control for potential biases due to observable and unobservable characteristics as early as possible (Rosenbaum, 2002; Heckman, LaLonde and Smith, 1999; Caliendo and Hujer, 2005). Furthermore, field visits and the collection of qualitative information with the aim to gain an understanding of the contextual background of the programme (and processes) under investigation are crucial, i.e. direct observation and interactions with participants, using 'thick description' (Geertz, 1973) based on thorough ethnographic understanding, comprehending the specific-country context and the economic conditions at the time of the evaluation, are important. Gathering qualitative information is particularly important in the case of microfinance where group formation processes are clearly driven by unobservable characteristics such as access to social networks, entrepreneurial skills and organisational abilities (Armendáriz de Aghion and Morduch, 2005; Coleman, 1999). Recent developments in behavioural economics appear to hold potential for a greater understanding of economic behaviour. Thus the presence, mechanisms and effects of variables that are unobservable to conventional survey based research could be further explored through behavioural and experimental games[66] as well as choice experiments played with current and/or potential microfinance participants as well as microfinance staff, to throw light on inclusion and exclusion from microfinance which is partially driven by the unobservables. This suggests that strictly quantitative approaches should be

[66] This is a rapidly growing area of research and it is beyond the scope of this thesis to explore this in depth. The interested reader is referred to Kagel and Roth (1995) for a good introduction to experimental and behavioural economics.

complemented with qualitative ones and/or games with the aim to illuminate the role of those unobservables.

For example, initial group formation could be explored through risk pooling games (Barr, Dekker and Fafchamps, 2010) where participants are asked to form credit groups that pool risk of default. Or stated choice experiments could be played with microfinance staff responsible for group formation. Choice experiments have been commonly used in health economics, environmental economics and marketing to capture people's preferences over hypothetical scenarios, objects and services (Street and Burgess, 2007). Thus, one could identify a set of hypothetical clients who differ on socio-economic attributes and who will be ranked by individual loan officers, or other potential or actual group members, according to their preference for loan approval. Regression methods could control for clients' attributes as well as characteristics of the loan officers that do the ranking, enabling the identification of important determinants of loan officers' (or potential group members') preferences, providing some insight into loan officers' (peers') influence on inclusion and exclusion from microfinance groups, and the role of the unobservables. These approaches can be complemented by post-game interviews (as done by Iversen et al, 2010) with participants in the games to elicit variables that may help identify potential instrumental variables for IV regressions with both the game and survey data. These interviews would focus on recording the social and economic statuses and attitudes to microfinance clients; their interpretation of the games; their employment histories, and in the case of microfinance clients, their entrepreneurial and credit histories. By enumerating variables that are not generally available from questionnaire surveys, and/or by specifying credible proxies for the unobservables, these non-standard data production methods may improve the analysis of observational data.

To conclude, the debate on the appropriateness of the evaluation methods used to account for selection (and placement) bias is far from over. The discussion in this thesis demonstrated that the evaluation techniques commonly employed with observational data have drawbacks in one way or another. This thesis argued that most techniques do not account for selection on unobservables and a clear-cut solution to this issue has not yet been found. There is no clear winner of this methodological debate between 'randomistas' (e.g. Banerjee et al, 2009; Karlan and Zinman, 2009; Duflo and Kremer,

2005) and advocates of observational designs (e.g. Deaton, 2009; Imbens, 2009; Pritchett, 2009). Despite the drawbacks of RCTs and the challenges of observational designs (as outlined in chapter 3), I argue that there is room for both. RCTs, which in principle have the best chance of meeting these challenges (i.e. accounting for selection and placement bias), are not always practical or desirable given a particular evaluation context, which means that the choice of the study design, whether to use a RCT or observational approach, heavily depends on the objectives of the evaluation, access to financial resources and time horizons. Hence, observational studies will continue to play an important role in evaluation (RnM) and value can be added by replicating and re-analysing (Hamermesh, 2007) existing observational data with, or indeed without, new methods, as done in this thesis.

Furthermore, I have argued that PSM and DID, which are relatively new but are being increasingly applied, are not the wondrous tools as advocated by many (e.g. Dehejia and Wahba, 1999 and 2002; Khandker, Koolwal and Samad, 2010) and the raw impact estimates presented in this thesis should be taken with the appropriate qualifications suggested by sensitivity analysis. Not only do these data and methods not provide robust support for the idea that microfinance is highly beneficial to the poor, rather than perhaps benefitting a slightly better off group, or being no better than alternative, less hyped, credit sources, but they leave open the question of whether microfinance is of any real benefit at all. In fact, much of the apparent difference between microfinance participants and controls is likely due to differences in their unobserved characteristics rather than the intervention per se - as suggested by the qualitative evidence (Ito, 2003; Fernando, 1997), and by proper application of PSM with sensitivity analysis to the data analysed here. If indeed there is no good evidence to support the claim that microfinance has an effect on the well-being of the poor or empowers women, then it might have been better to explore alternative interventions over the last decade or so that could have better benefited the poor (or empowered women). It is still unclear under what circumstances, and for whom, microfinance has been and could be of real rather than imagined benefit to the poor. Thus, to get a clearer picture on the impact of microfinance, the quantitative evidence should be complemented with qualitative tools and possibly with behavioural and experimental games to gain a better understanding of the selection mechanisms underlying microfinance participation and the role of the unobservables in this context. Exploring why what appears to have been inappropriate optimism towards microfinance came to be so widespread would also be a suitable subject for further research.

Bibliography

Abadie, A., Drukker, D., Herr, J. L. & Imbens, G. W., 2004. Implementing Matching Estimators for Average Treatment Effects in STATA. *The STATA Journal,* 4 (3), p.290-311.

Abou-Ali, H., El-Azony, H., El-Laithy, H., Haughton, J. & Khandker, S. R., 2009. Evaluating the Impact of Egyptian Social Fund for Development Programs. World Bank Policy Research Working Paper No. 4993, July.

Adams, D. W., 1978. Mobilizing Household Savings through Rural Financial Markets. *Economic Development and Cultural Change,* 26 (3), p.547-560.

Adams, D. W., 1988. The Conundrum of Successful Credit Projects in Floundering Rural Financial Markets. *Economic Development and Cultural Change,* 36 (2), p.355-367.

Adams, D. W., Graham, D. H. & von Pischke, J. D. eds., 1984. *Undermining Rural Development with Cheap Credit.* Boulder: Westview Press.

Adams, D. W. & von Pischke, J. D., 1992. Microenterprise Credit Programs: Déja Vu. *World Development,* 20 (10), p.1463-1470.

Ahlin, C. & Townsend, R. M., 2007. Using Repayment Data to Test Across Models of Joint Liability Lending. *The Economic Journal,* 117 (517), p.F11-F51.

Ahmed, Z., 2002. Revisiting the Politics of Fieldwork: Experience from Bangladesh. In Nurul Alam, S. M., ed. *Contemporary Anthropology. Theory and Practice.* Dhaka: The University Press Limited.

Akerlof, G. A., 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics,* 84 (3), p.488-500.

Alexander, G., 2001. An Empirical Analysis of Microfinance: Who are the Clients? *Northeastern Universities Development Consortium Conference.* Boston, 28-30 September 2001.

Alexander-Tedeschi, G. & Karlan, D. S., 2007. Cross Sectional Impact Analysis: Bias from Dropouts. Unpublished mimeo.

Angrist, J. D., Imbens, G. W. & Rubin, D. B., 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association,* 91 (434), p.444-455.

Angrist, J. D. & Krueger, A. B., 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *The Journal of Economic Perspectives,* 15 (4), p.69-85.

Angrist, J. D. & Lavy, V., 1999. Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, p.533-575.

Angrist, J. D. & Pischke, J.-S., 2010. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives,* 24 (2), p.3-30.

Aportela, F., 1999. Effects of Financial Access on Savings by Low-Income People. *Available at: http://www.lacea.org/meeting2000/FernandoAportela.pdf.*

Armendáriz de Aghion, B. & Gollier, C., 2000. Peer Group Formation in an Adverse Selection Model. *The Economic Journal,* 110 (465), p.632-643.

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

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

Arun, T., Imai, K. & Sinha, F., 2006. Does the Microfinance Reduce Poverty in India? Propensity Score Matching based on a National-Level Household Data. Economics Discussion Paper, The University of Manchester, September.

Ashenfelter, O., 1978. Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics,* 60 (1), p.47-57.

Ashraf, N., Karlan, D. S. & Yin, W., 2006. Female Empowerment: Impact of a Commitment Savings Product in the Philippines. *Available at: http://www.econ.yale.edu/growth_pdf/cdp949.pdf.*

Augsburg, B., 2006. Econometric Evaluation of the SEWA Bank in India: Applying Matching Techniques based on the Propensity Score. Working Paper MGSoG/2006/WP003, Maastricht University, October.

Banerjee, A., Besley, T. & Guinnane, T. W., 1994. The Neighbor's Keeper: The Design of a Credit Cooperative with Theory and a Test. *The Quarterly Journal of Economics,* 109 (2), p.491-515.

Banerjee, A. & Duflo, E., 2010. Giving Credit Where it is Due. *Available at: http://econ-www.mit.edu/files/5415.*

Banerjee, A., Duflo, E., Glennerster, R. & Kinnan, C., 2009. The Miracle of Microfinance? Evidence from a Randomized Evaluation. *Available at: http://econ-www.mit.edu/files/4162.*

Barnes, C. & Sebstad, J., 2000. Guidelines for Microfinance Impact Assessments. Discussion Paper for the CGAP 3 Virtual Meeting October 18-19, 1999 submitted to USAID.

Barr, A., Dekker, M. & Fafchamps, M., 2010. Who Shares Risk with Whom under Different Enforcement Mechanisms? *Available at: http://ipl.econ.duke.edu/bread/papers/working/267.pdf.*

Basu, P., 2006. *Improving Access to Finance for India's Rural Poor.* Washington D.C.: The World Bank.

Basu, A., Heckman, J. J., Navarro-Lozano, S. & Urzua, S., 2007. Use of Instrumental Variables in the Presence of Heterogeneity and Self-selection: An Application to Treatments of Breast Cancer Patients. *Health Economics,* 16 (11), p.1133-1157.

Bateman, M. & Chang, H.-J., 2009. The Microfinance Illusion. *Available at: http://www.econ.cam.ac.uk/faculty/chang/pubs/Microfinance.pdf.*

Becker, S. O. & Caliendo, M., 2007. Sensitivity Analysis for Average Treatment Effects. *The STATA Journal,* 7 (1), p.71-83.

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

Behrman, J. R., Sengupta, P. & Todd, P., 2005. Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico. *Economic Development and Cultural Change,* 54 (1), p.237-275.

Berg, G., 2010. Evaluating the Impacts of Microsaving: The Case of SEWA Bank in India. *Journal of Economic Development,* 35 (1), p.75-96.

Bertrand, M., Duflo, E. & Mullainathan, S., 2004. How much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics,* 119 (1), p.249-275.

Besley, T. & Coate, S., 1995. Group Lending, Repayment Incentives and Social Collateral. *Journal of Development Economics,* 46 (1), p.1-18.

Bhatt, E., 2006. *We Are Poor but So Many: The Story of Self-Employed Women in India.* New Delhi: Oxford University Press.

Björklund, A. & Moffitt, R., 1987. The Estimation of Wage Gains and Welfare Gains in Self-Selection Models. *The Review of Economics and Statistics,* 69 (1), p.42-49.

Blundell, R. & Costa Dias, M., 2000. Evaluation Methods for Non-Experimental Data. *Fiscal Studies,* 21 (4), p.427-468.

Blundell, R. & Costa Dias, M., 2002. Alternative Approaches to Evaluation in Empirical Microeconomics. The Institute for Fiscal Studies, Department of Economics, University College London, Cemmap Working Paper No. CWP 10/02.

Blundell, R. & Costa Dias, M., 2008. Alternative Approaches to Evaluation in Empirical Microeconomics. The Institute for Fiscal Studies, Department of Economics, University College London, Cemmap Working Paper No. CWP 26/08.

Bouman, F. J. A., 1989. *Small, Short and Unsecured: Informal Rural Finance in India.* New Delhi: Oxford University Press.

Browning, M. & Lusardi, A., 1996. Household Saving: Micro Theories and Micro Facts. *Journal of Economic Literature,* 34 (4), p.1797-1855.

Bryson, A., Dorsett, R. & Purdon, S., 2002. The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies. Policy Studies Institute and

National Centre for Social Research, Working Paper No. 4, Department for Work and Pensions.

Burgess, R. & Pande, R., 2003. Do Rural Banks Matter? Evidence from the Indian Social Banking Experiment. Discussion Paper No. DEDPS/40, London School of Economics and Political Science, August.

Burgess, R. & Pande, R., 2005. Do Rural Banks Matter? Evidence from the Indian Social Banking Experiment. *The American Economic Review*, 95 (3), p.780-795.

Burtless, G., 1995. The Case for Randomized Field Trials in Economic and Policy Research. *The Journal of Economic Perspectives,* 9 (2), p.63-84.

Caliendo, M., 2006. *Microeconometric Evaluation of Labour Market Policies.* Berlin: Springer.

Caliendo, M. & Hujer, R., 2005. The Microeconometric Estimation of Treatment Effects - An Overview. Forschungsinstitut zur Zukunft der Arbeit (IZA) Discussion Paper No. 1653, July.

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

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

Campbell, D. T., 1969. Reforms as Experiments. *American Psychologist*, 24, p.409-429.

Cassar, A., Crowley, L. & Wydick, B., 2007. The Effect of Social Capital on Group Loan Repayment: Evidence from Field Experiments. *The Economic Journal,* 117, p.F85-F106.

Census of India, 2001. *District Census Handbook, Part XII A&B.* Ahmedabad: Government of India.

Chakrabarti, R., 2004. The Indian Microfinance Experience - Accomplishments and Challenges. *Available at: http://ssrn.com/abstract=649854.*

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

Chen, M. A. & Snodgrass, D., 1999. An Assessment of the Impact of SEWA Bank in India: Baseline Findings. Report submitted to USAID Assessing the Impact of Microenterprise Services (AIMS), August.

Chen, M. A. & Snodgrass, D., 2001. Managing Resources, Activities, and Risk in Urban India: The Impact of SEWA Bank. Report submitted to USAID Assessing the Impact of Microenterprise Services (AIMS), September.

Chowdhury, I. R., 2010. Understanding the Grameen Miracle: Information and Organisational Innovation. *Economic and Political Weekly,* 45 (6), p.66-73.

Coleman, B. E., 1999. The Impact of Group Lending in Northeast Thailand. *Journal of Development Economics,* 60 (1), p.105-141.

Coleman, B. E., 2006. Microfinance in Northeast Thailand: Who Benefits and How Much? *World Development,* 34 (9), p.1612-1638.

Coleman, J. S., 1988. Social Capital in the Creation of Human Capital. *The American Journal of Sociology,* 94 (Supplement), p.S95-S120.

Collins, D., Morduch, J., Rutherford, S. & Ruthven, O., 2009. *Portfolios of the Poor: How the World's Poor Live on $2 a Day.* Princeton: Princeton University Press.

Cook, T. D. & Wong, V. C., 2008. Empirical Tests of the Validity of the Regression Discontinuity Design. *Available at: http://www.northwestern.edu/ipr/publications/papers/cook_empirical_tests.pdf.*

Cornfield, J., Haenszel, W., Hammond, E. & Lilienfeld, A., 1959. Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *Journal of the National Cancer Institute,* 22, p.173-203.

Cox, D. R., 1958. *Planning of Experiments.* New York: Wiley.

Cull, R., Demirguc-Kunt, A. & Morduch, J., 2009. Microfinance Meets the Market. *Journal of Economic Perspectives,* 23 (1), p.167-192.

Dawid, A. P., 1979. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological),* 41 (1), p.1-31.

Deaton, A., 2009. Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. *Available at: http://www.princeton.edu/~deaton/downloads/Instruments_of_Development.pdf.*

Dehejia, R., 2005. Practical Propensity Score Matching: A Reply to Smith and Todd. *Journal of Econometrics,* 125, p.355-364.

Dehejia, R. & Wahba, S., 2002. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economic Studies,* 84 (1), p.151-161.

Dehejia, R. H. & Wahba, S., 1999. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association,* 94 (448), p.1053-1062.

Deininger, K. & Liu, Y., 2009. Economic and Social Impacts of Self-Help Groups in India. World Bank Policy Research Working Paper No. 4884, March.

Devaney, P. L., 2006. Microsavings Programs: Assessing Demand and Impact, A Critical Review of the Literature. Assessing the Impact of Innovation Grants in Financial Services, IRIS Center, June.

Dichter, T. & Harper, M. eds., 2007. *What's Wrong with Microfinance?* Warwickshire: Practical Action Publishing.

Dieckmann, R., 2007. Microfinance: An Emerging Investment Opportunity: Uniting Social Investment and Financial Returns. Report completed for Deutsche Bank Research, December.

DiPrete, T. A. & 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), p.271-310.

Dreze, J., 1990. Poverty in India and the IRDP Delusion. *Economic and Political Weekly,* 25 (39), p.A95-A104.

Duflo, E., Glennerster, R. & Kremer, M., 2008. Using Randomization in Development Economics Research: A Toolkit. In Schultz, T. P. & Strauss, J., eds. *Handbook of Development Economics, Volume 4.* Amsterdam: Elsevier.

Duflo, E. & Kremer, M., 2005. Use of Randomization in the Evaluation of Development Effectiveness. In Pitman, G. K., Feinstein, O. N. & Ingram, G. K., eds. *Evaluating Development Effectiveness.* New Brunswick: Transaction Publishers.

Dupas, P. & Robinson, J., 2009. Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya. NBER Working Paper No. w14693.

Ellis, F., 1992. *Agricultural Policies in Developing Countries.* Cambridge: Cambridge University Press.

Feder, G., Lau, L. J., Lin, J. Y. & Luo, X., 1990. The Relationship between Credit and Productivity in Chinese Agriculture: A Microeconomic Model of Disequilibrium. *American Journal of Agricultural Economics,* 72 (5), p.1151-1157.

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), p.150-177.

Fischer, G., 2010. Contract Structure, Risk Sharing, and Investment Choice. *Available at: http://personal.lse.ac.uk/fischerg/Research.htm.*

Fisher, R. A., 1935. *The Design of Experiments.* London: Oliver and Boyd.

Fisher, T. & Sriram, M. S., 2002. *Beyond Micro-Credit: Putting Development Back into Micro-Finance.* New Delhi: Vistaar Publications.

Fraker, T. & Maynard, R., 1987. The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs. *The Journal of Human Resources,* 22 (2), p.194-227.

Friedlander, D. & Robins, P. K., 1995. Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods. *The American Economic Review,* 85 (4), p.923-937.

Gaile, G. L. & 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.

Galasso, E. & Ravallion, M., 2004. Social Protection in a Crisis: Argentina's Plan Jefes y Jefas. *World Bank Econ Rev,* 18 (3), p.367-399.

Gangopadhyay, S., Ghatak, M. & Lensink, R., 2005. Joint Liability Lending and the Peer Selection Effect. *The Economic Journal,* 115 (506), p.1005-1015.

Geertz, C., 1973. Thick Description: Toward an Interpretive Theory of Culture. In Geertz, C., ed. *The Interpretation of Cultures: Selected Essays.* New York: Basic Books.

Gertler, P., 2004. Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment. *The American Economic Review,* 94 (2), p.336-341.

Gertler, P. & Boyce, S., 2001. An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico. Unpublished mimeo.

Ghatak, M., 1999. Group Lending, Local Information and Peer Selection. *Journal of Development Economics,* 60 (1), p.27-50.

Ghatak, M., 2000. Screening by the Company You Keep: Joint Liability Lending and the Peer Selection Effect. *The Economic Journal,* 110 (465), p.601-631.

Ghatak, M. & Guinnane, T. W., 1999. The Economics of Lending with Joint Liability: Theory and Practice. *Journal of Development Economics,* 60 (1), p.195-228.

Ghate, P., 2007. Consumer Protection in Indian Microfinance: Lessons from Andhra Pradesh and the Microfinance Bill. *Economic and Political Weekly,* 42 (13), p.1176-1184.

Gine, X. & Karlan, D. S., 2007. Group versus Individual Liability: A Field Experiment in the Philippines. *Available at: http://134.245.95.50:8080/dspace/bitstream/10419/26981/1/593239520.PDF.*

Gine, X. & Karlan, D. S., 2009. Group versus Individual Liability: Long Term Evidence from Philippine Microcredit Lending Groups. *Available at: http://www.econ.yale.edu/growth_pdf/cdp970.pdf.*

Goetz, A. M. & Sen Gupta, R., 1996. Who Takes the Credit? Gender, Power, and Control Over Loan Use in Rural Credit Programs in Bangladesh. *World Development,* 24 (1), p.45-63.

Goldacre, B., 2008. *Bad Science.* London: Fourth Estate.

Goldberg, N., 2005. Measuring the Impact of Microfinance: Taking Stock of What We Know. Grameen Foundation USA Publication Series, December.

Green, D. P., Leong, T. Y., Kern, H. L., Gerber, A. S. & Larimer, C. W., 2009. Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks. *Political Analysis,* 17 (4), p.400-417.

Grootaert, C. & Bastelaer, T. v. eds., 2002. *Understanding and Measuring Social Capital: A Multidisciplinary Tool for Practitioners.* Washington D.C.: The World Bank.

Gu, X. S. & Rosenbaum, P. R., 1993. Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms. *Journal of Computational and Graphical Statistics,* 2, p.405-420.

Gujarati, D. N., 2003. *Basic Econometrics.* New York: McGraw-Hill.

Hahn, J., Todd, P. & van der Klaauw, W., 1999. Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design. NBER Working Paper No. 7131.

Hahn, J., Todd, P. & van der Klaauw, W., 2001. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica,* 69 (1), p.201-209.

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

Harper, M., 2002. Self-help Groups and Grameen Bank Groups: What are the Differences? In Fisher, T. & Sriram, M. S., eds. *Beyond Micro-credit: Putting Development Back into Micro-finance.* New Delhi: Vistaar Publications.

Harriss, J. & de Renzio, P., 1997. "Missing Link" or Analytically Missing? The Concept of Social Capital. *Journal of International Development,* 9 (7), p.919-937.

Hartmann, B. & Boyce, J. K., 1983. *A Quiet Violence: View from a Bangladesh Village.* London: Zed Press.

Hashemi, S. M., Schuler, S. R. & Riley, A. P., 1996. Rural Credit Programs and Women's Empowerment in Bangladesh. *World Development*, 24 (4), p.635-653.

Heckman, J. J., 1974. Shadow Prices, Market Wages, and Labor Supply. *Econometrica,* 42 (4), p.679-694.

Heckman, J. J., 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models. *Annals of Social and Economic Measurement,* 5 (4), p.475-492.

Heckman, J. J., 1978. Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica,* 46 (4), p.931-959.

Heckman, J. J., 1979. Sample Selection Bias as a Specification Error. *Econometrica,* 47 (1), p.153-161.

Heckman, J. J., 1997. Instrumental Variables: A Study of Implicit Behavioral Assumptions Used In Making Program Evaluations. *Journal of Human Resources,* 32 (3), p.441-462.

Heckman, J. J., Ichimura, H., Smith, J. & Todd, P., 1998. Characterizing Selection Bias Using Experimental Data. *Econometrica,* 66 (5), p.1017-1098.

Heckman, J. J., Ichimura, H. & Todd, P., 1997. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies,* 64, p.605-654.

Heckman, J. J., Ichimura, H. & Todd, P., 1998. Matching as an Econometric Evaluation Estimator. *The Review of Economic Studies,* 65 (2), p.261-294.

Heckman, J. J., LaLonde, R. & Smith, J., 1999. The Economics and Econometrics of Active Labor Market Programs. In Ashenfelter, O. & Card, D., eds. *Handbook of Labor Economics, Volume 3A.* Amsterdam: Elsevier.

Heckman, J. J., Lochner, L. & Taber, C., 1999. Human Capital Formation and General Equilibrium Treatment Effects: A Study of Tax and Tuition Policy. *Fiscal Studies,* 20 (1), p.25-40.

Heckman, J. J. & Robb Jr., R., 1985. Alternative Methods for Evaluating the Impact of Interventions. An Overview. *Journal of Econometrics,* 3, p.239-267.

Heckman, J. J. & Smith, J. A., 1995. Assessing the Case for Social Experiments. *The Journal of Economic Perspectives,* 9 (2), p.85-110.

Heckman, J. J. & Urzua, S., 2009. Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. NBER Working Paper No. 14706.

Heckman, J. J. & Vytlacil, E., 1999. Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences,* 96, p.4730-4734.

Heckman, J. J. & Vytlacil, E., 2000. The Relationship Between Treatment Parameters Within a Latent Variable Framework. *Economics Letters,* 66 (1), p.33-39.

Heckman, J. J. & Vytlacil, E., 2001. Policy-Relevant Treatment Effects. *The American Economic Review,* 91 (2), p.107-111.

Heckman, J. J. & Vytlacil, E., 2005. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica,* 73 (3), p.669-738.

Heckman, J. J. & Vytlacil, E., 2007a. Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. In Heckman, J. J. & Leamer, E. E., eds. *Handbook of Econometrics, Volume 6B.* Amsterdam: North-Holland.

Heckman, J. J. & Vytlacil, E., 2007b. Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments. In Heckman, J. J. & Leamer, E. E., eds. *Handbook of Econometrics, Volume 6B.* Amsterdam: North-Holland.

Hermes, N. & Lensink, R., 2007. The Empirics of Microfinance: What Do We Know? *The Economic Journal,* 117 (517), p.F1-F10.

Hirano, K. & Imbens, G. W., 2004. The Propensity Score with Continuous Treatments. In Gelman, A. & Meng, X. L., eds. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives.* New York: Wiley.

Hoddinott, J. & Skoufias, E., 2004. The Impact of PROGRESA on Food Consumption. *Economic Development and Cultural Change,* 53 (1), p.37-61.

Hoff, K. & Stiglitz, J. E., 1990. Introduction: Imperfect Information and Rural Credit Markets - Puzzles and Policy Perspectives. *The World Bank Economic Review,* 4 (3), p.235-250.

Holland, P. W., 1986. Statistics and Causal Inference. *Journal of the American Statistical Association,* 81 (396), p.945-960.

Holmes, J., Isham, J. & Wasilewski, J., 2005. Overcoming Information Asymmetries in Low-Income Lending: Lessons from the "Working Wheels" Program. *Southern Economic Journal,* 72 (2), p.329-351.

Hossain, M., 1988. Credit for Alleviation of Rural Poverty: The Grameen Bank in Bangladesh. IFPRI, Research Report 65, February.

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

Hulme, D. & Mosley, P., 1996. *Finance against Poverty.* London: Routledge.

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.

Imai, K. S., Arun, T. & Annim, S. K., 2010. Microfinance and Household Poverty Reduction: New Evidence from India. *World Development,* 38 (12), p.1760-1774.

Imai, K. S. & Azam, M. S., 2010. Does Microfinance Reduce Poverty in Bangladesh? New Evidence from Household Panel Data. Discussion Paper, DP2010-24, Kobe University, September.

Imbens, G., 2009. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). NBER Working Paper No. 14896.

Imbens, G. & Wooldridge, J., 2008. Recent Developments in the Econometrics of Program Evaluation. The Institute for Fiscal Studies, Department of Economics, University College London, Cemmap Working Paper No. CWP 24/08.

Imbens, G. W., 2000. The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika,* 87 (3), p.706-710.

Imbens, G. W. & Angrist, J. D., 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica,* 62 (2), p.467-475.

Imbens, G. W. & Lemieux, T., 2008. Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics,* 142 (2), p.615-635.

Ito, S., 2003. Microfinance and Social Capital: Does Social Capital Help Create Good Practice? *Development in Practice,* 13 (4), p.322-332.

Iversen, V., Jackson, C., Kebede, B., Munro, A. & Verschoor, A., 2010. Do Spouses Realise Cooperative Gains? Experimental Evidence from Rural Uganda. *Available at: http://www.uea.ac.uk/ssf/cbess/working_papers/Iversen,+Jackson,+Kebede,+Munro+and+Verschoor+(2010).*

Johnson, S., 1998. Programme Impact Analysis in Micro-Finance: The Need for Analysis of Real Markets. *IDS Bulletin,* 29 (4), p.21-31.

Johnson, S. & Rogaly, B., 1997. *Microfinance and Poverty Reduction.* Oxford: Oxfam.

Kagel, J. H. & Roth, A. E. eds., 1995. *The Handbook of Experimental Economics.* Princeton: Princeton University Press.

Karlan, D. S., 2001. Microfinance Impact Assessments: The Perils of Using New Members as a Control Group. *Journal of Microfinance,* 3 (2), p.75-85.

Karlan, D. S., 2007. Social Connections and Group Banking. *The Economic Journal,* 117 (517), p.F52-F84.

Karlan, D. & Goldberg, N., 2006. The Impact of Microfinance: A Review of Methodological Issues. Unpublished mimeo.

Karlan, D. S. & Morduch, J., 2009. Access to Finance. *Available at: http://karlan.yale.edu/p/HDE_June_11_2009_Access_to_Finance.pdf.*

Karlan, D. S. & Zinman, J., 2009. Expanding Microenterprise Credit Access: Using Randomized Supply Decisions to Estimate the Impacts in Manila. *Available at: http://karlan.yale.edu/p/expandingaccess_manila_jul09.pdf.*

Keynes, J. M., 1936. *The General Theory of Employment, Interest and Money.* London: Macmillan.

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.

Khandker, S. R., 2000. Savings, Informal Borrowing and Microfinance. *Bangladesh Development Studies,* 26 (2 & 3).

Khandker, S. R., 2003. Micro-finance and Poverty: Evidence Using Panel Data from Bangladesh. World Bank Policy Research Working Paper No. 2945, January.

Khandker, S. R., 2005. Microfinance and Poverty: Evidence Using Panel Data from Bangladesh. *The World Bank Economic Review,* 19 (2), p.263-286.

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

LaLonde, R. J., 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review,* 76 (4), p.604-620.

Lapenu, C. & Zeller, M., 2001. Distribution, Growth, and Performance of Microfinance Institutions in Africa, Asia, and Latin America. Food Consumption and Nutrition Division, Discussion Paper No. 114, IFPRI, June.

Leamer, E. E., 1983. Let's Take the Con Out of Econometrics. *The American Economic Review,* 73 (1), p.31-43.

Lechner, M., 2000. A Note on the Common Support Problem in Applied Evaluation Studies. *Available at: http://www.siaw.unisg.ch/org/siaw/webold.nsf/SysWebRessources/ML_2000_L_09/$FILE/00l09.pdf.*

Lechner, M., 2001. Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption. In Lechner, M. & Pfeiffer, F., eds. *Econometric Evaluations of Active Labor Market Policies in Europe.* Heidelberg: Physica.

Ledgerwood, J., 1999. *Microfinance Handbook: An Institutional and Financial Perspective.* Washington D.C.: The World Bank.

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.*

Levitt, S. D. & List, J. A., 2009. Was there Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. NBER Working Paper No. 15016.

Lewis, D. J., 1991. *Technologies And Transactions: A Study of the Interaction between New Technology and Agrarian Structure in Bangladesh.* Dhaka: Centre for Social Studies.

Madajewicz, M., 2004. Joint Liability versus Individual Liability in Credit Contracts. *Available at: http://www.econ.columbia.edu/RePEc/pdf/DP0304-18.pdf.*

Makina, D. & Malobola, L. M., 2004. Impact Assessment of Microfinance Programmes Including Lessons from Khula Enterprise Finance. *Development Southern Africa,* 21 (5), p.799-814.

Mannan, M., 2002. Bangsha: Islam, History and the Structure of Bengali Muslim Descent. In Nurul Alam, S. M., ed. *Contemporary Anthropology. Theory and Practice.* Dhaka: The University Press Limited.

Manski, C. F., 1995. *Identification Problems in the Social Sciences.* Cambridge: Harvard University Press.

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), p.93-115.

Meier, P., 1972. The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine. In Tanur, J., ed. *Statistics: A Guide to the Unknown.* San Francisco: Holden-Day.

Menon, N., 2006. Non-linearities in Returns to Participation in Grameen Bank Programs. *Journal of Development Studies,* 42 (8), p.1379 - 1400.

Miguel, E. & Kremer, M., 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica,* 72 (1), p.159-217.

Misra, B. S., 2010. *Credit Cooperatives in India: Past, Present and Future.* London: Routledge.

Molyneux, M., 2002. Gender and the Silences of Social Capital: Lessons from Latin America. *Development and Change,* 33 (2), p.167-188.

Montgomery, H., 2005. Serving the Poorest of the Poor: The Poverty Impact of the Khushhali Bank's Microfinance Lending in Pakistan. *Poverty Reduction Strategies in Asia: Asian Development Bank Institute (ADBI) Annual Conference.* Tokyo, 9 December 2005.

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,* XXXVII December, p.1569-1614.

Morduch, J. & Haley, B., 2002. Analysis of the Effects of Microfinance on Poverty Reduction. NYU Wagner Working Paper No. 1014, June.

Morgan, S. L. & Harding, D. J., 2006. Matching Estimators of Causal Effects. Prospects and Pitfalls in Theory and Practice. *Sociological Methods & Research,* 35 (1), p.3-60.

Morgan, S. L. & Winship, C., 2007. *Counterfactuals and Causal Inference. Methods and Principles for Social Research.* Cambridge: Cambridge University Press.

Mosley, P., 1996. Metamorphosis from NGO to Commercial Bank: The Case of BancoSol in Bolivia. In Hulme, D. & Mosley, P., eds. *Finance against Poverty.* London: Routledge.

Nair, T. S., 2005. The Transforming World of Indian Microfinance. *Economic and Political Weekly,* 40 (17), p.1695-1698.

Nair, T. S., 2006. Financial Intermediation for Rural Development: Exploring the Role and Contribution of Commercial Banks. In Indian Institute of Banking and Finance, ed. *Readings on Financial Inclusion.* New Delhi: Taxmann Publications.

Nannicini, T., 2007. Simulation-based Sensitivity Analysis for Matching Estimators. *The STATA Journal,* 7 (3), p.334-350.

Neyman, J. & Pearson, E. S., 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London,* 231, p.289-337.

Neyman, J. S., 1923. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Translated in Statistical Science,* 5 (4), p.465-480, 1990.

North, D. C., 1990. *Institutions, Institutional Change and Economic Performance.* Cambridge: Cambridge University Press.

Odell, K., 2010. Measuring the Impact of Microfinance: Taking Another Look. Grameen Foundation USA Publication Series, May.

Olson, M., 1982. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities.* New Haven: Yale University Press.

Pathak, B. V., 2003. *Indian Financial System.* New Delhi: Pearson Education.

Pearl, J., 2000. *Causality: Models, Reasoning, and Inference.* Cambridge: Cambridge University Press.

Pitt, M. M., Khandker, S. R. & Cartwright, J., 2006. Empowering Women with Micro-finance: Evidence from Bangladesh. *Economic Development and Cultural Change,* p.791-831.

Pitt, M. M., Khandker, S. R., Chowdhury, O. H. & Millimet, D. L., 2003. Credit Programmes for the Poor and the Health Status of Children in Rural Bangladesh. *International Economic Review,* 44 (1), p.87-118.

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), p.15-48.

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), p.958-996.

Pitt, M. M. & Khandker, S. R., 2002. Credit Programmes for the Poor and Seasonality in Rural Bangladesh. *Journal of Development Studies,* 39 (2), p.1-24.

Pitt, M. M., Khandker, S. R., McKernan, S.-M. & 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), p.1-21.

Pritchett, L., 2002. It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation. *Journal of Economic Policy Reform,* 5 (4), p.251 - 269.

Pritchett, L., 2009. The Policy Irrelevance of the Economics of Education: Is "Normative as Positive" Just Useless, or Worse? In Cohen, J. & Easterly, W., eds. *What Works in Development? Thinking Big and Thinking Small.* Washington D.C.: Brookings Institution Press.

Puhani, P. A., 2000. The Heckman Correction for Sample Selection and its Critique. *Journal of Economic Surveys,* 14 (1), p.53-68.

Putnam, R. D., 1993. *Making Democracy Work: Civic Traditions in Modern Italy.* Princeton: Princeton University Press.

Ramachandran, V. K. & Swaminathan, M. eds., 2005. *Financial Liberalization and Rural Credit in India.* New Delhi: Tulika Books.

Ravallion, M., 2001. The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation. *The World Bank Economic Review,* 15 (1), p.115-140.

Ravallion, M., 2008. Evaluating Anti-Poverty Programs. In Schultz, T. P. & Strauss, J., eds. *Handbook of Development Economics, Volume 4.* Amsterdam: Elsevier.

Rogg, C. S., 2000. The Impact of Access to Credit on the Saving Behavior of Microentrepreneurs: Evidence from 3 Latin American Countries. *Available at: http://idbdocs.iadb.org/wsdocs/getdocument.aspx?docnum=1481486.*

Roodman, D., 2006. How to Do xtabond2: An Introduction to "Difference" and "System" GMM in Stata. Center for Global Development, Working Paper No. 103, December.

Roodman, D., 2009. Estimating Fully Observed Recursive Mixed-Process Models with cmp. Center for Global Development, Working Paper No. 168, April.

Roodman, D. & 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., 1987. Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies. *Biometrika,* 74 (1), p.13-26.

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

Rosenbaum, P. R., 2010. *Design of Observational Studies.* New York: Springer.

Rosenbaum, P. R. & Rubin, D. B., 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika,* 70 (1), p.41-55.

Rosenbaum, P. R. & Rubin, D. B., 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association,* 79 (387), p.516-524.

Rosenbaum, P. R. & Silber, J. H., 2001. Matching and Thick Description in an Observational Study of Mortality After Surgery. *Biostatistics,* 2 (2), p.217-232.

Rosenzweig, M. R., 2001. Savings Behaviour in Low-Income Countries. *Oxford Review of Economic Policy,* 17 (1), p.40-54.

Rubin, D. B., 1973a. Matching to Remove Bias in Observational Studies. *Biometrics,* 29 (1), p.159-183.

Rubin, D. B., 1973b. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics,* 29 (1), p.185-203.

Rubin, D. B., 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology,* 66 (5), p.688-701.

Rubin, D. B., 1977. Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics,* 2 (1), p.1-26.

Rubin, D. B., 1978. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics,* 6 (1), p.34-58.

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

Saretsky, G., 1975. The John Henry Effect: Potential Confounder of Experimental vs Control Group Approaches to the Evaluation of Educational Innovations. *The American Educational Research Association's Annual Meeting.* Washington, D.C., 2 April 1975.

Schuler, S. R. & Hashemi, S. M., 1994. Credit Programs, Women's Empowerment, and Contraceptive Use in Rural Bangladesh. *Studies in Family Planning,* 25 (2), p.65-76.

Scriven, M., 2008. A Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research. *Journal of MultiDisciplinary Evaluation,* 5 (9), p.11-24.

Sebstad, J. & Chen, G., 1996. Overview of Studies on the Impact of Microenterprise Credit. Report submitted to USAID Assessing the Impact of Microenterprise Services (AIMS), June.

Sebstad, J., Neill, C., Barnes, C. & Chen, G., 1995. Assessing the Impacts of Microenterprise Interventions: A Framework for Analysis. Center for Development Information and Evaluation, Working Paper No. 7, USAID, March.

Setboonsarng, S. & Parpiev, Z., 2008. Microfinance and the Millennium Development Goals in Pakistan: Impact Assessment Using Propensity Score Matching. Asian Development Bank Institute (ADBI) Discussion Paper No. 104, March.

Shadish, W. R., Cook, T. D. & Campbell, D. T., 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin Company.

Shah, M., Rao, R. & Shankar, P. S. V., 2007. Rural Credit in 20th Century India: Overview of History and Perspectives. *Economic and Political Weekly,* 42 (15), p.1351-1364.

Sinha, S., 2000. India. In Asian Development Bank, ed. *The Role of Central Banks in Microfinance in Asia and the Pacific: Vol. 2, Country Studies.* Manila: Asian Development Bank.

Smith, J. A. & Todd, P., 2005. Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics,* 125, p.305-353.

Snodgrass, D. & Sebstad, J., 2002. Clients in Context: The Impacts of Microfinance in Three Countries: Synthesis Report. Report submitted to USAID Assessing the Impact of Microenterprise Services (AIMS), January.

Sriram, M. S., 2005. Microfinance and the State: Exploring Areas and Structures of Collaboration. *Economic and Political Weekly,* 40 (17), p.1699-1704.

Stiglitz, J. E., 1990. Peer Monitoring and Credit Markets. *World Bank Economic Review,* 4 (3), p.351-366.

Stiglitz, J. E. & Weiss, A., 1981. Credit Rationing in Markets with Imperfect Information. *The American Economic Review,* 71 (3), p.393-410.

Street, D. J. & Burgess, L., 2007. *The Construction of Optimal Stated Choice Experiments, Theory and Methods.* Hoboken: Wiley.

Takahashi, K., Higashikata, T. & Tsukada, K., 2010. The Short-Term Poverty Impact of Small-Scale, Collateral-Free Microcredit in Indonesia: A Matching Estimator Approach. *The Developing Economies,* 48 (1), p.128-155.

Tedeschi, G. A., 2008. Overcoming Selection Bias in Microcredit Impact Assessments: A Case Study in Peru. *Journal of Development Studies,* 44 (4), p.504-518.

Thistlethwaite, D. & Campbell, D., 1960. Regression-Discontinuity Analysis; An Alternative to the Ex Post Facto Experiment. *Journal of Educational Psychology,* 51 (6), p.309-317.

Todd, H., 1996. *Women at the Center: Grameen Bank Borrowers After One Decade.* Boulder: Westview Press.

van der Klaauw, W., 2002. Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach. *International Economic Review,* 43 (4), p.1249-1287.

Varian, H. R., 1990. Monitoring Agents With Other Agents. *Journal of Institutional and Theoretical Economics,* 146 (2), p.153-174.

Venkata, N. A. & Yamini, V., 2010. Why Do Microfinance Clients Take Multiple Loans? MicroSave India Focus Note 33, February.

von Pischke, J. D., 1983. Towards an Operational Approach to Savings for Rural Developers. In Von Pischke, J. D., Adams, D. W. & Donald, G., eds. *Rural Financial Markets in Developing Countries.* Baltimore: Johns Hopkins University Press.

von Pischke, J. D., 1991. *Finance at the Frontier: Debt Capacity and the Role of Credit in the Private Economy.* Washington D.C.: The World Bank.

von Pischke, J. D., Adams, D. W. & Donald, G., 1983. *Rural Financial Markets in Developing Countries.* Baltimore: Johns Hopkins University Press.

Vytlacil, E., 2002. Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica,* 70 (1), p.331-341.

White, H., 2009. Some Reflections on Current Debates in Impact Evaluation. 3ie Working Paper No. 1.

World Bank, 1998. The Initiative on Defining, Monitoring and Measuring Social Capital: Overview and Program Description. Social Capital Initiative, Working Paper No. 1, The World Bank, April.

Wydick, B., 2001. Group Lending under Dynamic Incentives as a Borrower Discipline Device. *Review of Development Economics,* 5 (3), p.406-420.

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

Yunus, M., 2007. Remarks by Muhammad Yunus, Managing Director, Grameen Bank. *Microcredit Summit E-News,* 5 (1).

Zhao, Z., 2003. Data Issues of Using Matching Methods to Estimate Treatment Effects: An Illustration with NSW Data Set. Working Paper, China Center for Economic Research (CCER), Peking University, July.

Appendix

Appendix 1: Descriptive statistics for all outcome variables across round 1 - 3

Household level hypotheses	**All respondents**			**Borrowers**			**Savers**			**Control**		
Data collection round	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**
Sample size	**786**	**786**	**210**	**264**	**264**	**70**	**260**	**260**	**70**	**262**	**262**	**70**
Household level hypotheses												
Total household income per annum in Rupees	42,582	48,487	51,804	51,417	59,715	51,657	40,401	47,388	51,966	35,845	38,264	51,789
Total household income per annum per capita in Rupees	7,878	8,827	12,363	9,268	10,535	11,673	7,827	8,911	12,767	6,526	7,023	12,648
Total household income per annum per capita in Rupees - equivalence scale adjusted	12,871	14,447	18,389	15,239	17,346	17,728	12,694	14,454	18,835	10,661	11,519	18,604
Inverse Simpson index	1.810	3.067	0.010	1.767	2.914	0.014	1.817	3.340	0.000	1.847	2.953	0.014
Expenditure for housing improvements in Rupees	4,566	8,596	1,149	7,391	12,671	2,578	4,291	8,769	-198	1,992	4,317	1,066
Expenditure on household assets in Rupees	1,671	2,602	11,433	2,475	2,617	13,993	1,363	3,088	9,853	1,167	2,106	10,452
School enrolment for girls aged 5 to 10 years	0.327	0.312	0.143	0.322	0.284	0.271	0.312	0.338	0.071	0.347	0.313	0.086
School enrolment for boys aged 5 to 10 years	0.422	0.385	0.219	0.417	0.360	0.286	0.423	0.404	0.200	0.427	0.393	0.171
School enrolment for girls aged 11 to 17 years	0.313	0.303	0.171	0.311	0.284	0.186	0.315	0.319	0.157	0.313	0.305	0.171
School enrolment for boys aged 11 to 17 years	0.342	0.388	0.133	0.352	0.352	0.200	0.338	0.415	0.057	0.336	0.397	0.143
Food expenditure per day per capita in Rupees	11.66	12.53		13.04	13.06		11.40	12.44		10.54	12.10	

Enterprise and individual level hypotheses	All respondents			Borrowers			Savers			Control		
Data collection round	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**
Sample size	**786**	**786**	**210**	**264**	**264**	**70**	**260**	**260**	**70**	**262**	**262**	**70**
Enterprise level hypotheses												
Informal sector income of whole household - per month in Rupees	5,649	7,400		6,758	10,793		5,275	7,599		4,903	3,784	
Informal sector income of respondent only - per month in Rupees	3,324	2,683		4,558	3,238		3,204	3,482		2,199	1,330	
Microenterprise revenues of all enterprises in household - per month in Rupees	4,911	6,446		6,013	9,331		4,698	6,657		4,012	3,329	
Microenterprise revenues of microenterprises for which respondent is primarily responsible - per month in Rupees	2,816	2,218		4,047	2,903		2,470	2,594		1,920	1,155	
Current value of fixed assets of all microenterprises in household in Rupees	1,639	599		1,924	1,027		1,294	522		1,693	243	
Current value of fixed assets of microenterprises for which respondent is primarily responsible in Rupees	857	138		1,285	334		741	67		542	12	
Hours worked in previous week in all microenterprises in household	35.51	45.73		40.67	60.06		36.37	44.39		29.45	32.63	
Days worked in previous month in all microenterprises in household	24.97	27.53		29.92	36.09		25.03	26.92		19.90	19.52	
Main types of suppliers - inferior suppliers? Yes=1, No=0	0.363	0.391		0.360	0.436		0.385	0.381		0.344	0.355	
Main types of customers - inferior customers? Yes=1, No=0	0.482	0.469		0.553	0.557		0.454	0.446		0.439	0.405	
Individual level hypotheses												
Respect by other household members? Yes=1, No=0	0.926	0.955		0.943	0.970		0.927	0.958		0.908	0.939	
Prepared to deal with future? Yes=1, No=0	0.882	0.817		0.905	0.852		0.877	0.800		0.863	0.798	

Appendix 2: Cross-section data results: Detailed household, enterprise and individual level hypotheses - microfinance participants versus controls – without sampling weights

Household level hypotheses	Round 1	Round 2	Round 3
Total household income per annum in Rupees			
USAID	10,090***	15,302***	N/A
PSM - 5 nearest neighbour matching	8,944***	14,635***	7,030
PSM - kernel matching, bandwidth 0.01	8,638***	13,786***	9,355*
Total household income per annum per capita in Rupees			
USAID	2,063***	2,685***	N/A
PSM - 5 nearest neighbour matching	2,019***	2,486***	1,805
PSM - kernel matching, bandwidth 0.01	1,913***	2,537***	2,222
Inverse Simpson index			
USAID	0.11***	0.025	N/A
PSM - 5 nearest neighbour matching	0.022	0.195	N/A
PSM - kernel matching, bandwidth 0.01	0.013	0.252	N/A
Expenditure for housing improvements in Rupees			
USAID	3,748***	5,871	N/A
PSM - 5 nearest neighbour matching	3,701***	6,546***	1,150
PSM - kernel matching, bandwidth 0.01	3,484***	6,504***	1,191
Expenditure on household assets in Rupees			
USAID	752***	545	N/A
PSM - 5 nearest neighbour matching	606	799*	45
PSM - kernel matching, bandwidth 0.01	415	463	414
School enrolment for girls aged 5 to 10 years			
USAID	-0.020	-0.005	N/A
PSM - 5 nearest neighbour matching	0.011	0.052	0.029
PSM - kernel matching, bandwidth 0.01	0.010	0.028	-0.012
School enrolment for boys aged 5 to 10 years			
USAID	0.065	0.005	N/A
PSM - 5 nearest neighbour matching	-0.027	0.021	-0.057
PSM - kernel matching, bandwidth 0.01	-0.007	-0.004	-0.042
School enrolment for girls aged 11 to 17 years			
USAID	0.015	-0.015	N/A
PSM - 5 nearest neighbour matching	0.028	0.012	0.014
PSM - kernel matching, bandwidth 0.01	0.006	0.009	-0.031
School enrolment for boys aged 11 to 17 years			
USAID	-0.075	-0.020***	N/A
PSM - 5 nearest neighbour matching	-0.025	-0.012	-0.000
PSM - kernel matching, bandwidth 0.01	-0.045	-0.019	0.000

Household level hypotheses – cont.	Round 1	Round 2	Round 3
Food expenditure per day per capita in Rupees			
USAID	1.30***	1.07	N/A
PSM - 5 nearest neighbour matching	1.35***	1.10	N/A
PSM - kernel matching, bandwidth 0.01	1.34***	1.02	N/A
Mechanisms used for dealing with shocks			
USAID split shock mechanism in stage 1 and stage 2 strategies without explicitly defining each of those strategies. Stage 2 strategies imply a loss of productive assets. 93% of the households in round 1 and 96% of the households in round 2 relied on stage 1 strategies. It was not particularly meaningful to further subject this hypothesis to PSM.			

Enterprise level hypotheses	Round 1	Round 2	Round 3
Informal sector income of whole household - per month in Rupees			
USAID	1,581	4,966	N/A
PSM - 5 nearest neighbour matching	1,848	5,164***	N/A
PSM - kernel matching, bandwidth 0.01	1,528	5,198***	N/A
Informal sector income of respondent only - per month in Rupees			
USAID	1,254***	2,668	N/A
PSM - 5 nearest neighbour matching	2,372**	1,946**	N/A
PSM - kernel matching, bandwidth 0.01	2,212**	1,836**	N/A
Microenterprise revenues of all enterprises in household - per month in Rupees			
USAID	-39.5	5,765	N/A
PSM - 5 nearest neighbour matching	1,930	4,327***	N/A
PSM - kernel matching, bandwidth 0.01	1,810	4,421***	N/A
Microenterprise revenues of microenterprises for which respondent is primarily responsible - per month in Rupees			
USAID	1,007	3,774	N/A
PSM - 5 nearest neighbour matching	1,960**	1,506***	N/A
PSM - kernel matching, bandwidth 0.01	1,820**	1,405***	N/A
Current value of fixed assets of all microenterprises in household in Rupees			
USAID	-1,135	606	N/A
PSM - 5 nearest neighbour matching	100	747*	N/A
PSM - kernel matching, bandwidth 0.01	-144	560	N/A
Current value of fixed assets of microenterprises for which respondent is primarily responsible in Rupees			
USAID	130	354	N/A
PSM - 5 nearest neighbour matching	401	187	N/A
PSM - kernel matching, bandwidth 0.01	439	188	N/A
Hours worked in previous week in all microenterprises in household			
USAID	0.8	14.2***	N/A
PSM - 5 nearest neighbour matching	12.4***	20.3***	N/A
PSM - kernel matching, bandwidth 0.01	11.6**	18.5***	N/A

Enterprise level hypotheses – cont.	Round 1	Round 2	Round 3
Days worked in previous month in all microenterprises in household			
USAID	5.5***	9.0	N/A
PSM - 5 nearest neighbour matching	9.5***	12.7***	N/A
PSM - kernel matching, bandwidth 0.01	9.4***	11.5***	N/A
Main types of suppliers - inferior suppliers? Yes=1, No=0			
USAID	0.062***	0.055	N/A
PSM - 5 nearest neighbour matching	0.063	0.057	N/A
PSM - kernel matching, bandwidth 0.01	0.064	0.050	N/A
Main types of customers - inferior customers? Yes=1, No=0			
USAID	0.024	0.056	N/A
PSM - 5 nearest neighbour matching	0.089*	0.095**	N/A
PSM - kernel matching, bandwidth 0.01	0.088**	0.090**	N/A

Individual level hypotheses	Round 1	Round 2	Round 3
Decision to take last loan, self? Decision to spend last loan, self? Decision to spend income/revenue, self?			
The decision-making questions were highly problematic and only applicable to borrowers who had repaid and borrowed again (N=83). Moreover, there was a high degree of non-response. On average only around 50% of the borrowers replied to those questions. In addition, the degree of non-response was much higher in round 2 than in round 1. The PSM results are not particularly meaningful in this context; hence this hypothesis was not further tested.			
Feelings with regard to contribution to the household?			
The USAID results could not be verified. The following question was asked: 'Do you feel you make an important contribution to the household?' Following this, USAID must have created a dummy variable with 0=negative response, 1=positive response. However, the raw data indicates that, apart from one respondent in round 1 and four respondents in round 2, all others responded that they made a contribution to the household. Hence, based on the raw data it was meaningless to investigate this further since the USAID results could not be replicated in the first place.			
Respect by other household members? Yes=1, No=0			
USAID	0.026	0.025	N/A
PSM - 5 nearest neighbour matching	0.025	0.016	N/A
PSM - kernel matching, bandwidth 0.01	0.025	0.020	N/A
Existence of personal savings?			
This hypothesis is obsolete since all SEWA Bank clients, be it borrowers or just savers are required to build up savings.			
Prepared to deal with future? Yes=1, No=0			
USAID	0.029	0.028	N/A
PSM - 5 nearest neighbour matching	-0.011	0.122	N/A
PSM - kernel matching, bandwidth 0.01	0.005	0.006	N/A

Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%.
The results in this table refer to the differences in the mean values between matched samples; they were obtained using the STATA command psmatch2. I also ran the STATA command pscore with the objective to cross-check the psmatch2 results across the various matching algorithms. The results I obtained from the different STATA routines displayed minor differences in terms of the size of coefficient and the level of significance. The results presented in Appendix 2 do not apply sampling weights. Results are bootstrapped.

Appendix 3: Sub-group comparisons: Detailed household, enterprise and individual level hypotheses, without sampling weights

All sub-group comparisons with control as a base – household level:

	Borrowers vs. control			Savers vs. control			One-time borrower vs. control			Repeat borrower vs. control		
	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2#	R3#
Household level hypotheses												
Total household income per annum in Rupees												
PSM - 5 nearest neighbour	12,323***	17,915***	1,084	7,236**	10,162***	3,767	11,196**	30,099***	171	17,556***	2,319	23,800
PSM - kernel, bandwidth 0.01	12,323***	18,256***	3,806	6,472**	10,085***	4,299	12,212***	27,700***	5,038	15,738***		
Total household income per annum per capita in Rupees												
PSM - 5 nearest neighbour	2,364***	3,222***	885	1,545**	1,899***	311	1,998**	5,500***	642	3,410***	1,112	3,067
PSM - kernel, bandwidth 0.01	2,347***	3,378***	1,459	1,431**	1,909***	758	2,186**	5,669***	1,574	3,203***		
Inverse Simpson index												
PSM - 5 nearest neighbour	0.013	0.252	N/A	-0.017	0.313	N/A	0.021	-0.251	N/A	0.037	0.363	N/A
PSM - kernel, bandwidth 0.01	0.006	0.199	N/A	-0.024	0.343	N/A	0.025	-0.299	N/A	0.034		N/A
Expenditure for housing improvements in Rupees												
PSM - 5 nearest neighbour	5,046***	8,137**	2,652	2,212**	5,044***	-171	5,125***	18,619*	2	6,059***	-825	29,866*
PSM - kernel, bandwidth 0.01	5,069***	8,160**	2,615	1,858*	4,508**	-171	5,107***	18,468*	-2	6,027***		
Expenditure on household assets in Rupees												
PSM - 5 nearest neighbour	1,146***	110	1,135	288	1,041*	41	1,066	1,082	2,584	1,140	1,150	-1,494
PSM - kernel, bandwidth 0.01	1,357**	-37	2,848	296	1,015*	-1,037	1,286*	1,707*	2,569	1,221		
School enrolment for girls aged 5 to 10 years												
PSM - 5 nearest neighbour	0.050	0.024	0.143*	-0.026	0.042	0.021	0.035	0.110	0.175**	-0.020	-0.400	-0.033
PSM - kernel, bandwidth 0.01	0.059	0.025	0.128	-0.001	0.049	0.043	0.064	0.033	0.135	0.022		
School enrolment for boys aged 5 to 10 years												
PSM - 5 nearest neighbour	0.049	-0.019	0.054	-0.011	0.027	0.018	0.007	0.074	0.053	-0.004	-0.200	-0.033
PSM - kernel, bandwidth 0.01	0.014	-0.016	0.019	-0.011	0.018	0.055	-0.024	0.032	0.067	-0.007		
School enrolment for girls aged 11 to 17 years												
PSM - 5 nearest neighbour	0.018	0.044	-0.051	0.013	0.000	0.009	0.020	-0.051	0.035	0.053	0.000	-0.033
PSM - kernel, bandwidth 0.01	0.005	0.000	-0.124	-0.000	0.018	-0.035	-0.012	0.008	-0.001	-0.153		

	Borrowers vs. control			Savers vs. control			One-time borrower vs. control			Repeat borrower vs. control		
	R1	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2#**	**R3#**
Household level hypotheses – cont.												
School enrolment for boys aged 11 to 17 years												
PSM - 5 nearest neighbour	-0.075	-0.005	0.137	-0.006	0.006	-0.158**	-0.036	-0.072	0.126	-0.099	-0.200	0.133
PSM - kernel, bandwidth 0.01	-0.100	-0.036	0.131	-0.026	0.014	-0.197**	-0.045	-0.003	0.141	-0.080		
Food expenditure per day per capita in Rupees												
PSM - 5 nearest neighbour	1.75***	1.63***	N/A	1.02*	0.29	N/A	1.46**	1.36	N/A	2.74**	-0.94	N/A
PSM - kernel, bandwidth 0.01	1.97***	1.50	N/A	0.85	0.55	N/A	1.62***	2.20	N/A	2.53**		N/A
Enterprise level hypotheses												
Informal sector income of whole household - per month in Rupees												
PSM - 5 nearest neighbour	1,394	7,105***	N/A	1,129	3,547*	N/A	2,356	11,036**	N/A	2,378	1,205	N/A
PSM - kernel, bandwidth 0.01	1,463	6,597***	N/A	1,185	3,573*	N/A	3,005	10,515**	N/A	2,034		N/A
Informal sector income of respondent only - per month in Rupees												
PSM - 5 nearest neighbour	2,956***	2,117***	N/A	1,134	1,897	N/A	3,179**	959	N/A	2,219**	204	N/A
PSM - kernel, bandwidth 0.01	3,075***	1,693***	N/A	1,361	2,016	N/A	3,498**	878	N/A	2,373		N/A
Microenterprise revenues of all enterprises in household - per month in Rupees												
PSM - 5 nearest neighbour	1,982	6,062***	N/A	1,484	2,989*	N/A	2,786	8,361*	N/A	2,332	1,205	N/A
PSM - kernel, bandwidth 0.01	2,078	5,566***	N/A	1,456	3,106**	N/A	3,200**	7,695*	N/A	2,231		N/A
Microenterprise revenues of microenterprises for which respondent is primarily responsible - per month in Rupees												
PSM - 5 nearest neighbour	2,688***	1,955***	N/A	721	1,178	N/A	2,799**	978	N/A	2,332**	204	N/A
PSM - kernel, bandwidth 0.01	2,805***	1,556***	N/A	910	1,281	N/A	3,033**	870	N/A	2,491*		N/A
Current value of fixed assets of all microenterprises in household in Rupees												
PSM - 5 nearest neighbour	-143	748	N/A	-262	105	N/A	400	2,737*	N/A	695	-13	N/A
PSM - kernel, bandwidth 0.01	371	854	N/A	-287	222	N/A	668	2,681*	N/A	302		N/A
Current value of fixed assets of microenterprises for which respondent is primarily responsible in Rupees												
PSM - 5 nearest neighbour	490	321	N/A	7.29	53	N/A	595	977	N/A	1,083	0	N/A
PSM - kernel, bandwidth 0.01	701	322	N/A	123	54	N/A	441	956	N/A	1,146		N/A

	Borrowers vs. control			Savers vs. control			One-time borrower vs. control			Repeat borrower vs. control		
	R1	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2#**	**R3#**
Enterprise level hypotheses – cont.												
Hours worked in previous week in all microenterprises in household												
PSM - 5 nearest neighbour	10.41*	29.08***	N/A	8.39	9.70*	N/A	5.37	24.82**	N/A	15.89*	38.90	N/A
PSM - kernel, bandwidth 0.01	10.85*	22.54***	N/A	9.21*	10.07*	N/A	10.43*	24.74**	N/A	17.75*		N/A
Days worked in previous month in all microenterprises in household												
PSM - 5 nearest neighbour	10.02***	17.40***	N/A	6.50**	6.59**	N/A	9.48**	15.84***	N/A	13.96***	12.60	N/A
PSM - kernel, bandwidth 0.01	10.70***	13.56***	N/A	7.42**	6.67**	N/A	10.65***	17.04***	N/A	14.99***		N/A
Main types of suppliers - inferior suppliers? Yes=1, No=0												
PSM - 5 nearest neighbour	0.028	0.088*	N/A	0.064	0.027	N/A	0.067	0.044	N/A	0.110	0.300	N/A
PSM - kernel, bandwidth 0.01	0.050	0.052	N/A	0.070	0.022	N/A	0.052	0.071	N/A	0.093		N/A
Main types of customers - inferior customers? Yes=1, No=0												
PSM - 5 nearest neighbour	0.092*	0.157***	N/A	0.031	0.042	N/A	0.134**	0.144**	N/A	0.12	0.400	N/A
PSM - kernel, bandwidth 0.01	0.097*	0.112**	N/A	0.049	0.031	N/A	0.141**	0.176**	N/A	0.16**		N/A
Individual level hypotheses												
Respect by other household members? Yes=1, No=0												
PSM - 5 nearest neighbour	0.040	0.017	N/A	0.045	0.019	N/A	0.015	0.062***	N/A	0.042	0.000	N/A
PSM - kernel, bandwidth 0.01	0.032	0.017	N/A	0.032	0.017	N/A	0.022	0.045*	N/A	0.030		N/A
Prepared to deal with future? Yes=1, No=0												
PSM - 5 nearest neighbour	0.016	0.072*	N/A	-0.008	-0.028	N/A	0.007	0.121***	N/A	0.050	0.200	N/A
PSM - kernel, bandwidth 0.01	0.026	0.044	N/A	0.002	-0.010	N/A	0.015	0.089*	N/A	0.097*		N/A

Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%.
The results in this table refer to the differences in the mean values between matched samples; they were obtained using the STATA command psmatch2. I also ran the STATA command pscore with the objective to cross-check the psmatch2 results across the various matching algorithms. The results I obtained from the different STATA routines displayed minor differences in terms of the size of coefficient and the level of significance. The results presented in Appendix 3 do not apply sampling weights. # No values for kernel matching in Round 2 and in Round 3, the sample was too small with propensity scores outside the common support region, no adequate matches could be found. Results are bootstrapped.

All sub-group comparisons with saver as a base & one-time versus repeat borrowers– household level:

	Borrowers vs. saver			One-time borrower vs. saver			Repeat borrower vs. saver			One-time vs. repeat borrower		
	R1	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3+**	**R1**	**R2**	**R3+**
Household level hypotheses												
Total household income per annum in Rupees												
PSM - 5 nearest neighbour	9,152***	11,014***	-9,594	5,159	23,843***	-8,865	12,440***	12,778	17,080	-5,313	22,654***	-25,950
PSM - kernel, bandwidth 0.01	9,020***	10,141***	-1,164	7,411*	24,857***	1,472	13,457***	12,667		-6,170	20,001***	
Total household income per annum per capita in Rupees												
PSM - 5 nearest neighbour	1,567**	1,634**	-1,823	1,118	4,025***	-762	2,241**	2,301	-393	-647	4,284***	-4,658
PSM - kernel, bandwidth 0.01	1,405*	1,634**	-192	1,294	3,926***	1,028	2,204**	2,029		-676	4,511***	
Inverse Simpson index												
PSM - 5 nearest neighbour	0.004	-0.102	N/A	0.020	-0.723**	N/A	0.096	0.215	N/A	-0.106	-0.731**	N/A
PSM - kernel, bandwidth 0.01	0.012	-0.192	N/A	-0.006	-0.776***	N/A	0.048	0.501	N/A	-0.056	-0.504	N/A
Expenditure for housing improvements in Rupees												
PSM - 5 nearest neighbour	3,700**	4,547	3,045*	3,663**	14,389	141	4,723***	31,664	30,630*	-3,362	9,417	-35,752**
PSM - kernel, bandwidth 0.01	3,349**	4,115	3,045**	3,442*	14,448	141	4,738**	29,829		-1,328	10,593	
Expenditure on household assets in Rupees												
PSM - 5 nearest neighbour	511	-315	3,817	909	1,350	5,599*	474	233	-1,912	424	1,777**	6,116
PSM - kernel, bandwidth 0.01	722	-371	5,592	940	1,302	5,975	531	-514		-619	1,168	
School enrolment for girls aged 5 to 10 years												
PSM - 5 nearest neighbour	0.002	-0.011	0.165**	0.018	0.023	0.175**	-0.086	-0.033	-0.033	0.058	0.027	0.175
PSM - kernel, bandwidth 0.01	0.014	-0.013	0.030	0.056	0.033	0.018	-0.050	-0.084		0.008	-0.010	
School enrolment for boys aged 5 to 10 years												
PSM - 5 nearest neighbour	0.034	-0.032	-0.038	-0.042	0.041	-0.039	0.123	0.700	0.167	-0.297***	0.052	0.175
PSM - kernel, bandwidth 0.01	0.019	-0.032	-0.192	-0.054	0.052	-0.156	0.110	0.574		-0.294***	0.052	
School enrolment for girls aged 11 to 17 years												
PSM - 5 nearest neighbour	-0.034	-0.090*	0.019	-0.014	-0.067	-0.007	0.084	-0.100	0.167	-0.146	0.032	0.050
PSM - kernel, bandwidth 0.01	-0.026	-0.090*	0.000	-0.030	-0.020	0.013	0.046	-0.105		-0.177*	0.015	

	Borrowers vs. saver			One-time borrower vs. saver			Repeat borrower vs. saver			One-time vs. repeat borrower		
	R1	R2	R3	R1	R2	R3	R1	R2	R3+	R1	R2	R3+
Household level hypotheses – cont.												
School enrolment for boys aged 11 to 17 years												
PSM - 5 nearest neighbour	0.009	-0.042	0.156**	-0.028	0.023	0.144	0.033	-0.433***	0.333	-0.024	0.084	-0.275
PSM - kernel, bandwidth 0.01	0.002	-0.069	0.099	-0.016	0.045	0.096	0.034	-0.398***		0.002	0.099	
Food expenditure per day per capita in Rupees												
PSM - 5 nearest neighbour	1.22	0.69	N/A	1.12	1.65**	N/A	2.00	8.02	N/A	-0.79	0.76	N/A
PSM - kernel, bandwidth 0.01	1.13	0.67	N/A	1.03	1.74**	N/A	1.91	8.48	N/A	-0.60	0.88	N/A
Enterprise level hypotheses												
Informal sector income of whole household - per month in Rupees												
PSM - 5 nearest neighbour	2,444	3,782	N/A	2,469	8,590	N/A	1,627	-10,229	N/A	-27	9,667*	N/A
PSM - kernel, bandwidth 0.01	1,497	3,106	N/A	1,800	8,713	N/A	1,162	-137	N/A	626	8,625	N/A
Informal sector income of respondent only - per month in Rupees												
PSM - 5 nearest neighbour	2,345	175	N/A	2,398	-697	N/A	607	-11,997	N/A	204	94	N/A
PSM - kernel, bandwidth 0.01	1,260	-260	N/A	1,655	-253	N/A	507	-2,084	N/A	800	-126	N/A
Microenterprise revenues of all enterprises in household - per month in Rupees												
PSM - 5 nearest neighbour	1,962	3,095	N/A	2,351	6,177	N/A	1,271	-4,394	N/A	-321	7,019	N/A
PSM - kernel, bandwidth 0.01	1,253	2,581	N/A	1,935	6,180	N/A	688	923	N/A	391	6,004	N/A
Microenterprise revenues of microenterprises for which respondent is primarily responsible - per month in Rupees												
PSM - 5 nearest neighbour	2,215**	545	N/A	2,201	-303	N/A	1059	-6,397	N/A	-515	166	N/A
PSM - kernel, bandwidth 0.01	1,449	257	N/A	1,673	102	N/A	968	-1,212	N/A	104	-15	N/A
Current value of fixed assets of all microenterprises in household in Rupees												
PSM - 5 nearest neighbour	609	718	N/A	979	2,631*	N/A	-661	141	N/A	-28	2,427*	N/A
PSM - kernel, bandwidth 0.01	779	635	N/A	834	2,589*	N/A	25	-193	N/A	-752	2,499**	N/A
Current value of fixed assets of microenterprises for which respondent is primarily responsible in Rupees												
PSM - 5 nearest neighbour	328	287	N/A	118	950	N/A	-498	0	N/A	-906	912	N/A
PSM - kernel, bandwidth 0.01	441	265	N/A	223	938	N/A	112	-23	N/A	-1,671	923	N/A

	Borrowers vs. saver			One-time borrower vs. saver			Repeat borrower vs. saver			One-time vs. repeat borrower		
	R1	R2	R3	R1	R2	R3	R1	R2	R3+	R1	R2	R3+
Enterprise level hypotheses – cont.												
Hours worked in previous week in all microenterprises in household												
PSM - 5 nearest neighbour	5.43	15.45**	N/A	-2.17	21.61*	N/A	8.60	-12.43	N/A	-4.80	19.56*	N/A
PSM - kernel, bandwidth 0.01	3.98	15.48**	N/A	-0.33	22.41**	N/A	7.89	7.37	N/A	-5.29	15.26	N/A
Days worked in previous month in all microenterprises in household												
PSM - 5 nearest neighbour	5.30	9.12**	N/A	1.69	15.27***	N/A	6.59	10.23	N/A	-5.34	14.42**	N/A
PSM - kernel, bandwidth 0.01	3.42**	9.69***	N/A	2.47	14.02**	N/A	6.39	21.83**	N/A	-4.23	10.92*	N/A
Main types of suppliers - inferior suppliers? Yes=1, No=0												
PSM - 5 nearest neighbour	-0.005	0.095**	N/A	0.018	0.092	N/A	-0.026	0.367	N/A	0.017	0.052	N/A
PSM - kernel, bandwidth 0.01	-0.013	0.096**	N/A	0.006	0.081	N/A	-0.010	0.355	N/A	-0.039	0.013	N/A
Main types of customers - inferior customers? Yes=1, No=0												
PSM - 5 nearest neighbour	0.130***	0.131***	N/A	0.124**	0.192***	N/A	0.059	0.500***	N/A	0.017	0.126	N/A
PSM - kernel, bandwidth 0.01	0.104**	0.133***	N/A	0.119**	0.195***	N/A	0.092	0.528***	N/A	-0.027	0.081	N/A
Individual level hypotheses												
Respect by other household members? Yes=1, No=0												
PSM - 5 nearest neighbour	0.012	0.008	N/A	0.006	0.026	N/A	0.042	0.033	N/A	-0.04	0.002	N/A
PSM - kernel, bandwidth 0.01	0.012	0.014	N/A	-0.001	0.021	N/A	0.034	0.032	N/A	-0.04	0.000	N/A
Prepared to deal with future? Yes=1, No=0												
PSM - 5 nearest neighbour	0.013	0.039	N/A	0.039	0.059	N/A	0.046	0.200**	N/A	0.013	0.104**	N/A
PSM - kernel, bandwidth 0.01	0.022	0.041	N/A	0.029	0.097**	N/A	0.043	0.239***	N/A	0.002	0.077	N/A

Source: Author's calculations.

Notes: *statistically significant at 10%, **statistically significant at 5%, ***statistically significant at 1%.

The results in this table refer to the differences in the mean values between matched samples; they were obtained using the STATA command psmatch2. I also ran the STATA command pscore with the objective to cross-check the psmatch2 results across the various matching algorithms. The results I obtained from the different STATA routines displayed minor differences in terms of the size of coefficient and the level of significance. The results presented in Appendix 3 do not apply sampling weights. + No values for kernel matching in Round 3, the sample was too small with propensity scores outside common support region, no adequate matches could be found. Results are bootstrapped.

Appendix 4: Weighted means and standard deviations, PnK and RnM

Variables	PnK 1998[1]		RnM 2009[2]	
	Mean	Standard deviation	Mean	Standard deviation
Age of all individuals	23	18	23	18
Schooling of individual aged 5 or above (years)	1.377	2.773	2.066	3.136
Parents of HH head own land?	0.256	0.564	0.254	0.563
Brothers of HH head own land?	0.815	1.308	0.810	1.305
Sisters of HH head own land?	0.755	1.208	0.750	1.206
Parents of HH head's spouse own land?	0.529	0.784	0.529	0.783
Brothers of HH head's spouse own land?	0.919	1.427	0.919	1.427
Sisters of HH head's spouse own land?	0.753	1.202	0.753	1.202
Household land (decimals)	76.142	108.540	76.145	108.052
Highest grade completed by HH head	2.486	3.501	2.523	3.525
Sex of household head (male=1)	0.948	0.223	0.948	0.223
Age of household head (years)	40.821	12.795	40.874	12.789
Highest grade completed by any female HH member	1.606	2.853	1.664	2.999
Highest grade completed by any male HH member	3.082	3.081	3.277	4.016
Adult female not present in HH?	0.017	0.129	0.017	0.129
Adult male not present in HH?	0.035	0.185	0.035	0.185
Spouse not present in HH?	0.126	0.332	0.123	0.329
Amount borrowed by female from BRAC (Taka)	350	1,574	349	1,564
Amount borrowed by male from BRAC (Taka)	172	1,565	173	1,575
Amount borrowed by female from BRDB (Taka)	114	747	114	746
Amount borrowed by male from BRDB (Taka)	203	1,573	204	1,576
Amount borrowed by female from GB (Taka)	956	4,293	972	4,324
Amount borrowed by male from GB (Taka)	374	2.923	360	2,895

Notes:
1. Source: PnK, table A1, p. 993, based on R1.
2. Source: RnM, table 1, p. 15, based on R1.

Morduch and Pitt do not provide any descriptive statistics.

Appendix 5: Simplified summary overview of headline results for all PnK related studies dealing with R1-3

	PnK 1998	Morduch 1998	Pitt 1999	Chemin 2008	RnM 2009	Author 2010
Method	WESML-LIML-FE	DID	Expansion of PnK model and comparison to Morduch using a simulation-based approach	PSM	cmp	PSM, DID, cmp
Particularities	Impact assessed by gender and separately for all three microcredit programmes, disregards other non-microcredit sources of borrowing, eligibility criteria not always strictly enforced	Impact assessed by various eligibility criteria; de jure and de facto, separately for all three microcredit programmes; no impact assessed by gender of borrower, disregards other non-microcredit sources of borrowing	Impact assessed by gender and separately for all three microcredit programmes, disregards other non-microcredit sources of borrowing, eligibility criteria refined, Pitt confirms PnK's results and refutes Morduch's claims	No impact assessed by gender of borrower, disregards other non-microcredit sources of borrowing, disregards the various eligibility criteria, all three microcredit programmes are pooled and their combined impact is assessed	Impact assessed by gender and separately for all three microcredit programmes, disregards other non-microcredit sources of borrowing	Impact assessed by gender and across all borrowers, various eligibility criteria used, other non-microcredit sources are considered and various treatment group comparisons are conducted, all three microcredit programmes are pooled and their combined impact is assessed

	PnK 1998	Morduch 1998	Pitt 1999	Chemin 2008	RnM 2009	Author 2010
Outcome variables						
Variation of log per capita expenditure (Taka)	N/A	No impact	N/A	No impact	N/A	No impact
Log per capita expenditure (Taka)	Significantly positive impacts	Significantly negative impacts	Significantly positive impacts	N/A	Significantly negative impacts	No impact
Log women non-landed assets (Taka)	Significantly positive impacts	N/A	Significantly positive impacts	No impact	Significantly positive impacts, in particular when women are involved in borrowing	Significantly positive impacts, in particular when women are involved in borrowing
Female labour supply, aged 16-59 years, hours per month	Significantly positive impacts	Significantly negative impacts, in particular when men are involved in borrowing	Significantly positive impacts	No impact	Significantly negative impacts, in particular when men are involved in borrowing	Significantly positive impacts only when women are involved in borrowing, otherwise significantly negative impacts when men are involved
Male labour supply, aged 16-59 years, hours per month	No impact	No impact	No impact	Significantly positive impacts	No impact	Significantly negative impacts, in particular when women are involved in borrowing
Girl school enrolment, aged 5-17 years	Significantly positive impacts	No impact	Significantly positive impacts	Significantly positive impacts	No impact	Significantly positive impacts
Boy school enrolment, aged 5-17 years	Significantly positive impacts	No impact	Significantly positive impacts	Significantly positive impacts	No impact	Significantly positive impacts

Source: Author's illustration.