Essays on Microfinance Repayment Behaviour:

An Evaluation in Developing Countries

Thesis submitted to The University of Reading

for the degree of Doctor of Philosophy

HENLEY BUSINESS SCHOOL

THE UNIVERSITY OF READING

ICMA Centre

Guan Huang

February 2018

Abstract

Microfinance research concerns addressed in this thesis relate to: the associations between the individual characteristics of borrowers and the probabilities of being in delinquent or default; the determinants for the financial awareness of interest repayment; and the application and comparison of modern missing data techniques (Multiple Imputation, Maximum Likelihood Estimation, and Predictive Mean Matching) with incomplete loan book data. The thesis emphasises credit scoring issues that affect repayment performance and revolves around three empirical chapters that seek to address the above research concerns.

Survey and loan book data from individuals in 51 MFIs across 27 developing countries. The data were compiled by the MFIs and collected by Micro Finanza Rating. Varied micro-econometric techniques (ordinary least squares, Logit regression, Tobit regression, Two-Part model, Double-Hurdle model, Box-Cox transformation, and three missing data imputation methods: Multiple Imputation, Maximum Likelihood Estimation, and Predictive Mean Matching) are used depending on the hypotheses being considered in each chapter.

The main findings are: engaging in agriculture is related to a lower probability of default that measured by the amount of arrear in general; besides, the association between agriculture and the length of delayed repayment is insignificant; previous access to micro-finance has positive association with the financial awareness of the clients who lived in urban areas; in addition, previous access to saving service has positive effect on the clients with at least primary education; when the missing microfinance data is semi-continuous, PMM outperforms MI and ML in most simulations; for binary or ordinal categorical data, PMM performance surpass MI and ML only when the sample sizes of data are large, the missing rates are low, and the missing mechanism is MAR.

The thesis suggests the following recommendation both for management of MFIs and government: we need to make financial services for poor farm households and farm-related business more attractive to the MFIs; financial awareness can be improved by access to microfinance services, hence extra learning programmes may be unnecessary; Two-Part Model should be applied to credit scoring; and PMM imputation is the best technique to be applied to deal with the missing data issues and improve data quality in microfinance.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other institute of learning.

Copyright Statement

List of Contents

Abstract……i

Declaration……ii

Copyright Statement……ii

List of Tables……v

List of Abbreviations ……ix

Chapter 1: Introduction……1

1.1 Motivation……1

1.2 Objectives……4

1.3 Structure of the Thesis……5

Chapter 2: Individual and Household Level Impacts of Microfinance Programmes on the Well-being of the Poor……6

2.1 Introduction……6

2.2 Methodologies of the Impact Assessment Studies……7

2.2.1 Definitional and conceptual Issues……7

2.2.2 Research designs and validity……10

2.2.3 Statistical methods of analysis……11

2.2.4 Common bias in microfinance assessments……12

2.3 Major Empirical Results and Controversies……13

2.3.1 Significant impacts of micro loans ……13

2.3.2 Insignificant impacts of micro loans ……14

2.3.3 Impacts of interest rate policy and group lending……15

2.3.4 Impacts of access to finance ……15

2.3.5 The issues of data quality……16

2.4 Limitations of the Current Literature……17

2.4.1 Quasi-experimental studies……17

2.4.1.1 With/without and before/after studies……17

2.4.1.2 Pipelines studies……19

2.4.2 Experimental studies (RCTs)……21

2.4.3 Women empowerment studies……23

2.5 Conclusions and Implications……24

Chapter 3: Delinquency of Microfinance: A Two-part Probit Analysis of Cross-MFI Data……25

3.1 Introduction……25

3.2 Literature Review……27

3.2.1 The concepts of loan delinquency and loan default……27

3.2.2 Determinants of loan delinquency or default in microfinance……28

3.2.2.1 Main findings related to individual socio-geographic characteristics……28

3.2.2.2 Main findings related to business characteristics……31

3.3 Data……33

3.4 Methodology……36

3.4.1 Econometrics models for censored data……36

3.4.2 The Tobit model……37

3.4.3 The Two-Part model (2PM)……40

3.4.4 Cragg's Double-Hurdle model (DH)……42

3.4.5 The Box-Cox transformation and likelihood functions……44

3.5 Empirical Results and Discussion……45

3.5.1 Model comparison and relations between marital status and loan delinquency……45

3.5.2 Subsample analysis between different microfinance institutions……48

3.5.3 Interaction terms analysis of marital status……50

3.5.4 Further analysis of the impacts of loan classification standards……52

3.6 Conclusions and Discussion……54

Chapter 4: What Drives Financial Awareness in Microfinance? ……64

4.1 Introduction……64

4.2 Literature Review……66

4.2.1 Financial awareness and the framework of financial literacy……66

4.2.2 Main findings in the previous studies……67

4.2.3 Limitations of the previous studies……69

4.2.4 Motivations and research questions……70

4.3 Data and Methodologies……71

4.3.1 Descriptive statistics……71

4.3.2 Missing data imputation methods……73

4.3.3 Estimation methods……74

4.4 Results and Discussion……76

4.4.1 Relation between access to credits and financial awareness……76

4.4.2 Split-sample Analysis……78

4.4.3 Interaction effects……80

4.4.4 Further Analysis……81

4.5 Conclusions and limitations……82

Chapter 5: Multiple Imputation, Maximum Likelihood and Predictive Mean Matching for Semi-continuous Missing Data: A Study of A Microfinance Administrative Loan Book……95

5.1 Introduction……95

5.2 Goals of this research……96

5.3 Literature Review……97

5.3.1 The Distribution of Missing Data……97

5.3.2 Multiple Imputation and Maximum Likelihood Estimation……99

5.3.3 Predictive mean matching……100

5.4 Data and Missingness Simulation……102

5.4.1 Modifying the population……102

5.4.2 Sampling benchmark datasets and skewness preservation……102

5.4.3 Generating missingness……104

5.5 Missing Data Imputation Methods……107

5.5.1 Multiple Imputation……107

5.5.2 Predictive Mean Matching……109

5.5.3 Maximum Likelihood Estimation……109

5.5.4 Completed Case Analysis……112

5.5.5 Evaluation of Imputation Performance……113

5.6 Empirical Results……114

5.6.1 Semi-continuous variable in univariate missing data……114

5.6.2 Semi-continuous variable in multivariate missing data……116

5.6.3 Binary variable in univariate missing data……117

5.6.4 Binary variable in multivariate missing data……119

5.6.5 An Ordinal variable in univariate missing data……120

5.6.6 An Ordinal variable in multivariate missing data……121

5.7 Conclusions and Discussion……123

Chapter 6: Conclusion……144

6.1 Introduction……144

6.2 Summary of Results……145

6.2.1 Delinquency of microfinance……145

6.2.2 What drives the financial awareness in microfinance? ……146

6.2.3 MI, ML and PMM for semi-continuous missing data in loan book……147

6.3 Implications and Recommendations……148

6.4 Limitations and Future Considerations……150

Appendix A: Top 25 Business Activities in Microcred and Finca Peru by Population……152

Appendix B: Financial knowledge, attitudes and behaviour (average scores); Stacked points (weighted data): all respondents, sorted by overall score……153

Appendix C: Basic Information of the Microfinance survey database……154

References……155

List of Tables

Table 3.1: Summary Statistics……..56

Table 3.2: Correlations Analysis……57

Table 3.3 Panel A: MLEs for Four Models (Dependent: Amount in Arrears (in USD))……..58

Table 3.3 Panel B: MLEs for Four Models (Dependent: Delayed Repayment (in Days))……..59

Table 3.4: Subsample Analysis between MFIs (2PM)……60

Table 3.5 Panel A: Interactions Analysis (Dependent: Amount in Arrears (in USD))……..61

Table 3.5 Panel B: Interactions Analysis (Dependent: Delayed Repayment (in Days))……...62

Table 3.6: Logistic Regression Analysis Based on Credit Collection Process……63

Table 4.1: Descriptive Statistics of the Data before Appling Multiple-Imputation……..84

Table 4.2: Descriptive Statistics of the Data after Appling Multiple-Imputation……..84

Table 4.3: Previous Access to Credits and Financial Awareness of Interest Rate (Part 1)……..85

Table 4.3: Previous Access to Credits and Financial Awareness of Interest Rate (Part 2)……..86

Table 4.4: Previous Access to Credits Regional and Religious Split-Sample Analysis (Part 1)..87

Table 4.4: Previous Access to Credits Regional and Religious Split-Sample Analysis (Part 2)..88

Table 4.5: Interactions on the Relation between Previous Access to Credits and Socio-demographic Characteristics (Part 1)……89

Table 4.5: Interactions on the Relation between Previous Access to Credits and Socio-demographic Characteristics (Part 2)……90

Table 4.5: Interactions on the Relation between Previous Access to Credits and Socio-demographic Characteristics (Part 3)……91

Table 4.5: Interactions on the Relation between Previous Access to Credits and Socio-demographic Characteristics (Part 4)……92

Table 4.6 Robustness Check with Alternative Proxy of Interest Repayment Awareness……93

Table 5.1: Summary Statistics of 11 Variables (N=3,200)……125

Table 5.2 Panel 1: RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Univariate MCAR Missing Data) – Grouping by Sample sizes……126

Table 5.2 Panel 2: RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Univariate MAR Missing Data) – Grouping by Sample sizes……..127

Table 5.2 Panel 3: RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Univariate MNAR Missing Data) – Grouping by Sample sizes……128

Table 5.3 Panel 1: RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Multivariate MCAR Missing Data) – Grouping by Sample sizes……129

Table 5.3 Panel 2: RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Multivariate MAR Missing Data) – Grouping by Sample sizes……..130

Table 5.3 Panel 3: RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Multivariate MNAR Missing Data) – Grouping by Sample sizes……..131

Table 5.4 Panel 1: RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Univariate MCAR Missing Data) – Grouping by Sample sizes……132

Table 5.4 Panel 2: RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Univariate MAR Missing Data) – Grouping by Sample sizes……..133

Table 5.4 Panel 3: RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Univariate MNAR Missing Data) – Grouping by Sample sizes……..134

Table 5.5 Panel 1: RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Multivariate MCAR Missing Data) – Grouping by Sample sizes……..135

Table 5.5 Panel 2: RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Multivariate MAR Missing Data) – Grouping by Sample sizes……..136

Table 5.5 Panel 3: RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Multivariate MNAR Missing Data) – Grouping by Sample sizes……..137

Table 5.6 Panel 1: RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Univariate MCAR Missing Data) – Grouping by Sample sizes……138

Table 5.6 Panel 2: RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Univariate MAR Missing Data) – Grouping by Sample sizes……..139

Table 5.6 Panel 3: RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Univariate MNAR Missing Data) – Grouping by Sample sizes……..140

Table 5.7 Panel 1: RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Multivariate MCAR Missing Data) – Grouping by Sample sizes……..141

Table 5.7 Panel 2: RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Multivariate MAR Missing Data) – Grouping by Sample sizes……..142

Table 5.7 Panel 3: RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Multivariate MNAR Missing Data) – Grouping by Sample sizes……..143

List of Abbreviations

2SLS	Two-Stage Least Square
2PM	Two-Part Model
AIC	Akaike's Information Criterion
CCA	Complete Case Analysis
CI	Confidence Interval
CP	Coverage Probability
DH	Cragg's Double-Hurdle Model
DID	Difference in Differences
FDH	Full Doule-Hurdle Model
GE	Generalization Error
IADB	Inter American Development Bank
INFE	International Network on Financial Education
IV	Instrumental Variables
LD	Listwise Deletion
MAR	Missing at Random
MCAR	Missing Completely at Random
MDT	Missing Data Techniques
MFI	Microfinance Institutions
MI	Multiple Imputation
MI-LOGIT	Multiple Imputation for Logistic Variables
MI-OLOGIT	Multiple Imputation for Ordered Logistic Variables
MICE	Multiple Imputation by Chained Equations
ML	Maximum Likelihood
MNAR	Missing Not at Random
MSE	Mean-Square Error
NGO	Non-Governmental Organisation

NPL	Nonperforming Loans to Total Gross Loans
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Square
PaR	Portfolio at Risk
PaR30	Portfolio at Risk for 30 days
PMM	Predictive Mean Matching
PSM	Propensity Score Matching
RCT	Randomised Control Trials
RMSE	Root Mean Square Error
SME	Micro, Small and Medium-Sized Enterprises

Chapter 1:

Introduction

1.1 Motivation

Microfinance is an emerging market particularly amongst the urban and peri-urban populations in developing countries. Private microfinance institutions and local governments are the primary support for the growth of this sector. Technically speaking, microfinance is a business in which the lenders provide short-term loans to small or micro enterprises or low-income households, and characterised by the use of collateral substitutes. Microfinance is a way of supplying small credits to finance small projects to help the poor have an income through forming their own small-scale business to earn their daily bread and improve their living standards. Microfinance Institutions (MFIs) use social sanctions and credit denial as punishments for defaulting borrowers. These punishments serve the role of collateral substitutes. However, a successful social sanction requires navigating a delegation problem. Besides, the credit denial lacks market value. It may lead to adverse selection and a higher probability of a non-repayment equilibrium.

A delayed instalment is said to be delinquent, and a repayment that has not been made is said to be in default. The possibilities of being delinquent or defaulted in the microfinance industry are controversial. The rapid proliferation of MFIs has drawn criticism. Howard et al. (2006) indicate that some people fear that it has outpaced the capacity of the developing country governments to implement regulatory measures, and it created a wild environment in which borrowers with limited financial knowledge may be exploited by incompetent or immoral lenders. In order to alleviate poverty, provision of subsidised credit was embraced by lots of countries during the period from 1950' to 1980'. The repayment rates often dropped below 50%. These experiences were almost disasters (Morduch, 1999). Loan delinquency and default have continued to cause severe challenges to most MFIs.

It is in this regard that the first objective of this thesis was designed to investigate the determine factors associated with loan delinquency among the microfinance participants. Regarding the current literature, most empirical studies focus on the effects of business

characteristics of micro-enterprises and the credit policies of a single Microfinance Institution. Studies related to individual socio-geographic characteristics with high-quality cross-MFI data are scarce. On the other hand, the indicator of delinquency or default is usually dichotomised to a dummy variable in the prior studies. Information lost is severe. These two issues are the motivations for the first empirical study in this thesis - A two-part probit analysis which focuses on the individual characteristics of microfinance borrowers.

Regarding the determinants of loan delinquency and default, one of the most well-known factors is financial literacy. Traditionally, financial literacy refers to the sets of knowledge and skills that allow an individual to make effective decisions with his/her financial resources. However, there is no universal definition of it. In the previous literature, most authors established their versions of measurements for financial literacy based on their research objectives that linked to specific financial education programmes provided by MFIs.

One of the most widely used frameworks to measure the financial literacy is suggest by Lusardi and Mitchell (2008). In their framework, there are three basic financial questions corresponding to interest rates, inflation, and diversification. It is obvious that such a framework does not include a measurement of financial awareness. In fact, Carpena et al. (2011) claim that the financial literacy programmes may affect a client's financial decision-making process through other channels besides developing his/her computational capability and common sense. Access to finance can make individuals and households more aware of their financial conditions and available products, and reshape their attitudes towards financial behaviours. INFE (2011) also defines financial literacy as a combination of awareness, knowledge, skill, attitude and behaviour necessary to make financial decisions.

Considering the characteristics of microfinance, financial awareness might be the most important factor associated with loan delinquency or default. Different from loans from commercial banks, the loans from MFIs usually have small amounts, high-interest rates, short instalment intervals, short repayment cycles, and low levels of collateral. In this case, the capability of self-control and personal preference outweigh the financial capability and skill. The microfinance studies that focus on the financial literacy are rare, don't even bother the financial awareness, financial attitude, etc. It motivates me to study the relations between the clients' individual or household level characteristics and their financial awareness of interest repayment in the microfinance industry specifically. In the second empirical study, I focus on the relations between financial awareness and a client's previous access to microfinance services.

During the studying of loan delinquency and financial awareness in the microfinance industry, an unexpected issue arises. I found that almost all variables in the administrative loan book data and client survey data have a certain number of missing values, and the missing percentages for some variables are even higher than 20%. In fact, the situation of incomplete data exists in many areas of empirical research, especially prevalent in social and behavioural studies. In many cases, missing values simply happen when respondents are not available for the surveys, or there is a human mistake when collecting the data and the information is damaged.

To deal with the missing data, the simple complete case analysis (CCA) techniques such as listwise and pairwise deletions are still popular in many papers nowadays. King et al. (2001) have reviewed a great number of studies and concluded that data analysis in political science research typically loses a third of the cases due to listwise deletion of missing data. The increase in MSE is comparable to what we can expect from omitted variable bias. CCA may generate significant biases when the percentage of missing values reaches about 20%. For instance, when a dataset has ten variables and 3% of data randomly missing in each variable, then the total missing percentage for the dataset may vary from 15% to 30% if CCA is applied. As results, the researchers who process the missing data with CCA have to put up with either the severe information loss caused by a high percentage of missing data or dropping a great number of incomplete explanatory variables.

With modern missing data techniques (MDT), we can impute the missing values, so the maximum amount of information is restored and keep the data less biased at the same time. It helps us to perform more robust empirical analysis and obtain more convincing results. Popular MDT which are potentially suitable for microfinance loan book data and survey data include Multiple Imputation (MI), Maximum Likelihood (ML) estimation, and Predictive Mean Matching (PMM). While the previous literature suggests that these MDT outperform the traditional CCA in most cases, it is unclear whether they are preferable when the real missing mechanism is unknown, and the assumption of a normal distribution is violated, such as the semi-continuous variable ‘Arrears’ in our data. It motivates me to implement a systematic evaluation of the missing data imputation performances of MI, ML and PMM with semi-continuous data.

1.2 Objectives

The aim of the first empirical chapter lies in addressing the individual level determinants relevant to the microfinance delinquency. The second empirical chapter sets out to be the first rigorous cross-MFI study of the relation between a client's individual/household level characteristics and financial awareness of interest repayment. Finally, the third empirical chapter provides a systematic evaluation for the imputation performances of MI, ML and PMM with actual administrative loan book data, as there are so few performance comparison studies of different missing data techniques available in the current literature.

More specifically, these three chapters in the thesis respectively study the following eleven hypotheses and four research questions:

H1. Married individuals have a lower probability of default and lower intensity of delinquency.

H2. The youngest and oldest borrowers have a lower probability of default and a smaller intensity of delinquency, while the middle-age group of consumers have a higher probability of default and a larger intensity of delinquency.

H3. Female borrowers have a lower probability of default and a smaller intensity of delinquency.

H4. Borrowers with higher educational levels have a lower probability of default and a smaller intensity of delinquency.

H5. The credit destined to an agricultural sector has a lower probability of default and a smaller intensity of delinquency.

H6. Women have a lower probability of being aware of their interest rate.

H7. Older borrowers have a lower probability of being aware of their interest rate

H8. Less educated borrowers have a lower probability of being aware of their interest rate.

H9. Borrower living in rural areas have a lower probability of being aware of their interest rate.

H10. Clients who have saving account before (or previously accessed to moneylenders, previously accessed to MFIs, previously accessed to formal banks) have a higher probability of being aware of their interest rate.

H11. Borrowers who have no education, but have saving account before (or previously accessed to the moneylenders, previously accessed to MFIs, or previously accessed to formal banks) have a higher probability of being aware of their interest rate.

Q1. Will Predictive Mean Matching consistently outperform Multiple Imputation, Maximum Likelihood estimation, and Complete Case Analysis, across different types of data especially for semi-continuous variables?

Q2. Will Predictive Mean Matching consistently outperform Multiple Imputation, Maximum Likelihood estimation, and Complete Case Analysis, across different missing mechanisms?

Q3. Will Predictive Mean Matching consistently outperform Multiple Imputation, Maximum Likelihood estimation, and Complete Case Analysis, across different sample sizes?

Q4. Will Predictive Mean Matching consistently outperform Multiple Imputation, Maximum Likelihood estimation, and Complete Case Analysis, across different missing data rates?

All hypotheses and research questions will be discussed in detail and motivated based on existing literature in the empirical chapter 3 to 5.

1.3 Structure of the Thesis

The thesis is structured around three empirical chapters that seek to address the above research questions. Chapter 2 presents a systematic review of the prior literature related to impact assessment of microfinance programmes on the well-being of the poor. It investigates the methodologies, empirical results, and potential biases of the previous studies. The empirical Chapter 3 to 5 present the main body this thesis. Chapter 3 provides a discussion on the relationships between individual or household level characteristics and loan delinquency. Chapter 4 assesses the influences of previous access to credit on the financial literacy of a client. In Chapter 5, I evaluate the imputation performances of MI, ML and PMM on a real microfinance loan book data under various combinations of sample sizes, missing rates, missing mechanisms, and data types. The last chapter of the thesis provides a summary of the empirical chapter, implication recommendations, and areas for further work.

Chapter 2:

Literature Review: Individual and Household Level Impacts of Microfinance Programmes on the Well-being of the Poor

---

2.1 Introduction

Since the Nobel Peace Prize winner, Muhammad Yunus firstly introduced the concept of microfinance into Grameen Bank 39 years ago, whether microfinance programmes could generate positive impacts have been studied for a long time. Theoretically, it has the potential to enable income-generating investments, smooth consumption and reduce financial vulnerability. In 2011, the United Nations Capital Development Fund even tried to explore microfinance as a practical social protection tool. Beginning with the traditional financial intervention which only provided credit to the poor, microfinance has evolved over decades and now includes many services, such as micro-savings, micro-leasing, micro-insurance and financial training programmes. In general, microfinance is apparently successful and promising, at least in the early evaluations.

However, according to the recent reviews of literatures related to microfinance impact assessments (Gaile and Foster, 1996; Sebstad and Chen, 1996; Goldberg, 2005; Odell, 2010; Duvendack et al., 2011; Orso, 2011; Stewart et al., 2012), we have no convincing objective evidence of either positive or negative impacts. Rigorous quantitative results are rare and inconclusive (Armendariz and Morduch, 2005). Also, whether microfinance programmes that focused on women were more effective was unclear (for instance, Pitt and Khandker, 1998 vs Karlan and Zinman, 2009). Overall, the empirical findings of the effectiveness of microfinance programmes are still controversial.

This review set out to discuss and summarise not only the major findings in previous literature but also their research designs, statistical analysis methods, limitations and potential biases. These technical challenges could provide a better view of the current research situation and lay a solid foundation for the further impact assessment studies. The rest of the review is structured as follow: Section 2.2 introduces the methodologies that used in the previous literature in details. Section 2.3 presents the major empirical impact evaluation

findings and controversies. As a base of further research, the bias and limitations are discussed in the context of the representative papers in Section 2.4. The conclusion and implications for research are presented at the end.

2.2. Methodologies of the Impact Assessment Studies

In general, while various methodologies have been implemented to the microfinance impact assessments, a few of them are found to be dominating the studies throughout the years. Micro-credit was the most widely studied financial intervention, following by micro-saving and micro-leasing in sequence. Other interventions have rarely been explored. Income, enterprise profits/revenues, housing improvements, education, and women empowerment were the dominating dependent variables. In terms of research designs and statistical analysis methods, with/ without (before/after) comparisons and Propensity Score Matching were the mainstream techniques. It is also noticeable that a high proportion of reviewed studies exposed to the risk of selection bias. All these features and more details of the methodologies are discussed in the subsections below.

2.2.1 Definitional and conceptual Issues

The key econometric characteristics of the literature reviewed and the relationships between the central concepts are outlined and defined as follows:

Participants of Microfinance Programmes (Treatments and Controls): The papers reviewed in this chapter mainly focus on individuals living in 40 low income and 56 lower-middle income countries with very few assets that can be used as collaterals. As defined by the World Bank, GNI per capita was the main criterion to classify countries. Participants of microfinance programmes have to be identified as poor or vulnerable within their society. Target groups might include individuals, households or microenterprises that were exposed to the influence of particular microfinance services.

Microfinance Interventions (Independents): Microfinance interventions are complex and diverse. For instance, a credit product may involve savings, training and etc. The papers included in this review focus on three of the largest financial inclusion services: micro-credit, micro-savings and micro-leasing (e.g., Stewart et al., 2012). Micro-credit is the provision of small loans to the poor, usually in cash, with considerably varying interest rates between 20% and 40%. While some MFIs charge a fixed rate on the amount borrowed, a floating rate

is more commonly used. Micro-savings is a deposit service which is usually linked to credit as a compulsory condition of an individual loan or a pool of shared group savings resources. It protects participants from unexpected shocks and encourages them to build an asset base (Hulme et al., 2009). Micro-leasing is a contractual arrangement which allows the lessee to use an asset owned by the lessor in exchange for specified periodic payments (Gallardo, 1997). It enables the poor to access productive assets and to generate income. All these financial services mentioned in previous studies were provided by basic, transformed or commercial MFIs, NGO MFIs, commercial banks, credit cooperatives and other public sector financial services providers.

Economic, Social and Empowerment Outcomes (Dependents): There are hundreds of outcome variables have been tested in the reviewed papers. Hence, I need to organise them into groups for presentation. In terms of category, the outcomes can be classified into economic (mainly credit received from microfinance, business inputs, production, sales, profits, expenditures, housing, durables and assets), social (mainly health and education expenditures) and empowerment (control power of home expenditure and strength of social interaction, exclusively of women). Economic indicators have been dominating microfinance assessments for long. Measuring changes in income is the first choice of many researchers though changing income alone is insufficient to draw conclusions about the status of household members. Social indicators became popular in the 1980s and had been introduced into microfinance as an attempt to examine if microfinance could contribute to empowerment (Goetz and Sen Gupta, 1996; Schuler and Riley, 1996; Mayoux, 1997). This development has led to the new measurements such as individual control over their resources, discursive power in the household decision and community participation, and permitted the developmental impacts to be assessed in a much more sophisticated manner.

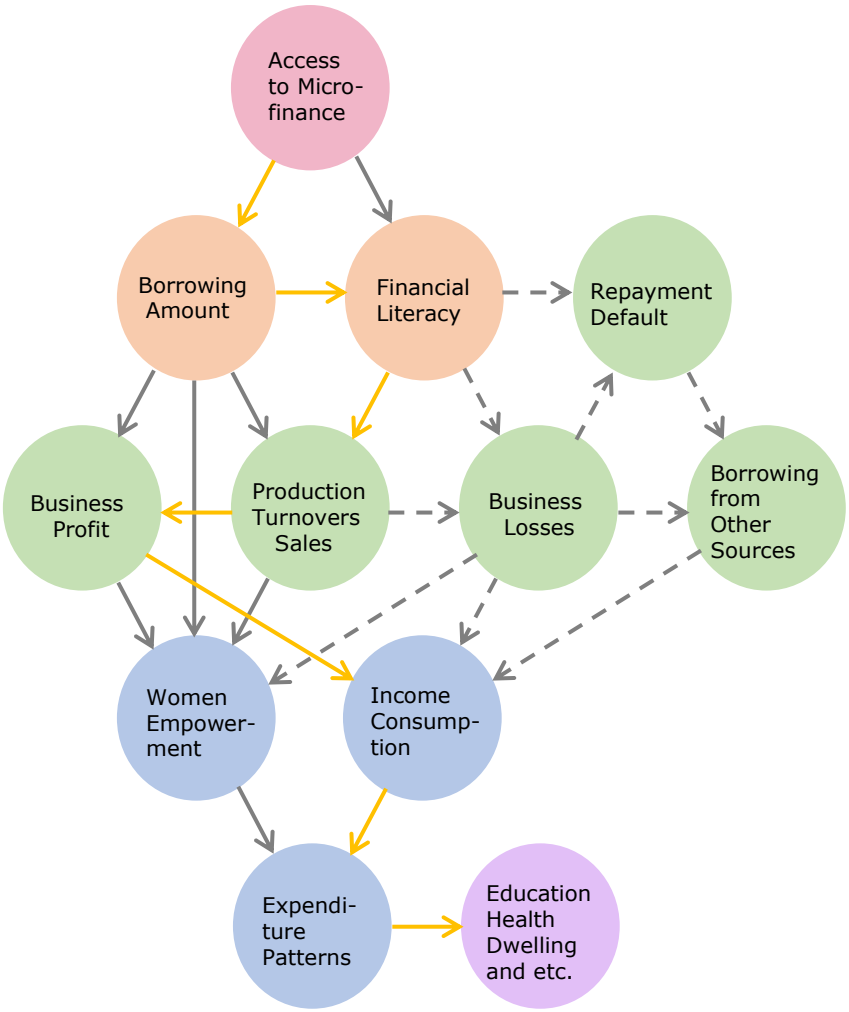
Lots of studies, such as those will be discussed in subsection 2.3.2, have conducted impact assessments on the same data with different sub-samples and methods of estimation. Two of the most iconic databases are: 1. the cross-sectional data from a World Bank funded research which conducted a survey in 1991-1992 on three leading microfinance group-lending programmes in Bangladesh (Pitt and Khandker, 1998); 2. the three longitudinal studies in the late 1990s on Peru, India & Zimbabwe funded by USAID. Regarding the impact assessment studies reviewed in this chapter, our statistics show that 75% of them were implemented on economic indicators. Only 16% and 9% of them were implemented on social and empowerment indicators respectively. Besides, more than 40% of the economic and empowerment impact assessments have found significant results, while the percentage for social was just

25%. It appears that trying to identify the relationships between the microfinance interventions and the improvements of programme participations' livelihoods is a challenging task.

Hierarchy of the Microfinance Outcomes (Theory Map): While a number of outcomes reflect the direct influences of access to microfinance, such as increases in borrowing, most of the others have few specific implications for the value of microfinance, such as improved dwelling conditions. In order to review the previous studies that spread on the different levels of the causal chain of microfinance impacts more systematically, I have developed a rough network which illustrates the hierarchy and the potential relationships between various outcomes are presented in **Figure 2.1**. Red and violet represent the beginning and the end of the entire impact casual chain respectively. Orange, green and blue indicate the outcomes' closeness to personal/household well-being. The common positive and negative relationships between outcomes are marked by solid and dash lines. Starting at the top is the effect of access to microfinance, which leads to increased borrowing. Through this route, access to credit will increase the cash balances, improve the financial literacy, and indirectly influences sales and revenues. If the enterprise succeeds, it will lead to greater profit, income and consumption. Otherwise, access to credit may lead to a series of negative outcomes shown on the right of the figure. Finally, access to microfinance reshapes the expenditure patterns of participants and hence improves their livelihoods qualities. Despite the complications shown here, most of the research included in this review was framed by the most simplistic causal models that directly link access to credit to the final-stage indicators of well-being. Therefore, the subsequent theory can be expanded on those indicators of effect intermediate, such as the attitudinal impacts on clients.

Figure 2.1

Hierarchy, Relationships and Outcomes of Microfinance



2.2.2 Research designs and validity

Because of the difficulties of assessing impacts in development, it is not easy to solve the uncertainty of microfinance's effectiveness (Haddad, 2011, Karlan and Appel, 2011). To isolate the impact of microfinance is the key to identify what would happen without it during assessments. This section presents the features of the major research designs encountered in the previous microfinance impact assessment studies.

Based on a hierarchical order of internal validity (from strongest to weakest), research designs in the reviewed papers can be classified into three levels: experimental, quasi-experimental and simple comparison studies. Simple comparison studies investigate groups with and without interventions by using the data collected after interventions. Thus, their results

can only be considered as the association between variables but not strong evidence of causality. On the contrary, experimental studies provide the most robust results with random allocation to intervention and comparison groups including with/ without (before/after) data. The Randomised Control Trials (RCTs), which constitute the only purely experimental method, have been widely debated as an assessment tool of social and development interventions. While many supports highly appreciate RCTs and use RCTs as the standards to judge other research designs, there are still limitations, such as double-blinding, pseudo (meaning) effects, experiment effects and the assumption of no spill-over effects (Blundell and Costa Dias, 2009). In fact, there are few studies have been done using RCTs as they are costly and laborious. Quasi-experimental studies are the compromised form of the pure experiments. They are either unable to randomise the participants of an intervention or unable to acquire the ex-ante data. All the research designs described as pipelines, with/without (before/after) comparisons, panel, longitudinal or natural experiments are quasi-experiments. The common threats to their validity are: 1. non-random allocation; 2. the risk of confounding; and 3. the bias of selections and programme placements. Despite these disadvantages, Cook et al. (2008) and Kunz et al. (2007) have shown that quasi-experiments can generate similar findings to RCTs by using appropriate statistical methods.

In summary, none of the research designs is absolutely superior. Roger (2010) reminds us that the quality of evidence should be judged by whether the internal and external validity has been systematically checked, instead of whether a particular method has been used. However, it is found that the with/without (before/after) comparisons and panels have been used in over 80% of the reviewed papers, while that percentage for pipelines and RCTs were just 13% and 4% respectively.

2.2.3 Statistical methods of analysis

This subsection discusses the characteristics, limitations, and applications of the most commonly used statistical methods used in the microfinance impact assessment studies.

Propensity Score Matching (PSM): The basic idea of matching on microfinance is to estimate the effect of an intervention of a particular programme by accounting for a group of covariates that influence receiving the intervention. Thus, PSM can account for the selection on observables and reduce the bias caused by confounding variables during the estimations. Noticeably, the selection on unobservable remains unaccounted for. The drawback of PSM is that matching estimators are sensitive the choice of data and not robust enough. It means that matching is appropriate only when high-quality data are available (Smith and Todd,

2005). Therefore, a sensitivity analysis which explores the robustness of matching estimators is crucial to obtain rigorous results.

Difference in Differences (DID): In contrast to the with/without and the before/after estimate of intervention effect, the DID estimators represent the difference between the differences of the treatment and control groups. Hence DID can be used to control the fact that microfinance interventions are more likely on some types of people, and create similar effects of using PSM. Smith and Todd (2005) have found evidence that a DID approach is more appropriate as an evaluation strategy by replicating Dehejia and Wahaba's (2002) study, in which the authors claimed that PSM results are good approximations to those estimated by experimental approaches. However, there is no conclusive evidence on either side of this debate in terms of the current literature.

Instrumental Variables (IV): The IV approach has the function to control the selection of observables and unobservables simultaneously (Basu et al., 2007). The instruments are a set of variables which influence people's decisions to participate in specific a programme but have no impacts on the final outcomes. Therefore, exogenous is the key for a valid instrument (Caliendo, 2006). Examinations on the qualities of instruments can be done by over-identification tests (e.g., Hansen-Sargan test). Nonetheless, Deaton (2010) has queried the reliability of these tests as he proved that the invalid instruments were able to pass the tests in some cases. In addition, a number of researchers such as Heckman and Vytlacil (2007) argued that Two-Stage Least Square (2SLS) is not always better than the Ordinary Least Square (OLS) especially when the instruments are weak.

2.2.4 Common bias in microfinance assessments

The most common biases that exist in the literature reviewed are introduced in this section. According to the common classification scheme of bias from the Cochrane Handbook for Systematic Reviews of Interventions (2011), the key components of bias can be defined as: 1. selection bias (systematic differences between baseline characteristics of the treatment and control groups); 2. performance bias (systematic differences between groups received different amounts of treatment); 3. detection bias (systematic differences between groups as some are affected by the experiment itself along with the interventions of interest); 4. attrition bias (systematic differences between groups with different numbers of withdrawal members); 5. reporting bias (systematic differences between reported and unreported findings due to selective behaviour).

Regarding the five types of bias, selection bias is of particular importance in studies of microfinance because who engage in microfinance programmes and are successful in business are impossible to have lots of same characteristics as those who do not. This makes microfinance impact assessment an extremely difficult task. In context to the previous sections, it is presented that 93% of the studies included in this review are quasi-experiments and the bias of selection is one of the main threats to their validity.

2.3 Major Empirical Results and Controversies

2.3.1 Significant impacts of micro loans

A large number of individual/household level microfinance impact assessment studies have found significant positive effects of expanding access to finance to the poor. Most of these studies focused on Bangladesh because of the success story of a local microfinance institution - Grameen Bank - has successfully extended credit to more than 2.6 million people to reduce poverty. Pitt and Khandker (1998) conducted a multipurpose quasi-experimental household survey on 87 villages in rural Bangladesh and found that credit is an important determinant of many outcomes, especially for women. By using the same database, Khandker (2005) examined the effects of microfinance on poverty reduction at both the individual level and village level. The results are consistent with the former one and suggest that microfinance can help the local economy.

Besides Bangladesh, similar researches have been conducted in other developing countries as well, while the number of studies is relatively small. Karlan and Zinman (2010) have conducted another survey on 787 marginal applicants (new, rejected, but potentially creditworthy) in South Africa, linked it with loan repayment data, and estimated the impact of credit supply expansion using field experiment. They came to the similar conclusions as Pitt and Khandker (1998). In addition, the marginal loans were found to be profitable for the lenders as well with some evidence. Lensink and Pham (2012) have examined the impact of microcredit on self-employment profits based on a huge sample of 9,189 households in Vietnam. Their findings also reveal positive effects of access to credit on self-employed households. As a representative of cross-country analysis, McIntosh et al. (2011) conducted a field research in Guatemala, India, and Ghana. They estimated the effects of development programmes by the “Retrospective Analysis of Fundamental Events Contiguous to Treatment” method and

found the strongest relationship between credit and household improvement when using the endogenous measure.

2.3.2 Insignificant impacts of micro loans

Nevertheless, some studies such as Roodman and Morduch (2014) have indicated that Pitt & Khandker's (1998) and Khandker's (2005) evidence for impacts are weak and fail in expunging endogeneity. By using a field experiment and follow-up survey that measured impacts of credit expansion for micro-entrepreneurs in Philippines, Karlan and Zinman (2009) found surprising result that creditworthy customer who randomly receive credit shrink their businesses relative to the control group. Expanding access to credit increases profits for male but not for female borrowers. Besides, they found no evidence that increased access to credit improves well-being; rather, they find some evidence of a small decline in self-reported well-being.

Some other studies which go against the findings of Khandker (2005) claim that the microfinance programmes, in reality, have little impact on the poorest or the most vulnerable. Navajas et al. (2000) have analysed the evidence of the depth of outreach for five MFIs in Bolivia with the random sample of 622 active borrowers. They indicated that most of the poor households reached by the MFIs were just near the poverty line – the richest of the poor. By conducting a survey on the 444 households in Thailand, Coleman (2006) has evaluated the impacts of two microfinance programmes with controls on the endogenous self-selection and indicated that wealthier people are more likely to participate than the poor.

On the other aspect, Amin et al. (2003) have assessed the impacts of microcredit programmes on both the relatively poor and vulnerable by surveying 120 households in Bangladesh. They found that microcredit was less successful at reaching the vulnerable comparing to reaching the poor. The contradictions among all research introduced in subsections 2.3.1 and 2.3.2 have produced controversy and confusion for some time.

Some other studies have discovered that the impacts of microfinance programmes can be affected by external factors. For instance, Imai and Azam (2010) have examined whether microfinance reduces poverty in Bangladesh drawing upon the national representative panel data that covers more than 3,000 households with the treatment effects model. It found that simple household access to general loans from MFIs could not increase the household

income, while household access to general loans for productive purposes from MFIs significantly increased household income.

2.3.3 Impacts of interest rate policy and group lending

Besides directly assessing the impacts of microfinance/microcredit, many studies also tried to indirectly assess the impacts in terms of the changing interest rate policy or the group lending method.

By examining in the interest rate on microfinance loans in the slums in Bangladesh with the loan book data of 5,147 clients of SafeSave programme, Dehejia et al. (2012) have studied the price elasticity of credit demand of the poor. It was found that target clients took smaller and more frequent loans and repaid faster as a reaction to the increased interest rate.

Karlan (2007) exploited a quasi-random group formation process with 2,054 loan book records of FINCA-Peru to find evidence to support peer monitoring and joint-liability methods. He indicated that individuals with stronger social connections to their fellow group members have higher repayment and higher savings, as the social connection would deteriorate after default and the method of peer monitoring let the individuals know who should be punished after default. In contrast, Coleman (1999) claimed that most of the group lending impact studies neglected the issues of self-selection and endogenous programme placement and the programme loans have very little impacts according to his findings.

2.3.4 Impacts of access to finance

Analysing the determinants of access to finance (or creditworthiness) is another interesting area that has attracted many researchers in the recent years. Johnston and Morduch (2008) used a survey including 1,438 households in Indonesia to analyse the prospects for expanding financial access. They found that about 40% poor households were judged creditworthy according to the criteria but fewer than 10% borrowed from a micro-bank or formal lender.

Some studies also try to identify the relations between gender difference and access to credit. Agier and Szafarz (2013) have investigated whether men and women benefit from the same credit conditions by establishing their original model and testing its predictions on a loan book data comprising more than 34,000 applications from an MFI in Brazil. A loan size gender gap was detected, and it would increase disproportionately with respect to the scale of potential projects.

There are many other interesting combinations of topics that worth studying, such as Beisland and Mersland's (2012) recent study that investigated the use of microfinance services among economically active disabled people in Uganda. However, these non-mainstream studies usually have little empirical supports behind them.

2.3.5 The issues of data quality

In terms of individual/household level studies of microfinance impact assessment or access to microfinance, high-quality data is the determining factor as it is extremely difficult to acquire. Unlike enterprise level microfinance studies that often use similar databases such as Mix-Market and Microcredit Summit, the data used at the individual level are always distinctive in each paper.

First, most of the data are related to specific small areas, provinces, villages and etc., such as the local field experiments conducted by Karlan and Zinman (2009; 2010), and Roodman and Morduch (2014). Bangladesh has attracted the attention of many researchers while the studies about other countries are scarce. Cross-MFI analyses are so rare that it is difficult to find any relevant papers.

Second, individual/household surveys are widely used while personal loan book data are used in very few studies: Storey (2004), Alesina et al. (2008), Bellucci et al. (2010), Agier and Szafarz (2013) and etc. All these loan book studies concentrate on the subject of access to finance but impact assessment. As the loan books are provided by specific MFIs with details of an enormous number of clients (sometimes greater than 50,000), the quality of data used in these papers is very high. The largest problem of loan book data may be the difficulty to generate a "without programme" control group for impact assessment.

Third, most individual/household surveys only covered a small number of respondents. Except for a few studies such as Lensink and Pham (2012) that interviewed over 9,000 households, the sample sizes of the majority of surveys are less than 500. For instance, Park and Ren's (2001) and Coleman's (2006) survey data only covered about 450 households. Insufficient participation of microfinance in remote areas heavily limits the sample size. However, with the capabilities to specialize in the unique natures of each treated group and to establish control groups, survey data is more suitable for impact assessment than loan book data. Combing the two types of data to support more complete analyses might be a trend for further studies.

2.4 Limitations of the Current Literature

In the following three sections, summary evidence from the studies and papers reviewed are organised and discussed by their fundamental research designs as presented in 2.2.2. As the amount of related literature is enormous, and many of them provided very similar results, it is impossible to talk or even mention about every single paper. To make this review as inclusive as possible, the studies presented below are selected by their influence in the area of microfinance impact assessment such as the series of Pitt and Khandker and those contributed by USAID. Sections 2.4.1 and 2.4.2 discuss the quasi-experimental and the experimental results respectively. The studies focusing on women empowerment are presented in the extra section 2.4.3.

2.4.1 Quasi-experimental studies

The results of the quasi-experimental studies are separated and discussed in two sub-sections: the with/without studies and the pipeline studies. Broadly speaking, the with/without studies have assessed a higher proportion of impacts on the later stage of the causal chain that highlighted by blue and purple in **Figure 2.1**, comparing to the studies using pipelines. Most of the economic outcomes of the with/without studies (by the IV methods) are significant and more likely to be positive. Nevertheless, there are few significant outcomes on the social side and on women's empowerment. On the other aspect, the vast majority of the outcomes assessed in pipeline studies are insignificant.

2.4.1.1 With/without and before/after studies

This section starts by briefly introducing the two iconic studies of Pitt and Khandker and USAID, and then discusses the influential studies developed based on them.

The Pitt and Khandker series of studies in Bangladesh (1998, 2002, 2003, 2006, and 2011): The fundamental cross-sectional data in these studies were collected from a survey conducted in 1991-1992 on three group lending programmes in Bangladesh. The survey included 87 villages and 1,789 households in rural areas. Labour supply, enrolment of education, expenditure per capita and non-financial assets were the main indicators. In 1998-1999, for the purpose of investing long-term microfinance impacts, Khandker resurveyed the same households and surveyed another 810 households from both the original and new villages in the original thanas.

Based on this data, Pitt and Khandker used a quasi-experiment to sample the targets. According to whether a target is living in the village with microfinance programmes and whether a target has a choice to join the programmes, all households were split into four sub-samples. By running the IV regressions, Pitt and Khandker found that microcredit has significant and positive influences on the indicators shown in the last paragraph. They stressed that larger positive influences were found when female clients were involved in the programmes. As an extension to the findings, Khandker re-examined the results with the 1998-1999 data and found that the impacts of microcredit on poverty reduction were sustainable in the long-run. Moreover, positive spill-over effects were found at the village level.

However, many associated studies of Pitt and Khandker's original data have failed to replicate the same findings, probably because of the complication and poor documentation of research design. Instead of the IV approach, Morduch (1998) and Roodman and Morduch (2014) have applied PSM in the re-examination. They found a contradictory result that there were hardly any impacts and argued that Pitt and Khandker have overestimated the impacts because the criteria for eligibility were not strictly implemented. Slightly different to Roodman and Morduch's study, Chemin (2008) has found significant and positive impacts for half of the outcomes by using PSM on the same data, though the results were lower than Pitt and Khandker's findings and the impacts on the other half were almost negligible.

Moreover, a number of studies such as Duvendack (2010) and Duvendack and Palmer-Jones (2012) indicated that the situation of multiple sources of borrowing had not been considered in the papers discussed above. By using a strategy named 'novel treatment' to obtain more homogeneous control groups, Duvendack found mixed results when he compared microcredit participants with who accessed to other sources of credit. Venkata and Yamini (2010) have pointed out that many microcredits were often too small to cover the costs of micro-entrepreneurship and multiple sources of borrowing help to smooth the borrowers' income and consumption. As another explanation, Coleman (1999) and Fernando (1997) indicated that it is common for debtors to use borrowing from one source to repay the loans of another. Based on these views, both the Pitt and Khandker studies and the associated criticisms are not convincing enough. In addition, using PSM to replicate the Pitt and Khander studies is doubtable, in terms of the limits that already discussed in 1.3 (requiring high-quality data). Further discussion on the analysis methods is beyond the scope of this section.

The USAID studies in India, Peru and Zimbabwe: The target of the three longitudinal studies was to evaluate microfinance impacts on the poor. All panel datasets were collected by USAID in the late 1990s. The data at all three different levels (individual, household and firm) were included in the studies. Similar to the issues encountered in the Pitt and Khandker data, the robustness of USAID's selection procedure of control groups is questionable. Some unobservable characteristics which account for why the eligible individuals/households did not participate in microfinance programmes made the sampling of USAID less convincing.

As the earliest studies using the USAID panel data, Chen and Snodgrass (1999, 2001) found evidence that microfinance led to changes at the household level and detected positive impacts on income, income per capita, income diversification, expenditure on food and resistance to shocks. However, the results at the individual and firm levels were insignificant. As explained by Chen and Snodgrass, a possible explanation was that most clients of the microfinance programmes were workers instead of entrepreneurs. By using PSM and DID to reduce the selection bias, Augsburg (2006) and Duvendack (2010) re-examined and broadly confirmed Chen and Snodgrass' findings. However, Duvendack also pointed out that the matching estimates (of PSM) were very sensitive to the selection on unobservables. Microfinance participants might have been superior to non-participants long before joining the programmes, in terms of social networks, wealth and skills (Armendariz and Morduch, 2010). In brief, the re-investigation of the USAID studies (and Pitt and Khandker's studies) have greatly weakened reliability of the empirical support for microfinance's poverty reduction function.

2.4.1.2 Pipelines studies

Coleman (1999, 2006 and etc.) was the very first researcher who tempted to apply pipeline designs in microfinance impact assessments. Since then his method has been widely used. He conducted the surveys on 455 households in North-eastern Thailand during the period 1995-1996. Self-selection and non-random programme placement bias were controlled by observable village-level fixed effects. The 1995 data were related to the participants and non-participants in villages where microfinance already activated, and the 1996 data identified potential participants and non-participants in villages where microfinance was planned to operate. By using DID to estimate the difference in different incomes between participants and non-participants with village controls, Coleman has found little impact of micro-

finance. Moreover, he concluded that micro-finance had positive impacts on increased borrowing activities and debts because it was discovered that many participants joined the programmes for consumption purposes instead of entrepreneurship or investment.

The second important series of studies was conducted by Copestake et al. (2001, 2002 and 2005). The 2001 paper reported microcredit impact in a group liability context in Zambia using two cross-sectional sample groups and a pipeline group. Only a number of hardly statistically significant outcomes were found. This finding was vitiated due to the high exit rate of the sample group clients between different loan cycles. As an improvement, the 2002 paper involved continuing borrowers to eliminate the problem of exits, drop-outs and graduates. Some initial levelling up effects on business income was found, but the microfinance impacts on the other variables such as business profit and transfers to the household budget were polarized. Different from the two previous studies, the 2005 paper estimated impacts in a basic DID model and a multivariate model by using panel data from Peru. The results suggested that the programmes have significant effects on individual and household income (more for richer than poorer ones), but no effects on business sales and profit.

Colman's and Copestake's studies are remarkable for two reasons, the very large number of assessed variables and the relatively slim and straightforward econometric analysis method – using DID without lots of control variables or 2SLS. In fact, Steele et al. (2001) have done something very similar to Colman's 1999 paper, but the more sophisticated methods included in that study made it harder to replicate. There are few studies that have applied other analysis methods besides DID. One noticeable example trying is that Setboonsarng and Parpiev (2008) applied PSM to pipeline data in the expectation that it would provide higher robustness. They did it, but at the cost of a dramatic loss of an important part of the data: a great number of participants (with low propensity scores) for whom there were lots of potential matches had been dropped. This was a preposterous basis to undertake further impact analysis and obtain convincing results. It showed us that excessive pursuit of statistical precision has an adverse effect on the pipeline studies of microfinance impact assessment, in which there were usually tons of unobservable.

Regarding the results that found in the other pipeline studies, most of them are very similar to Colman's findings: 1. microfinance has significant positive influence on the early stage (**Figure 2.1**) outcomes such as borrowing and business activities; 2. it has no statistically significant influence on the variables of well-being. All these studies have provided evidence that the earlier impact assessments made by other analysis methods were overoptimistic. In

addition, some of them argued that the clients of microfinance services were the riches of the poor instead of the very poorest ones. In contrast, there are very few papers that suggested significant positive effects. As a representative, Deininger and Liu (2013) tested a large number of variables with unique econometric specifications and detected a significant positive influence on the well-being of clients. However, their study is vulnerable to bias as the treatment and control groups have different locations.

The biggest constraint of the pipeline method is by the nature of itself that there is only a tiny period of time within which the treatment group and pipeline group can be considered to be different. Therefore, the impacts estimated by such method may only be effective in the short-term, while the majority of social influences are likely to be observable only in the long-term.

2.4.2 Experimental studies (RCTs)

While RCTs are recognised as the most robust methods for impact assessments in the development industry, the full potentials of RCTs are still waiting to be explored, and very few rigorous studies about the impact of access to microfinance relative to no access are found and included here. This section begins with the introduction of some details and threats to the validity of the essential papers and then discusses their findings.

As claimed by the authors, Banerjee et al. (2015) have conducted the first randomized experiment of the impact of introducing microfinance to a new market. The panel data used in this study was a subset of 104 slums (approximately 65 households in each of them) at the southern Indian state Hyderabad, where Spandana (an MFI that focuses on self-formed female borrowing groups) considered to select some areas for opening branches randomly. The baseline survey and end-line survey were conducted before and subsequent to the randomisation respectively. It is not clear that whether the selection of the baseline survey participants has been randomised (Type 1 bias, see 1.4). The second threat to validity was the possibility that potential participants in the control slums postponed business expansions because they expected for low-cost loans from Spandana in the near future (Type 3 bias) rather than pure commercial consideration. On the other hand, the risk of attrition bias was low, and the spill-over effect has been accounted by acknowledging the entrance of other MFIs in the sample slums.

Another fascinating series of RCT studies are conducted by Karlan and Zinman (2009 and 2010). Only the 2009 paper is presented in this review as both studies have taken a very similar approach. The authors used a field experiment and follow-up survey to measure impacts of credit expansion (by the First Macro Bank in Manila) for micro-entrepreneurs with mid-creditworthy (A credit scoring software was used by the MFI to render disposition based on applicants' household and business information. 31 and 59 were the automatic rejected and approved thresholds. Decisions for who scored 31-59, the mid-creditworthy applicants, depended on the MFI's loan officers' judgement). These applicants were randomly assigned to the approved (intent-to-treatment) groups with 60%, and 85% approval rates and the rest were assigned to the rejected (intent-to-control) groups for further assessments. The term 'intent' means that loan officers did not always make the offers as instructed by the software though it was highly possible. Randomisation might not be well achieved because of the loan officers' selections on unobservable information (Type 1 bias). Moreover, loan officers may dissimilarly treat the clients and paid extra attention to the mid-creditworthy clients who received loans compared to the high-end clients. This is also a potential threat to validity (Type 2 bias). Besides, a less creditworthy client who accepted the offer might attribute his/her success to the reason of being surveyed and altered behaviours accordingly (Type 3 bias). The issues of attrition bias and spill-over/in effect are unclear in this study as there is no evidence about how characteristics affected the attrition rate (30%) and whether the other MFIs have influences on the participants.

Putting aside the highlighted research designs used by these studies, very little significant impacts of microfinance were founded on the well-being outcomes. By testing a large number of variables, Banerjee et al. (2015) founded no discernible effect on education, health and women empowerment within the 15-18 month time period while the effect on household expenditure and expanding business was significant. It can be regarded as strong evidence that microfinance has no short-term impacts on well-being, which is a popular interpretation in the Economist. The findings of Karlan and Zinman (2009) were a bit more complex. They found some evidence that the borrowing amount and profit of clients did increase after participating in the microfinance programmes. However, they appeared to shrink their businesses by shedding unproductive employees. The effects of treatment were stronger for male and higher-income entrepreneurs. Besides the fact that borrowing households substituted away from labour into education, no evidence of significant increases in well-being was identified. In summary, the contribution of the previous RCTs studies was very limited, probably because of the intention-to-treat basis used in estimations and the

ignorance of spill-over/in effect. A potential solution would be to replace the mainstream well-being indicators by those in the earlier stages of **Figure 2.1**. Nonetheless, the unproved casual relationships between well-being and the early indicators may create another thorny problem.

2.4.3 Women Empowerment Studies

The issue of women's empowerment, which is one of the primary missions to introduce microfinance, has been addressed in many studies that mentioned previously. In terms of the with/without (before/after) research, the Pitt and Khandker serial studies (see 3.1.1), the papers developed based on their data and methods, and the serial studies contributed by USAID have tried to investigate this issue. All this literature used an indicator named 'household-decision-marking' as the major proxy for empowerment. The underlying data for such variable were simply collected by asking the participants if they considered themselves able to control or affect the household expenditure. While mixed results have been found in these quantitative studies, their validity is doubtful because of lack of precise empowerment measurements.

On the other hand, a great number of qualitative studies have found evidence that the perception of the female microfinance participants did change in their communities, and they were more involved in household and community decision-making. All this evidence, however, should be regarded as 'stories' because most of them were based on sample surveys or case studies, such as the studies of Goetz and Sen Gupta (1996) and Hashemi et al. (1996). Besides decision-making, the qualitative studies also used a wide range of variables, such as mobility, economic security, the freedom from family's domination, and participation in social and political life, to proxy the empowerment and investigate the microfinance impact in Bangladesh. These indicators, again, might lack credibility, because the relationships between women empowerment and them have yet to be proven.

In terms of the studies based on other research designs (pipeline or RCTs), Deininger and Liu (2013) are the only authors who found positive impacts on empowerment by examining a self-help group microfinance project in India using pipeline design. This study is, however, vulnerable to selection bias and the evidence are untrustworthy. As one of the few RCT studies, Banerjee et al. (2015) could not find any noticeable microfinance impacts on empowerment within 15-18 month time period of study. The authors themselves also pointed out that such a short period may be insufficient for the long-run influences to reach observable

levels. In addition, the statistical power of this study was limited by potential selection and detection bias (see 3.2).

2.5 Conclusions and Implications

This review has investigated the studies included by comparing their methodologies, results and potential biases in detail. Most of the assessed microfinance impacts are found to occur in the early stages in the casual chain (**Figure 2.1**). The studies that focused on the later stages were insufficient. Moreover, the majority of findings were statistically insignificant. It is also remarkable that a number of studies have detected significant negative influences. These results are consistent with some studies on the qualitative side.

By comparing and analysing different methodologies used in the previous literature, four implications for further research can be concluded as follows:

1. The indicators and measurements of microfinance impacts need to be more precise, and greater standardisation of them is necessary. Besides, researchers have to carefully consider if there are any potential long-term effects which may not reveal themselves in short experiment periods when using the social indicators.

2. Because the current evidence base for the impacts casual relationships is small, studies that are focusing deep on specific stages (**Figure 2.1**) are more necessary than those simply link microfinance to the final-stage indicators of well-being.

3. More studies should be implemented on different research designs, especially on the well-designed RCTs which use validated impact indicators, in order to reduce the systematic risks of bias and provide more convincing evidence.

4. At last, further comparisons between individual lending and group lending and between female and male clients are also needed in the new studies, instead of simply focusing on the female group lending as the current literature.

Chapter 3:

Delinquency of Microfinance:

A Two-part Probit Analysis of Cross-MFI Data

3.1 Introduction

Microfinance Institution (MFIs) can be defined as any financial institutions which offer not only loans to Micro, Small and Medium-sized enterprises (SMEs), groups and individuals, but also other financial services like savings, insurance, and investment advice including training programmes to their clients. There are more and more international organisations coming to the realisation that Microfinance Institutions (MFIs) are veritable and effective channels to improve the effectiveness of poverty alleviation programmes in developing countries (Okumadewa, 1998). According to Chossudovsky (1998), the World Bank Sustainable Banking with the Poor project (SBP) in 1996 has estimated that there were more than 1,000 MFIs in over 100 countries, and each MFI has a minimum of 1,000 members and with 3 years of experience.

The issue of loan delinquency among MFIs has been discussed in many previous studies and considered as the primary reason why commercial banks have not shown much interest in financing SMEs. According to Balogun and Alimi (1988), loan delinquency can be defined as the inability of a borrower to fulfil his/her loan obligation when instalments are due. Because of the unintended negative impacts on financing, the high frequencies of loan delinquency in SMEs lending should be of major concern to policymakers in developing countries. In fact, MFIs in developed countries are faced with the same challenge of loan repayment.

The chance that a lender does not receive its money (plus interest) back from borrowers is the most common and often the most serious vulnerability in the MFIs (Warue, 2012). Since most loans are unsecured, delinquency can rapidly spread from a few loans to a significant portion of the entire portfolio. This contagious effect will be strengthened by the fact that microfinance portfolios often have a high concentration in a small number of business sectors such as agriculture and food retail. As a result, borrowers may be exposed to the same

external threats such as lack of demand, livestock disease outbreak, bad weather and etc. These factors create volatility in loan portfolio quality and heighten the importance of credit risk control. In this regard, MFIs need a monitoring system that highlights repayment problems clearly and quickly, so that loan officers can focus on the delinquency of clients before it gets out of hand.

The sustainability of MFIs highly depends on their ability to collect their scattered loans as efficiently and effectively as possible. In other words, to be financially viable, MFIs must ensure high portfolio quality with a repayment rate closed to 100%, or at worst low default and cost recovery. In recent years, there have been more complaints by MFIs regarding the high default rates of their clients. Loan delinquency and hence default has started spilling over deeply into the operations of MFIs in developing countries.

A feature of many loan delinquency models which have been frequently used in prior empirical studies, such as straightforward binary or censored data models, is that the process which results in non-delinquency is strongly assumed to be the same as the process which determines the intensity of delinquency. For instance, if a borrower characteristic has a significant and positive effect on the intensity of delinquency, then a high value of this characteristic will inevitably lead to the prediction of being-delinquent for this client. Such an assumption might fail when there is a proportion of the population of borrowers who will never default under any conditions. There is no reason for us to expect this assumption a priori. In addition, the information loss is severe as we dichotomize the delinquency data into binary format. These considerations lead us to a class of model in which the probability and intensity of events can be estimated separately. This type of model is known as the 'Double-Hurdle' model which is proposed by Cragg (1971). The model assumes that a borrower must cross two hurdles in order to be delinquent. Borrowers who fall at the first hurdle are referred as 'never-delinquents' in this study.

On the other hand, in terms of the current literature, most of the empirical studies only focus on the effects of business characteristics of micro-enterprises and the credit policies of a single Microfinance Institution. Empirical studies related to individual-level socio-geographic characteristics with high-quality cross-MFI data is scarce. Moreover, the indicator of delinquency or default is usually dichotomized to a dummy variable in the prior studies, and it leads to severe information lost. With a unique cross-MFI loan book data which have never been used in other studies before, we try to analyse the determinants associated to loan de-

linquency among the borrowers based on the Double-Hurdle models in this chapter. Our results show that the Two-Part Model can be applied to heavily skewed loan book data, and implemented in establishing new credit scoring systems. We also found that engaging in agriculture is generally related to lower probability and intensity of being delinquency in terms of arrears. It indicates that governments and MFIs should provide greater supports for poor farm households and farm-related business.

The rest of this paper proceeds as follows: Section 3.2 reviews the literature related to loan delinquency and loan default. Section 3.3 describes the summary statistics of data. Section 3.4 presents the theories, models, and transformation techniques for estimation. Section 3.5 reports the empirical results. Conclusions and discussion are presented in the final section.

3.2 Literature Review

3.2.1 The concepts of loan delinquency and loan default

A loan is delinquent when an instalment payment is late. Delinquency is measured because it indicates an increased risk of loss, warnings of operational problems, and may help to predict how much of the portfolio will eventually be lost because it never gets repaid. There are three broad types of delinquency indicators: 1. collection rates which measure amounts actually paid against amounts that have fallen due; 2. arrears rates which measure overdue amounts against total loan amount; and 3. Portfolio at Risk in a certain period of time, which measure the outstanding balance of loans that are not being paid on time against the outstanding balance.

Loan delinquency becomes loan default as the chance of recovery becomes minimal. By definition, loan default occurs when the borrower does not make required payments or in some other way violate a loan covenant (conditions) of the debt contract (Ameyaw-Amankwah, 2011; Murray, 2011). The potential reasons for loan default can be either objective (unable to repay), or subjective (unwilling to repay), or more realistically a combination of both of them. In this study, ‘delinquent’ and ‘default’ have the same meaning and will be used interchangeably.

Moreover, Pearson and Greeff (2006) refine the standard of loan default as a risk threshold that describes the point in the borrower’s repayment history where he or she missed at least three instalments within a 24-month period. This represents a point in time and indicator of

behaviour; wherein there is a demonstrable increase in the risk that the borrower eventually will truly default, by ceasing all repayments. This definition is consistent with international standards. It is necessary because consistent analysis required a common definition. Such definition does not mean that the borrowers had entirely stopped paying the loans and therefore been referred to collection or legal processes, or from an accounting perspective that the loan had been classified as bad or doubtful, or actually written-off.

3.2.2 Determinants of loan delinquency or default in microfinance

A study conducted by Okorie (1986) in the Ondo State of Nigeria indicated that the repayment ability and consequently high default rates are associated with nature, time of disbursement, supervision and profitability of enterprises. Other critical factors contributed to loan delinquencies include: interest rate, type of loan, term of loan, borrowers' income, poor credit history, and transaction cost of the loans. According to another study conducted by Ahmad (1997), causes of loan default also include: lack of willingness to pay loans that coupled with the diversion of funds by borrowers, intended negligence and unsuitable appraisal by credit officers. Similarly, Kohansal and Mansoori (2009) considered that most defaults arose from an unwillingness to repay loans, loan diversion, and poor management procedures. According to their study, the most important factors that led to loan delinquencies include: interest rate ceilings imposed by the government, monopoly power in credit markets exercised by informal lenders, large transaction costs incurred in loan applications, moral hazard, and many more.

3.2.2.1 Main findings related to individual socio-geographic characteristics

Marital status is a very common variable in the default - repayment relevant literature. It is often considered a sign of responsibility, reliability or maturity on the part of borrowers. The relationship between the borrowers' marital status and loan repayment performance remains controversial. We can expect that the probability of default payment is higher for singles than married individuals. More often than not, single borrowers tend to be less responsible (Dunn and Kim, 1999; Vogelgesang, 2003). By analyzing U.S. consumer loans, Avery et al. (2004) suggest that married individuals are less likely to default compared to those who have never been married, because they may have a second income to rely on in case of unemployment or illness. Similarly, Kocenda and Vojtek (2009) indicate that married borrowers have a lower default rate in Czech retail banking. We should be aware that the assumption of a reliable secondary income of the spouse may not be plausible in developing countries. But Vigano (1993) and Vogelgesang (2003) do find that being married is a sign of financial

stability in developing countries. As another explanation, Sharma and Zeller (1997) indicate that borrowers with children do not wish to risk the privileges combined with the repayment of loans.

On the other hand, Dinh and Kleimeier (2007) claim that the probability of default is higher for married than single borrowers as the former are generally related to a greater number of dependents (such as children), which in turn reflects a financial pressure on a borrower's ability to repay a loan. Bandyopadhyay and Saha (2011) indicate that the risk of default will increase as the number of family members of the borrower increases, while a secondary income does lead to a lower default probability. According to the study of Ugbomeh et al. (2008), we can see that household size affects the loan repayment in a negative way, and a greater family size might induce the borrower to use the loan for unintended consumption.

But in this study, we still expect the positive effect to outweigh the negative one and therefore married clients to have a low probability of default.

H1. Married individuals have a lower probability of default and lower intensity of delinquency.

Concerning the associations between age and the repayment of a micro loan, the evidence is ambiguous. In the context of Vietnamese retail banking, Dinh and Kleimeier (2007) found that default rates increase steadily with age. Regarding the Indian housing loans, Bandyopadhyay and Saha (2011) have come to similar results. They found that younger borrowers are less likely to default on their loans than older ones. For these findings, there are three major explanations: 1. it can be assumed that younger borrowers are more independent, free from financial burden such as education expenditure of children, and will, therefore, be less likely to default; 2. older borrowers may already have one or more loans and overstretch their financial capabilities; and 3. borrowers in a high age bracket have fewer service years left and a limited ability to reduce financial constraints.

On the contrary, a number of studies indicate that probability of loan default is negatively related to age (Arminger et al., 1997; Dunn and Kim, 1999). Vogelgesang (2003) and Van Gool et al. (2012) found that age has a risk-reducing effect. Besides, Vigano (1993) also assumes that a higher age is a symptom of stability of finance, and it leads to a reduction of default rate in developing countries. In reality, it is often assumed that older borrowers are usually wiser, more risk averse, more knowledgeable, and more responsible than younger borrowers and therefore, will be less likely to default. As another possible explanation,

Reinke (1998) argues that older borrowers are less likely to look for better employment opportunities than younger people. Older clients rely heavily on their loan-supported businesses and are therefore less likely to fail on the repayment of a loan.

While the previous studies do not suggest a clear trend, we argue that age might have a non-monotonic effect on repayment rates. We can expect that the youngest and oldest groups of borrowers to have the highest repayment rates, while the middle-aged consumer group would have the lower repayment rates.

H2. The youngest and oldest borrowers have a lower probability of default and a smaller intensity of delinquency, while the middle-age group of consumers have a higher probability of default and a larger intensity of delinquency.

Based on the literature, it has been claimed that women demonstrate much better repayment behaviour in terms of microfinance is one of the most discussed facts (Dinh and Kleimeier, 2007; Roslan and Mohd Zaini, 2009; Salazar, 2008; Schreiner, 2004; Viganò, 1993). They default less frequently on loans probably because they generally enjoy the hard - work ethic and the culture of financial discipline (Bhatt and Tang, 2002; Pitt and Khandker, 1998). As another explanation, repayment rates may be expected to be higher for women simply because they are more likely to choose relatively less risky projects (Sharma and Zeller, 1997). Croson and Gneezy (2009) also suggest that women are more risk-averse compared to men.

The Food and Agriculture Organisation of the United Nations (2005) has conducted an in-depth analysis of female farmers in Nicaragua. According to this investigation, a great proportion of women in the northern regions of Nicaragua live in the role of being a housewife and mother. The local culture does not consider women to be professional farmers. Hence, women are often excluded from training, networking, and consultancy. In order to acquire income and feed their families with fewer opportunities and resources, the local women were more dedicated to the agricultural projects available for them.

H3. Female borrowers have a lower probability of default and a smaller intensity of delinquency.

In classical banking, a higher level education indicates a lower probability of default (Kocenda & Vojtek, 2009). Better educated individuals would have a higher ability to understand and analyze complex information, and have higher entrepreneurial social competence enabling him/her to make the right business decisions (Bhatt and Tang, 2002). On the other

side, whether such relation is consistent in the microfinance industry is doubtful. The first reason is that most of the clients who use micro loans participate in business activities that require very little knowledge but working experience and skills. For instance, it is sensible to assume a weak relationship between agricultural production and middle school education. The second reason is that better-educated borrowers have less difficulty to access to other sources of credit. Therefore, Borrowers with very limited education may highly depend on the micro loans and thus more stable. Nevertheless, as the mainstream empirical results about the associations related to education are positive, in this paper, we just keep our hypothesis like the ones in the prior studies for better comparison.

H4. Borrowers with higher educational levels have a lower probability of default and a smaller intensity of delinquency.

3.2.2.2 Main findings related to business characteristics

By surveying different formal banks in India, Berger and De Young (1995) identified the main causes of loan delinquencies from the industrial sector. These include an improper selection of entrepreneurs, deficient project viability analysis, inadequate collaterals against loans, inappropriate schedule of loan repayment, lack of follow up measures, and default due to natural disasters. Similarly, Sheila (2011) also stressed that inadequate financial analysis is a crucial cause of loan default. It happens when the officers in the loans department do not take a careful study of the applicants to ensure that they have sound financial bases such that the risk of loan default can be mitigated. Besides, he pointed out that in Uganda, the issue of inadequate loan support is another cause of loan default, and it is very important that the loan officers collectively ascertain the positions in which the borrowers find themselves. However, that was not the case, and the given support was irrelevant to which leaves the business crumbling and leads to loan default.

On the other hand, Sheila's (2011) study also pointed out that illiteracy and inadequate skills are another causes of loan default. A large proportion of borrowers are engaged in traditional and low paying businesses which are rarely diversified. It implies that they did not have enough alternative marketable skills that can benefit them as their current businesses do not function properly. In addition, most of them have no idea how to read, write, and make simple calculations. As a result, it was very difficult for the borrowers to account for their businesses when the lenders made mistakes, and they were held liable for the loan.

Poor business practice is yet another cause. According to Gorter and Bloem (2002), non-performing loans were usually caused by an inevitable number of wrong business decisions by borrowers and plain bad luck (unexpected price changes for materials, bad weather, etc.). Under such circumstances, the holders of loans can make allowances for a share of non-performance in the form of bad loan provisions. Alternatively, they may spread the risk by taking out insurance. Similarly, Kasozi (1998) indicated that there are considerable weaknesses of the borrowers over which the lenders have very little control. Business management is an essential part that needs to be emphasized. He found that many borrowers lack the technical skills such as keeping records and checking on business performance. Most borrowers never plough back the profits into business, and it leads to loan default in the long run.

The literature on SME loans in developing countries appears sparsely populated. The study conducted by Munene and Guya (2013) in Kenya shown that one of the causes of loan default is the characteristics of the business. Their study shows that probability loan default is extremely high (67.9%) in the manufacturing sector. This is followed by that of the service sector (64.0%), and then by the agriculture sector (58.3%). In comparison, the retail sector records the lowest loan default rate (34.9%). This could be attributed to the observation that the retail sector deals with fast moving products on high demand, which could transmit into good business performance and increase revenue that accounts for lower default rate. Using the dataset of a commercial MFI in Tanzania, Weber and Musshoff (2012) found that agricultural firms are less often delinquent when paying back their loans than non-agricultural firms. A possible explanation of these results is that agricultural firms face higher obstacles to access to credit. According to Baesens et al. (2011) and Viganò (1993), agriculture is assumed to be the safest sector due to the higher social control and typical lower volatility. Services and small trades are assumed to be positively related to the categories with high default risk owing to their inherent volatility and their dependence to a certain degree of technology.

However, Fidrmuc et al. (2010) have studied the loan default rates of 700 SMEs in Slovakia for the period from 2000 to 2005. They found that the default rates clearly differ between business sectors, and the service and agriculture sectors have higher probabilities of default than manufacturing, retail, and construction sectors. In another study with the dataset of an MFI in Madagascar, Weber and Musshoff (2013) indicate that bad weather conditions, such as an excessive amount of rain in the harvest period, will increase the default probabilities of loans granted to small-scale farmers. It seems that there are no consistent results on agriculture yet.

H5. Credit destined to the agricultural sector has a lower probability of default and a smaller intensity of delinquency.

3.3 Data

The raw data for four MFIs from different countries located at South America and Sub-Saharan Africa have been extracted from administrative loan books gathered by Micro Finanza Rating, which is a private and independent international rating agency specialized in microfinance. It contains two MFI types, with three NGOs (CACIL Honduras, INSOTEC Ecuador, and FINCA Peru) and one cooperative (MICROCRED Madagascar). All loan books were compiled by the MFIs and submitted to Micro Finanza Rating between 2010 and 2011. As the percentages of missing values are very low, the impact brought by missing values is marginal. Hence, a simple listwise deletion approach is applied here. On the other hand, the occurrence of outliers in the data used for this papers is limited. With no signs of correlated outliers, simple winsorizing and trimming (Wainer, 1976) are adopted. All the observations of loan amount under (or above) a 5% (95%) percentile are replaced by the limits. The data of age, time to maturity, arrearage, and the length of delayed repayment are trimmed in the same way with different percentiles (see footnotes of **Table 3.1**). In order to represent the actual population proportions of different countries (UN World Population Prospects, 2010), the raw data is also processed by weighted random selection. In the end, our sample consists of 32,673 clients. Ecuador, Honduras, Madagascar, and Peru, take up 21%, 11%, 28.4% and 39.6% in the sample respectively.

It should be mentioned that the analysis in this study is based on clients with approved loans only. The standard loan approval processes applied by the MFIs are unknown, and no generalisations can be made for a random sample of all microfinance applications. The issue of obtaining the default risk profile of rejected applicants is called reject inference. In general, absolute reliable reject inference cannot be achieved (Hand and Henley, 1993). Besides, the effect of sample election problem in credit portfolios depends on the rejection rate and becomes influential only when the rejection rate is extremely high (Crook and Banasik, 2004). Therefore, the models developed in this study are applied to the borrowers who have been approved by the four MFIs in our sample, and the problem of rejection inference has been assumed to be negligible.

There is no universally accepted approach to choose the explanatory variables for credit scoring (Dinh & Kleimeier, 2007). The explanatory variable choice in the case study presented in this paper is based on prior studies and expert advice from the microlender staff. All explanatory variables used in this study can be classified into two categories: 1. socio-demographic characteristics (marital status, gender, age, and education level); and 2. loan purpose (consumption, buy a fixed asset, agriculture, commerce, manufacture, service, and financing). The control variables include two categories as well: 1. loan status (loan amount and time to maturity); and 2. MFIs.

This is an unusually short list; most scorecards for microfinance institutions would also use the income, occupation, and the number of dependents; ownership of a phone, house, or car; and measures of the size and financial strength of the business. Therefore, the research in this paper is conservative and mainly focusing on the variables stated in the hypotheses. If a scorecard with these characteristics works, then a scorecard with a full complement of characteristics on the borrowers would work even better.

For the dependent variable of loan default or delinquency, authors often need to create their own proxies when the rating agency requires a specific variable or the required data is not directly available. For example, Schreiner (2004) defines a bad customer to be 15 days late on the repayment, Vogelgesang (2003) characterizes default loans by an average of 10 days overdue per payment, while some other authors like Van Gool et al. (2012) focus on late repayment that indicated by an average two days late on the installments. As results, the findings in these studies become incomparable, and their practical implementations are limited.

In order to standardize the measurement of loan default, there are three variables used in this paper: 1. the current amount of arrears; 2. the number of days of delayed repayment; 3 Portfolio at Risk (PaR), which is calculated by dividing the outstanding loan amount with arrears over a particular period (e.g., PaR30 denotes Portfolio at Risk over a 30-day period), plus all restructured loans, by the outstanding gross portfolio as of a certain date.

However, occasional late payment of a few days does not constitute a problem to MFIs, and our rigorous definition may lead to overestimation of default risk. To solve the issue, a separate regression analysis is necessary to capture the intensity of loan default. Therefore, the Two-Part Model and the Double-Hurdle Model (which will be discussed in the next section) should be applied in this case.

Table 3.1 summaries the mean, standard deviation, minimum, maximum, and quartiles, for all variables in our sample. We see that the average outstanding loan amount is $970 and the average time to maturity is 317 days. Meanwhile, these two numbers rise to $1360 and 457 days respectively in the subsample of clients with non-zero arrears. It indicates strong associations between the outstanding loan amount, time to maturity, and the probability of default. Comparing the medians and means, we also see that the outstanding loan amount is heavily skewed to the right. 25% of our clients borrowed less than $232 and 50% of them borrowed less than $580. These statistics well describe the primary mission of microfinance – providing small loans and saving facilities to those who are excluded from commercial financial services.

As far as the repayment variables are concerned, the dummy of late payment equals 0.07 on average, which means that 7% of clients are delinquent or have defaulted. For these clients specifically, the arrearage is $238, and the length of delayed repayment is 230 days on average. By measuring the delinquency loans based on the standardized schedule RC-N, the delinquency rate of the subsample[1] including CACIL and FINCA is 2.65% in 2010.

For socio-demographic variables, we see that 52% clients are married and 14% of them are cohabiting with their partners. The rest of them are either single, divorced, or widowed. The statistics show that the probability of loan default might associate with marriage and single status to some extent. The proportion of married clients for the defaulted group is 7% lower than that for the normal group, and the proportion of single clients for the defaulted group are 12% higher than that for the normal group. On average, 65% of clients in MFIs are women. The 75th percentile in the distribution of female clients is 1.0. What is more, the clients in our sample are 39 years old on average. 12% of clients are completely illiterate, and 56% of them have completed secondary or tertiary education.

For loan purposes variable, 56% of outstanding loans are invested in commercial activities. The second and the third biggest sectors are agriculture and service, which take up 16% and 13% of loans respectively. The proportion of investment in commerce for the defaulted group is 10% lower than that for the normal group. Hence, there might be a potential association between loan purpose and the probability of default as well.

Table 3.2 shows the correlation matrix of ten different loan default indicators and all explanatory variables. As can be seen from the table, the selection of loan default indicator has a

[1] Unable to calculate the default rate for the hold sample as INSOTEC and MICROCRED have not recorded the length of delayed repayment in their administrative loan books.

substantial influence on the correlation coefficients of all variables. For example, the relation between age and the probability of default is found to be significant and negative in columns (1) to (5) and (7), but it becomes insignificant in the other columns. We can see that the coefficients of some variables have opposite signs in different panels (PaR30 vs Delayed Repayment), such as cohabitation. These statistics bear out the former supposition that even a slight modification of the measurement of loan default may lead to completely different conclusions. The loopholes in definition make the MFIs extremely difficult to establish a reliable credit scoring model, in which the significances of explanatory variables are irrelevant to the risk preferences of microlenders.

As the measurement of loan default is standardized in this study, we focus on the results presented in column (1) and (6) only. As can be seen from column (1), there is a negative correlation between marriage and PaR30 (-0.07), indicating that married borrowers have better repayment rates (**H1**). The correlation between both age and education and the PaR30 are also negative (**H2**, **H4**). Loan purposes variables indicate that lending to clients engaged in agriculture occurs in significant lower PaR30 than lending to clients engaged in consumption, purchasing a fixed asset, and commercial activities (**H5**). Additionally, no significant relationship between gender and PaR30 is found (**H3**).

On the contrary, column (6) tells a completely different story. The length of being in delinquent is irrelevant to both marriage and age (H1, H2), while it positively associates with cohabitating and single clients. Unexpected positive relations have been found between education levels and the length of being in delinquent (H4). In terms of loan purposes, the correlations between the length of delayed repayment and both agriculture and commerce have changed signs (H5), which is unexpected as well. Both abnormal results will be discussed in detail in the empirical results section.

3.4 Methodology

3.4.1 Econometrics models for censored data

The objective of this study is to estimate a default intensity equation using individual-level data. However, such data are characterized by having a large cluster of zeros denoting “no arrears”, or it would denote that a large number of individuals will never default in any situation. In fact, some clients may deliberately choose to default, but in practice, the information is also set to zero. This feature of the data is known in the literature as ‘censoring’.

Censored data may appear to present some methodological difficulties. The cluster of zeros is too large to be ignored econometrically and so the conventional estimator, OLS, seems unsuitable for the purpose of this study. Hence an alternative estimator has to be considered.

When choosing an econometric estimator, one has to make an assumption of the mechanism explaining the zero. Although in practice the nature of this zero may not be entirely known, standard econometric approaches which are conventional in empirical work have attempted to deal with such an issue under different assumptions. The most common econometric approach is the Tobit model, although more flexible estimators have emerged over the years such as the Two-Part model and a closely related one, the Double Hurdle model.

For this analysis, the preference lays on the Two-Part model (2PM hereafter) to estimate the intensity of loan default in sample countries. The empirical evidence presented previously shows that this model provides the best fit given the data available. Although the Tobit estimator has been ruled out for this analysis, a discussion of this model has been included since it provides an ideal starting point for introducing the actual model being estimated. The exposition provided below is largely based on standard econometric text-books given that these estimators are conventional methodologies for the problem at hand. The exposition of the Tobit model and Double Hurdle model is mainly based on that provided in Moffatt (2005), whilst the discussion of the 2PM is based on the exposition by Cameron & Trivedi (2005). Given the differences in notation found in different studies, a common notation has been utilized.

3.4.2 The Tobit model

The analysis of series containing a high proportion of zeros has attracted the attention of researchers, not only for analyzing the intensity of loan default but for a wide range of economic applications. The reason lies in the observation that zeros may represent two different processes. Therefore statistical methods treating these by one distribution, which is in the case of OLS, appear to be limited (Pudney, 1994). The first econometric model to successfully treat the censoring information with two distributions is due to Tobin (1958). This model, commonly known in the literature as "Tobit" for its resemblance to the Probit model, would specify the intensity of default in terms of an index equation such as:

$$Q_i^* = \boldsymbol{x}'_i\beta + \sigma\varepsilon_i$$

$$Q_i = Q_i^* \quad if\ Q_i^* > 0;\ Q_i = 0 \quad otherwise \quad (1)$$

where Q_i^* is the latent dependent variable described before, $\boldsymbol{x}_i$ is a vector of individuals' socioeconomic and demographic characteristics affecting the intensity of loan default, σ is a scale parameter and ε_i is the error term which reflects the unobserved heterogeneity in the utility maximisation solution process. The model in (1) is linear in regressors with an additive error that is normally distributive and homoscedastic such that:

$$\sigma\varepsilon_i \sim NID(0, \sigma^2) \quad (2)$$

The model represented in (1) corresponds to the standard Tobit model (Tobit type I in the literature) where the non-negativity constraint is imposed. In order to estimate the parameters in (1), a Maximum Likelihood (ML) routine is usually applied. The log-likelihood function of the Tobit model can be written as:

$$logL(\beta, \sigma^2) = \sum_{i \in I_0} logP\{Q_i = 0\} + \sum_{i \in I_1} logf(Q_i) \quad (3)$$

where the indexes I_1 and I_0 represent the set of zeros and positive values respectively and f is a specified function. The likelihood function reveals several features of the model that are relevant for choosing an appropriate estimator for modelling the intensity of loan default. For example, it is easy to see that the Tobit model decomposes the two processes involved with two different densities. On one hand, the density that represents the probability $Q_i = 0$ is given by

$$P\{Q_i = 0\} = 1 - \Phi(\boldsymbol{x}'\beta/\sigma) \quad (4)$$

where Φ is the univariate standard normal cumulative distribution function (CDF). On the other hand, the density representing the distribution of positive, which is just the truncated (at zero) normal distribution. The conditional expectation is given by:

$$E\{Q_i | Q_i > 0\} = \boldsymbol{x}'\beta + \sigma \frac{\phi(\boldsymbol{x}'\beta/\sigma)}{\Phi(\boldsymbol{x}'\beta/\sigma)} \quad (5)$$

where ϕ is the standard normal probability distribution function (PDF). Technically then, this model accommodates the censoring of the information into a formal statistical model.

The Tobit model, however, relies upon several important assumptions that have been found unsuitable not only for this study but in many applications. First, it is important to point out that with this model the intensity of default is generated by the following process:

$$Q_i = max(Q_i^*, 0) \quad (6)$$

Therefore, it assumes that the nature of censoring corresponds to a 'corner solution'. Empirically, it would imply that 'at current age, gender, marital status and etc., the individual will

never default', and is, therefore, a corner solution to his or her utility maximisation problem. As a result, substantial changes in individual characteristics could result in positive repayment. This may not hold true in this case. In the literature, it is more common to assume that the zero arises because the individual deliberately chooses to default in microfinance. Thus, one of the main limitations of the Tobit model is that it rules out the possibility of a "true zero". In other words, it rules out the possibility that individuals do not repay purely by choice and not because of their current financial conditions. Perhaps failing to distinguish corner solutions from "true zeros" is one of the main reasons why the Tobit model is usually rejected.

Second, even if the assumption of the "corner solution" is accepted, the structure of the Tobit model is viewed as too restrictive given that this model encompasses the two distributional processes into a single equation. Verbeek (2008) also explains that 'exactly the same variables affecting the probability of a non-zero observation determine the level of a positive observation and, moreover, with the same sign' (p.227). Empirically, this has been found unsatisfactory especially within the context of loan repayment. It may be the case that factors determining to be in default and factors determining the level of loan default are different.

Thirdly, the assumptions on which the model relies on unbiased and consistent ML estimates are too strong to work empirically. It has been stated that the error term in (1) must be homoscedastic and normally distributed. The empirical evidence suggests that these conditions are difficult to meet largely because data, in which this type of regression has been considered, is usually by nature highly skewed. By far the biggest concern is the presence of non-normally distributed errors which, in such a case, calculated estimates are inconsistent.

For all the reasons mentioned above, the Tobit estimator is usually rejected in favour of its alternatives. Fundamentally, alternative estimators are models flexible enough in capturing the different determinants involving the process of the probability of encountering positive outcomes and the determinants involving the level of loan default. However, when the process in which the zeros are generated is unknown, most attention is paid to the convenience of alternative estimators simply because they rely on weaker distributional assumptions for consistent estimates. The 2PM presented below successfully addresses both issues.

3.4.3 The Two-Part model (2PM)

The underlying assumptions that motivate the use of the Tobit model within the context of loan repayment appears to be too restrictive. Fundamentally because of the empirical evidence
favours the view that the intensity of default arises from an individual's subjective choice of being in default.

As a result, there is an interest in disentangling the choice of default and the actual intensity of default which is in fact observed. Thus, an alternative estimator for the Tobit model is usually applied, namely the 2PM. This model provides more flexibility for determining the probability of observing default and the observed outcome. As an alternative estimator, the intensity of default is modelled by two separate processes: the first process denoted as "participation" which accounts for the censoring mechanism and the second process denoted as "intensity", which accounts for the outcome or level of loan default. In its general form, the model can be written as:

$$f(Q|x) = \begin{cases} Pr[d = 0|x] & if\ Q = 0 \\ Pr[d = 0|x] f(Q|d = 1, x) & if\ Q > 0 \end{cases} \quad (7)$$

where f is a specified density function and is an indicator variable equal to 1 for a non-defaulted client, 0 otherwise. This model is also usually referred to the literature as Cragg's model (Cragg, 1971) or simply the Hurdle model. The model is appealing for its simplicity in estimation. Usually the participation equation is estimated by means of a Probit model. In turn, the intensity equation can be estimated say, by OLS with the sub-sample of positive values of Q. The expression in (7) can be represented by:

PARTICIPATION FUNCTION

$$d_i = \boldsymbol{x}'_i\beta + \varepsilon_{1i} \quad \varepsilon_{1i} \sim N[0,1]$$

$$d_i = 1 \quad if\ Q_i > 0, \quad 0\ otherwise. \quad (8)$$

INTENSITY FUNCTION

$$Q_i = \boldsymbol{x}'_i\beta + \varepsilon_{2i} \quad \varepsilon_{2i} \sim N[0, \sigma^2]$$

$$Q_i = Q_i \quad if\ Q_i > 0, \quad 0\ otherwise. \quad (9)$$

One important feature of the 2PM is that it relies on the assumption that ε_{1i} and ε_{2i} are uncorrelated. This means that the intensity is identified based on "selection on observables"

(Cameron and Trivedi, 2005). There is, however, a discussion of whether correlation between ε_{1i} and ε_{2i} should be allowed which in such a case, a closely related estimator namely the 'heckit' estimator would become relevant. The heckit estimator is a consistent and alternative estimator to the Heckman model (Heckman, 1979) and although it is not an efficient estimator it is computationally simpler than the usual Heckman ML estimator. The heckit estimator is also a type of 2PM in which the participation equation is based on a Probit model just as in (8) but the, intensity equation also includes the Inverse Mills Ratio. It is usually motivated by the "sample selection" grounds which are a closely related issues to the two mechanisms explaining the cluster of zeros: the 'corner solution' which has been ruled out, and the 'abstention' or 'choice' which is more in accordance with the existent literature.

Even when sample selection is not a problem of concern, many researchers still see the heckit estimator as an ideal alternative to the 2PM. However, depending on the research question, caution should be taken when choosing the appropriate econometric model given each estimator produces results with different interpretations. An interesting discussion concerning which estimator should be used is given by Madden (2008). His analysis pointed out several criteria that should be taken into account before choosing between the 2PM and the heckit estimator. On theoretical grounds, Madden (2008) doubts whether the heckit estimator is relevant for analyzing, in particular, the intensity of default given that the prediction is based on "potential outcomes". This contrasts with the 2PM where prediction is based on "actual outcomes". In loan default studies, the main concerned is on the latter.

Moreover, at a more technical level, the issue of potential versus actual outcomes relates to the fact that the 2PM is better suited to estimate the "unconditional mean" of Q_i therefore inferences about unconditional partial effects can be made (Mullahay, 1998). In contrast, the heckit estimator is designed to estimate the "conditional mean" of Q_i and to correct for selectivity bias. Thus, "unconditional partial effects" are more difficult to calculate. Nevertheless, even if this difficulty is overcome, it remains an empirical question whether correlation between the two equations is relevant and/or sample selection turns out to be an issue that should be addressed given the problem at hand. Otherwise, the 2PM is the better alternative available to the Tobit model.

Lastly, in empirical applications, the practicality of the heckit estimator or the Heckman model, in general, has been questioned. This has to do with whether the same regressors should be used in the participation and intensity equations or if the exclusion of some varia-

bles should be imposed. This is important because under circumstances of no exclusion, adding Inverse Mills Ratio as an additional regressor may lead to a multicollinearity problem. Usually, the same regressors are used in both equations to test whether the factors affecting the probability of being defaulted are the same as those factors affecting the level of default, regarding both the sign and the statistical significance. For the problem at hand, it seems reasonable to follow this approach although there is no reason why it should be so even when collinearity is not suspected. However, when a collinearity problem is present, exclusion restrictions must be imposed. It is commonly imposed in the intensity equation. An empirical difficulty arises given that there is no clear guidance to which variables should be excluded. Thus, in the absence of clear choices for exclusions, particularly when collinearity persists, on practical grounds the heckit estimator is not an ideal estimator to use.

In summary, from the most common models of censored data, this analysis will employ the 2PM for modelling the intensity of default. This model has been found to be flexible enough to recognise the most plausible mechanism explaining the cluster of zeros which according to the existing literature is by "choice" or "abstention". Also, given that there is no reason to believe that correlation between the two equations (participation and intensity) and selection bias would be an issue of concern, the heckit estimator will not be considered. Therefore, the methodology will be restricted to estimate a Probit model for participation in the first stage and the second stage, an alternative procedure to the usual OLS is employed to avoid retransformation problem. This is further explained in the last subsection.

3.4.4 Cragg's Double-Hurdle model (DH)

In fact, besides the two models introduced above, the 2PM model and heckit model, the Double Hurdle family has two other members: Cragg's (1971) Double Hurdle model (DH), and the Full Double-Hurdle (FDH) model. The major differences and are based on two aspects: a) the independence between the residuals of participation and intensity function; and b) the dominance (or first hurdle dominance) which implies that no individual is observed at a standard corner solution and that once the first hurdle has been passed, standard Tobit censoring is no longer relevant (Jones, 1989). Therefore, the selection criteria of the four models are as follows:

- M1: Independent but NOT Dominance: Cragg's (1971) Double Hurdle model.
- M2: Dominance but NOT Independent: Heckman's selection model, or heckit model.
- M3: Independent AND Dominance: Two-Part model.
- M4: NEITHER Independent NOR Dominance: Full Double Hurdle model.

In this paper, we have assumed that the two equations (participation and intensity) are uncorrelated as discussed previously. Hence, M1 and M3 should be our potential choices because they rely on the assumption of independence. Regarding the second assumption, technically speaking, first hurdle domination is not very convincing. Empirically, it would imply that at current income, current loan amount, with certain characteristics (gender, age, etc.), the client will never repay the loans, and is at a corner solution to his or her utility maximisation. This may not hold true given the financial products in question. It is more common to assume that zero arises because of the client's financial decision: comparison between investment return and penalty of delay, balancing between livelihood expenditure and repayment. By releasing the assumption of first hurdle domination, zero can be generated from two unobservable sources. Therefore, it is interesting to include and compare the results of both M1 and M3 in this paper. If the results are significantly different from each other, we might infer that the second unobservable source (client's financial decision) is the main reason for loan default. If the results are very similar, we might infer that the clients expected to repay their loans as soon as possible. The insufficient fund is the main reason for loan default.

The basic form of the hurdle model was introduced by Cragg (1971), and its statistical properties are well established in the literature (see Pudney (1994) and Smith (2002). A representative example is given by García and Labeaga (1996). They define the two hurdles and the way in which they interact with the dependent variable as follows:

$$d_i^* = \boldsymbol{x}'_{1i}\beta_1 + \varepsilon_{1i}$$

$$y_i^* = \boldsymbol{x}'_{2i}\beta_2 + \varepsilon_{2i} \quad (10)$$

where d_i^* denotes whether individual is a defaulted client or reports non-zero delayed repayment of his/her loan (latent participation variable) therefore:

$$d_i = 1 \quad if\ d_i^* > 0, \quad d_i = 0\ otherwise. \quad (11)$$

and y_i^* is the latent dependent variable such as that:

$$y_i = y_i^* \quad if\ d_i^* > 0, \quad y_i = 0\ otherwise. \quad (12)$$

where y_i is the observed intensity of delayed repayment. x_i is a vector of individuals' characteristics (i.e socio-economic and/or demographic), β_i is a vector of parameters to be estimated and ε_i is the error term. As the model produces two error terms, these are assumed to follow a (bivariate) normal distribution with zero mean and constant variance such that:

$$\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix} \sim N\left\{\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho\sigma\\ \rho\sigma & \sigma^2\end{bmatrix}\right\} \quad (13)$$

3.4.5 The Box-Cox transformation and likelihood functions

Having a large cluster of zeros, however, is not the only methodological difficulty encountered. An additional characteristic commonly present is that for those individuals reporting delayed repayment, the distribution of the data appears (highly) skewed to the right and exhibits non-constant variance (Tauras, 2005). This implies that using data in its original structure may lead to inefficient or even inconsistent estimates depending on the econometric model used. In fact, all of the models outlined in this section rely heavily on the assumption of normality in the error terms: without normality, the property of consistency of the Maximum Likelihood estimation fails to hold. Researchers usually overcome this problem by applying a suitable variable transformation, and the most common one is a logarithmic transformation. However, the logarithmic transformation is inappropriate due to the presence of the zero observations in the sample, especially in the present situation in which the zeros are the focus of the analysis.

Instead, we follow Jones and Yen (2000) by applying the Box-Cox transformation, defined as

$$y^T = \frac{y^{\lambda}-1}{\lambda}, \quad 0 < \lambda \leq 1 \quad (14)$$

Note that the Box-Cox transformation (14) includes as special cases a straightforward linear transformation (λ = 1), and the logarithmic transformation (λ -> 0), but normally we would expect the parameter lambda to be somewhere between these limits. The transformation (14) can be applied to any of the models outlined in this section, including the Double Hurdle model introduced later. When it is applied to the dependent variable in the 2PM model, we obtain the Box-Cox 2PM model, defined as follows:

PARTICIPATION FUNCTION

$$d_i{}^* = \gamma + \delta_1 \boldsymbol{S}_i + \delta_2 \boldsymbol{C}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,\ 1), \quad d_i = 1\ if\ d_i{}^* > 0, \quad d_i = 0\ if\ d_i{}^* \leq 0 \quad (14)$$

INTENSITY FUNCTION

$$y_i{}^{**} = \alpha + \beta_1 \boldsymbol{S}_i + \beta_2 \boldsymbol{C}_i + u_i, \quad u_i \sim N(0, \sigma^2)$$

$$y_i{}^{*T} = max\left(y_i{}^{**T}, -\frac{1}{\lambda}\right) \quad (15)$$

OBSERVED $y_i{}^T$

$$y_i{}^T = y_i{}^{*T} \ if\ d_i = 1, \quad y_i{}^T = -\frac{1}{\lambda} \ if\ d_i = 0 \quad (16)$$

Note that the lower limit of the transformed variable is -1/λ rather than zero.

The log-likelihood function for the Box-Cox 2PM model is

$$Log\ L = \sum_0 ln[1 - \boldsymbol{\Phi}(\delta_1 \boldsymbol{S}_i + \delta_2 \boldsymbol{C}_i)] + \sum_+ ln\left[\boldsymbol{\Phi}(\delta_1 \boldsymbol{S}_i + \delta_2 \boldsymbol{C}_i)\frac{1}{\sigma}\right] \quad (17)$$

where $y_i{}^*$ is the latent variable representing propensity to delay; $\boldsymbol{S}_i$ and $\boldsymbol{C}_i$ are matrices of socio-demographic and credit related explanatory variables; λ is the Box-Cox Transformation parameter; 0 and + indicate summations over the zero observations and positive observations; $\boldsymbol{\Phi}$ and ϕ are the standard normal cdf and pdf.

When all these functions are applied to the Double Hurdle model, the participation and intensity function and observed $y_i{}^T$ are exactly the same as above. The only difference is the general log-likelihood function for the Box-Cox Double Hurdle model shown as follows:

$$Log\ L = \sum_0 ln\left[1 - \boldsymbol{\Phi}(\delta_1 \boldsymbol{S}_i + \delta_2 \boldsymbol{C}_i)\boldsymbol{\Phi}\left(\frac{\beta_1 \boldsymbol{S}_i + \beta_2 \boldsymbol{C}_i + \frac{1}{\lambda}}{\sigma}\right)\right]$$

$$+ \sum_+ ln\left[\boldsymbol{\Phi}(\delta_1 \boldsymbol{S}_i + \delta_2 \boldsymbol{C}_i) y_i{}^{\lambda-1}\frac{1}{\sigma}\phi\left(\frac{y_i{}^T - \beta_1 \boldsymbol{S}_i - \beta_2 \boldsymbol{C}_i}{\sigma}\right)\right] \quad (18)$$

Note that (21) is not very different from the log-likelihood function for the Box-Cox 2PM model (20). Two important differences are the involution of the second hurdle's cdf and pdf, and the Jacobian term $y_i{}^{\lambda-1}$ that required by the use of $y_i{}^T$ in the final term. The estimation of the 2PM and Double Hurdle are relatively easy in STATA as in-built commands for the Probit model and GLM are available.

3.5 Empirical Results and Discussion

3.5.1 Model comparison and relations between marital status and loan delinquency

The comparison of results from four models are reported in **Table 3.3** for both Amount in Arrears (Panel A) and Length of Delayed Repayment (Panel B). The sample sizes used in the estimations of the two panels are 17369 and 14170 respectively. Recall from the previous section that these samples have been artificially weighted to reflect the true population ratio of different countries. These explanatory variables are selected based on previous literature

and then tried out many different combinations in the Box-Cox 2PM and DH models. The different columns correspond to the different estimation methods (Probit, Tobit, 2PM, and DH).

Statistically, the results of the first hurdles of Box-Cox 2PM and DH model are the same as those of the Probit model. Considering the entire models, 2PM and DH appear superior to the Tobit model. From the Akaike's Information Criterion (AIC) at the foot of each panel, we can see that 2PM and DH models have much lower AIC (0.85 in Panel A and 0.48 in Panel B) compared to the Tobit model (1.07 in Panel A and 0.52 in Panel B). It provides evidence that the first hurdle should not be ignored from the estimation process, and the first hurdle dominance effect is relatively strong in our samples, especially when we use the amount of arrears as the proxy for loan default in regression analysis. As introduced in the previous section, selection between 2PM and DH models depends on the assumption of first hurdle dominance. However, **Table 3.3** shows that both models provide very similar results and AIC (column (4) and (6)).

Considering the practical applications in real life scenarios, the efficiency of the algorithm plays a crucial role in modelling. In this sense, the 2PM is faster than the DH model when there are hundreds of iterations, a great number of variables, and many nonlinear terms. Meanwhile, the estimation performances between 2PM and DH models are marginal. Therefore, by implementing the 2PM in credit scoring, the MFIs would obtain extra information related to the probability and intensity of default with moderate time investment. The rest of the regression analyses in this paper are mainly based on the 2PM method.

In columns (3) in **Table 3.3**, we see that married borrowers are less likely to be potential defaulters, measured by both the amount of arrears (Panel A) and the length of delayed repayment (Panel B). Even focusing on the respondents conditional on default (columns (4)), married borrowers still tend to have lower levels of arrears and shorter length of delayed repayment. On the other hand, while the relationship between cohabitation and length of delayed repayment is negative and significant in the first hurdle (Panel B), the relationship between cohabitation and amount in arrears is positive and significant in the second hurdle only (Panel A). Thus, the results for cohabiting borrowers are inconsistent in the two samples with different MFIs. One explanation for the different impacts of marriage and cohabitation could be the fact that cohabiting couples usually have fewer economic resources than married couples (Manning and Lichter, 1996). Besides, the nature of the cohabiting union has been described as having lower relationship quality compared with marriages (Brown and

Booth, 1996). For instance, couples in long-term cohabiting relationships have a higher probability of domestic violence than married couples (Kenney and McLanahan, 2001). Cohabitees may not share their income with their partners (Bauman, 1999) since cohabiting couples do not necessarily pool their resources in the same manner as married couples. After all, the significance of cohabitation depends on the different philosophical, political and religious ideas that dominate in the sample countries.

Examining the other demographic characteristics, we see that the results of gender, age and education background, are also inconsistent as we change the combination of sample MFIs. Panel A of **Table 3.3** shows that gender and age are insignificant to the probability of being in default, while educational background significantly relates to it. Regarding the intensity of default, female borrowers tend to have a lower level of arrears. For comparison, from Panel B we see that gender, age and educational background are significant at 5%, 10%, and 1% respectively. It is also found that age indeed has a U-shaped effect on both the probability and intensity of default, with a maximum at 0.03/(2*0.0006)=50.0. This implies that borrowers aged 50 years are the riskiest to be in the 'always-default' category. However, whether it is the true age effect or just a cohort effect would require additional observations taken in a different year, but it is beyond the scope of this chapter.

Also, it is surprising to find that education positively relates to the probability of potential default in Panel B of **Table 3.3**. According to the comparisons in **Table 3.4**, we found that the positive relationship between education and the probability of loan default only exists in FINCA Peru but other MFIs. Hence, it is reasonable to infer that the positive association between education and the probability of default presented in Table 3 is caused by the strong influence of FINCA Peru for which the sample size is much larger.

As introduced in the literature review section, most loan default empirical studies in the classical banking area suggest that a more educated borrower should have a lower probability of default and lower intensity of delinquency. However, our results show that it is not the case for FINCA Peru. Based on what we discussed previously, there are at least two possible explanations for the abnormal results for FINCA Peru: 1. their borrowers participate in business activities that require very little education, but experience and skills; and 2. borrowers with better education have access to more credit sources than their competitors, such as the other MFIs in Peru.

Regarding the attitudes towards education, the people of Peru are very different from the people in the other Latin American countries. In practice, parents paying for private education is common in most of the Latin American countries. Regarding the regional economic report provided by OECD (2011), disadvantaged families usually make a greater effort (over 13%) than the affluent families (less than 4%) regarding the percentage of household income devoted to education. The schooling gap caused by different education expenditures will greatly shrink, and the relationship between education and parental income is weak. However, it is not the case for Peru. In Peru, the affluent families invest more than 12% of household income on education, while the investment of the disadvantaged families is just 8%. Education is an indicator of parental income. Better education usually indicates better family background and thus higher creditworthiness. This is consistent with the second point presented in the last paragraph.

By examining the variables representing loan purposes at the lower section in Panel A of **Table 3.3**, we see that consumption and agriculture are related to a lower probability of default, and 'buying fixed asset' is related to a lower intensity of default. It means that agriculture is the safest sector to invest in compared to commerce, manufacture and service as we should expect. However, focusing on the respondents conditional on default specifically, the four sectors perform similarly in terms of the capability of reducing loan default intensity. What is more, the result of agriculture is inconsistent and becomes insignificant in Panel B. Hence, further subsample analysis for the influences of different investments is needed. There are no convincing conclusions at this stage.

3.5.2 Subsample analysis between different microfinance institutions

Table 3.4 presents split-sample regressions, where the main regressions are repeated for the four MFIs with two different measurements of loan default. The results, which substantially change across different sub-samples, reconfirm that, in general, married borrowers are associated with lower probabilities of being in default, but such a relation disappears in CACIL Honduras. A possible explanation to the results differ across countries is that the benefit of second income brought by marriage might be neutralised by the extra financial burden associated with the number of dependents. The summary statistics of our data shows that the average household size in CACIL Honduras is 2.84, while the numbers in the other MFIs are lower than 2.18.

What is more, the influence of marriage is especially pronounced for MICROCRED Madagascar, where married borrowers have not only a higher repayment rate but also a smaller intensity of default (Columns 3). These findings are consistent with the propositions of Dunn and Kim (1999) and Vogelgesang (2003) stated previously. Hypothesis 1 is accepted, in terms of 'marriage'. The most likely reason why married borrowers have better repayment records should be the high-risk tolerance brought by the community property regime.

However, when we look at the other variables of marital status, 'cohabitation' is only significant in MICROCRED and FINCA, while 'divorced' and 'widowed' are only significant in MICROCRED. Therefore, we may infer that in some countries a sense of responsibility might be irrelevant to the probabilities of default and delinquent, as there are no differences between single and cohabitation.

Regarding the other demographic characteristics such as gender, age and education level, the results are distinct from one country to another. Gender and age are insignificant in most of the subsamples. In FINCA Peru, female borrowers are associated with a lower probability of default as expected but a higher intensity of default. On the other hand, in CACIL Honduras, significant convex relations between age and loan default have been found in both hurdles. This finding contradicts Hypothesis 2 that the youngest and oldest groups of borrowers have the highest probabilities of repayment. It could be because the oldest borrowers have higher financial commitments to their family and business expenses, while the youngest borrowers are less responsible in repaying their loans and lack of experience in business and financing.

In this section, we also found that a higher education level is not always significant and positively related to better repayment performance across different MFIs. It is detected in the subsample of MICROCRED Madagascar only. In terms of CACIL Honduras, the clients who completed primary education have shorter periods of delayed repayment at the 1% level. However, there is no relationship between education level and the probability of loan default. These findings imply that, in many countries, financial literacy and financial awareness, which highly associate with repayment performance, are generated from other sources instead of formal education at school.

The results for the variables of loan purposes in **Table 3.4** are inconsistent to those in **Table 3.3**. We reconfirm that investing in agricultural business is associated with lower probability of delinquency in the subsamples of INSOTEC Ecuador and FINCA Peru. However, in CACIL Honduras, it is found that engaging in agriculture has a higher intensity of delinquency. In

terms of the probability of delinquency, the difference between the agricultural and service sectors is insignificant.

There are two possible explanations for the inconsistent results. First, borrowers involved in agricultural businesses usually use credit to purchase seeds, fertiliser, pesticides, livestock, machinery, etc. They need at least four months to one year to receive the revenue from harvesting. Hence, they cannot pay back the loan with a high repayment frequency, such as the weekly repayment plan with two weeks grace period which may sound feasible to the other business sectors. In our data, all microfinance institutions in the sample apply an indiscriminative and fixed frequency of repayment (monthly) to all clients. Considering the payback periods of different sectors, the institutional management should re-evaluate and modify their lending system to ease the burden on specific groups of borrowers.

Second, while agriculture is claimed to be the safest sector due to high social control and low volatility, it is in line with the prevailing weather conditions and indeterminate natural disasters. May 28 of 2009, an earthquake with a moment magnitude of 7.3 (very strong) occurred at the Caribbean Sea, 320 kilometres northeast of Tegucigalpa, the capital of Honduras. It caused an estimated $37 million worth of damage. Mar 25 of 2010, the National Congress of Honduras approved to declare a national emergency caused by a prolonged drought and famine. 7,000 families suffered from severe food shortages. Paradoxically, Honduras also experienced flooding and excessive rains in other regions within the same period. According to the statistics proved by Knoema, the cereal production had decreased by 9.9%, and the primary vegetable production had decreased by 7.8% in 2010. With the strong interventions of natural disasters, the regression results related to agriculture in Honduras may be bias and misleading. In fact, Honduras' vulnerability to natural disaster kept increasing dramatically in the recent decades. Its nominal losses were estimated at $4.7 billion, nearly half the losses for the entire Central American since 1974 (IADB, 2009). Therefore, the risk of investing in agricultural businesses may vary from a country to another.

3.5.3 Interaction terms analysis of marital status

Regression results concerning the interaction terms are presented in **Table 3.5**. Panel A analyses the probability of default regarding the amount in arrears, and Panel B measures the probability of default in terms of the number of days of delayed repayment. The different columns represent the different interaction terms that were added subsequently. In order to study the isolated impact of each interaction, all interaction variables are taken up simultaneously in the last column. The relations between marital status and the probability of

default for the different categories are now indicated by the sum of the coefficients for the interaction terms and the reference coefficients at the first row of each section.

As can be seen from column (1) of both Panels A and B of **Table 3.5**, the interaction term (married female clients) returns an insignificant coefficient, which indicates that the relation between marital status and the probability of default does not differ with gender. Therefore, the impact of marriage on the male clients is as strong as that on the female clients. On the other hand, when it comes to age, both Panels A and B of **Table 3.5** indicate that clients aged between 40 and 49 benefits the most from marriage when it comes to repayment performance. In contrast, the relation between marriage and the probability of default is weakened for clients aged between 22 and 29. These effects are not only found in the individual regressions but also persistent when the other interactions are added, as shown in column (8).

As for education background, columns (5)-(7) of Panel A indicate that the relation between marriage and the probability of default is strengthened if a client has completed a secondary school diploma or the equivalent. This finding may have two implications as follows: 1. when the education level of a client is too low or too high, his/her probability of default is mostly determined by education; 2. when the education level of a clients is moderate, his/her probability of default is generally determined by marital status. On the other hand, we can also see that the sign and significant level are upheld when all interaction variables are included (column 8). In addition, the results of the interaction terms with primary school education or tertiary school education are insignificant in the individual regressions.

These findings suggest that there are two mechanisms that can actively reinforce the positive relation between marital status and repayment performance: First, the ages of clients are likely to increase the positive impact of marriage and reduce the probability of loan default. According to the theory of economic resources to marriage (Becker, 1973, 1974) and the longitudinal study of well-being in young adults' marriage (Clarkberg, 1999), there are a series of fixed costs associated with marriage. These include the cost of the wedding, the purchase of a house, household equipment, and childbearing. Also, married clients are likely to have a larger family size and higher expenses compared to single clients. Therefore, single clients should have better repayment ability than married clients in young adulthood. Nevertheless, as the children mature and become the new labour force, the financial burden will turn into extra income, and lower the probability of loan default. Our findings illustrate that

the relationship between marriage and repayment performance greatly depends on the ages of clients.

Second, based on the current situation of relatively low levels of education in SMEs, marriage can effectively enhance financial awareness and improve repayment performance of business owners. Educational disparities across different firm sizes are striking at the university level. For example, there are only 21% of SME owners in Chile have Bachelor's degrees, compared to 55% of medium-firm owners (Alvarez and Crespi, 2003). A possible explanation of such a phenomenon is that the poor often create survival-oriented SMEs due to a lack of job opportunities. If an SME owner's education level is extremely low, he/she might be incapable of sustaining the business. If an SME owner's education level is higher than the country-specific threshold of education, he/she might start up the business in a rapidly growing sector and earn high revenues. In both cases, the relations between education and the probability of default do not differ with marital status. However, marital status will be influential once it interacts with secondary school attainment in developing countries. While secondary school attainment had no discernible impact on SME growth as tertiary education (Kantis et al., 2004), it provides basic numeracy skills for a business that not included in primary education, and noticeably increases the survivability of SMEs. In Madagascar and Peru (See Appendix. A), for instance, most SMEs owners completed secondary school and concentrate on a small group of business activities that grow relatively slowly. Under such circumstances, a strong sense of responsibility to the family and their creditworthiness might be the prime motivation for the SME owners to repay their loans on schedule.

3.5.4 Further analysis of the impacts of loan classification standards

A final robustness test has been carried out to analyse the findings more in detail. Specifically, since there is no universal standard of loan classification across different MFIs, we want to examine whether the results hold as we change the thresholds of sub-standard, doubtful and bad loans. While different of loan classifications are implemented, the frameworks and measuring methods of the criterions used by MFIs share similar specifications. They usually quantify the actual repayment capacities of borrowers and classify the loans into five categories based on risk: 1. standard; 2. special mention; 3. substandard; 4. doubtful; and 5. loss. The latter three credit grades are defined as 'bad loan'. The length of overdue repayment and PaR are the key performance indicators to quantify and classify the micro loans.

Some important issues may arise as we conduct loan classification with PaR in the microfinance industry and impair the robustness and the comparability between MFIs. First of all, none of the MFIs in our data can separate the restructured loans from their non-restructured loans. As a result, we have to assume that all loans are non-restructured in this paper. It might underestimate portfolio risk seriously, especially when restructured loans appear to be material (over 1%) for an MFI. Second, conventional measures of PaR (30, 90) are meaningless for a balloon payment at the end of the loan period, which is the case in agricultural lending when repayments are tied to the crop cycle. In our data, all MFIs follow a rigid monthly repayment contract model. Thus, the portfolio risks for some borrowers might be overestimated. Third, portfolios with different risk profiles may have the same PaR value. For instance, while the PaR measure is the same, a loan with a large concentration of seriously delinquent loans (affected by arrears of more than 180 days) will be riskier than a loan where arrears remain in the range within 60 days. Hence, PaR is a useful measure, but it does not tell the whole story.

Table 3.6 presents logistic regressions where the main regression is repeated for different thresholds of loan default. In this section, we define 30% of PaR30 and 90 days of delayed repayment as the thresholds of bad loans. Columns (3) and (8) show that the percentages of bad loans are 1.8% and 2% in the two groups of MFIs with different performance indicators respectively. For the subsample including CACIL, INSOTEC and MICROCRED, the nonperforming loans to total gross loans (NPL ratio) are just 0.85%. According to statistics from the World Bank, the average rate of NPL in the world gradually decreased from 4.01% to 3.89% for the period from 2010 to 2011. At the end of 2010, the NPL ratios of Honduras, Ecuador, and Peru are 3.7%, 3.4% and 3.0% respectively. By comparison, we found that the microfinance sector is less risky than the banking industry on average.

As can be seen from the results, marital status is associated with the probability of loan default across different thresholds in both samples. It means that the relation between marital status and repayment performance still holds as we change the standard of credit scoring. Similarly, we can see that the results of cohabitation, age, and education are consistent as well. For the subsample measured by PaR30, we found that gender, loan amount, and time to maturity seem to be more significant as we shift the risk tolerance to higher levels, while the tendency of loan purposes is just the opposite. For the subsample measured by the length of overdue repayment, gender, time to maturity, and loan purposes have inconsistent results, but no certain patterns are found as we alter the threshold of risk. In summary, the significances of gender, loan amount, time to maturity, and loan purposes highly depend on

the level of risk tolerance. Therefore, the results for these four independent variables can be manipulated and biased based on the current credit scoring standards.

3.6 Conclusions and Discussion

Microfinance loans are a major tool for developing countries to fight poverty. However, the balance between outreach to the poor and financial sustainability is hard to achieve. Therefore, a better understanding of the risk determinants of loan delinquency and default is of great importance in the area of microfinance. In a cohesive empirical study, we identify the individual socio-demographic and business characteristics that are associated with micro loans based on a high-quality administrative loan book data that stems from four MFIs from developing countries.

In this research, we replace the omnifarious binary default indicators used in previous studies with three semi-continuous default indicators: the amount of arrears, the number of days being in delinquent, and PaR30. There are many advantages of using these variables as default indicators: 1. no discretisation and no information lost; 2. easy to acquire and compare; and more importantly 3. who will never default can be separated from who has a low probability of default. In terms of the explanatory variables, most of them are already known from classical banking and prior literature of microfinance. According to the clustered structure and skewness of the data, a Two-Part model with the Box-Cox transformation is applied here.

First of all, our results show that the estimation performances between the Two-Part model and the Double Hurdle model are similar, while the algorithm of the Two-Part model is more efficient. By implementing the Two-Part model in credit scoring, MFIs would obtain more results for the probability and intensity of default with moderate time investment.

In general, married borrowers have a lower probability of default and a lower intensity of delinquency, measured by both the amount of arrears and the length of delayed repayment. In the subsample analyses, married borrowers have a lower probability of default in general. This relationship is especially pronounced in MICROCRED Madagascar but is insignificant in CACIL Honduras. What is more, we found that the relation between marital status and the probability of default does not differ by gender. In addition, the relation between marriage and the probability of default will be strengthened if a client has completed secondary school.

The results of gender and age are inconsistent across different MFIs, and they are insignificant in most of the subsamples. An interesting finding related to age is that the clients aged between 40 and 49 benefit the most from marriage when it comes to repayment performance.

It is surprising to find that education positively relates to the probability of default in FINCAR Peru, while the association is negative in MICROCRED Madagascar. For CACIL Honduras and INSOTEC Ecuador, no significant relations between education and loan default are detected. Possible explanations of the abnormal results for FINCA Peru include: 1. borrowers participate in business activities that require little education, but lots of experience and skills; 2. borrowers with better education are more likely to be over-indebted as they access to credit much easier, because education is highly related to parental income and creditworthiness in Peru.

Agriculture is related to a lower probability of default as measured by the amount of arrears. However, it becomes insignificant when we use the length of delayed repayment as a proxy for the probability of default. In the subsample analyses, we reconfirm that investing in agricultural business associated with lower probability of default in INSOTEC Ecuador and FINCA Peru. However, in CACIL Honduras, it is found that agriculture positively relates to both the probability and intensity of loan default. Possible explanations to the inconsistent results include: 1. borrowers involved in agricultural businesses cannot pay back the loan with a high repayment frequency; and 2. while agriculture is claimed to be the safest sector due to high social control and low volatility, it is in line with the prevailing weather conditions and indeterminate natural disasters that happen during the period of interest.

Overall, we have provided new insight into important characteristics and risk determinants of micro loans in developing countries. These can be applied to develop the current credit scoring systems implemented by MFIs and can contribute to achieving the ultimate objectives: improving the outreach of microfinance to people in need and reducing poverty.

Table 3.1

Summary Statistics.

'Married' is a dummy that is 1 if the client is married, and 0 otherwise; 'Cohabitation' is a dummy that is 1 if the client is living with his/her partner, and 0 otherwise; 'Divorced' is 1 if the client is divorced, and 0 otherwise; 'Single' is 1 if the client is single, and 0 otherwise; 'Widowed' is 1 if the client is a widow, and 0 otherwise; 'Gender' is 1 if the client is female, and 0 otherwise; 'Illiterate' is 1 if the client has never received any formal education, and 0 otherwise; 'Primary Completed' is 1 if the client has completed primary school, and 0 otherwise; 'Secondary Completed' is 1 if the client has completed secondary school, and 0 otherwise; 'Tertiary Completed' is 1 if the client has completed university, college or trade school education, and 0 otherwise; 'Loan Amount' is the loan outstanding per client measured in dollars; 'Maturity' is the number of days before maturity; 'Consumption' is 1 if the client uses the loan on consumption, and 0 otherwise; 'Buy Fixed Asset' is 1 if the client uses the loan to purchase fixed asset, and 0 otherwise; 'Agriculture' is 1 if the client uses the loan on agricultural production, and 0 otherwise; 'Commerce' is 1 if the client uses the loan on commercial activity, and 0 otherwise; 'Manufacture' is 1 if the client uses the loan on manufacturing, and 0 otherwise; 'Service' is 1 if the client uses the loan to provide services; 'Financing' is 1 if the client uses the loan to finance his/her business.

		Mean							
	n	All	Normal	Abnormal	*Q1*	*Q2*	*Q3*	St. dev.	Min.	Max.
Sociodemographic										
Marital Status										
Married	32673	0.52	0.53	0.46	0	1	1	0.50	0	1
Cohabitation	32673	0.14	0.14	0.09	0	0	0	0.34	0	1
Divorced	32673	0.04	0.04	0.05	0	0	0	0.19	0	1
Single	32673	0.27	0.26	0.38	0	0	1	0.44	0	1
Widowed	32673	0.03	0.03	0.03	0	0	0	0.16	0	1
Gender (Female -> 1)	32673	0.68	0.67	0.56	0	1	1	0.47	0	1
Age	32673	39.35	39.38	38.97	31	38	47	10.46	22	65
Education Levels										
Illiterate	32673	0.12	0.11	0.13	0	0	0	0.32	0	1
Primary Completed	32673	0.32	0.32	0.32	0	0	1	0.47	0	1
Secondary Completed	32673	0.43	0.43	0.43	0	0	1	0.50	0	1
Tertiary Completed	32673	0.13	0.13	0.11	0	0	0	0.34	0	1
Loan Status										
Loan Amount (in USD)	32673	970.31	942.00	1362.20	232.00	580.00	1300.00	1053.31	107.00	6420.00
Maturity (in Days)	32673	316.80	306.69	456.71	180	300	360	202.87	90	1080
Loan Purposes										
Consumption	32673	0.04	0.04	0.11	0	0	0	0.21	0	1
Buy Fixed Asset	32673	0.01	0.01	0.03	0	0	0	0.11	0	1
Agriculture	32673	0.16	0.16	0.16	0	0	0	0.37	0	1
Commerce	32673	0.56	0.57	0.47	0	1	1	0.50	0	1
Manufacture	32673	0.08	0.08	0.07	0	0	0	0.27	0	1
Service	32673	0.13	0.13	0.13	0	0	0	0.34	0	1
Financing	32673	0.01	0.00	0.04	0	0	0	0.07	0	1
Default Indicators										
Delay or Not	32637	0.07	0	1	0	0	0	0.25	0	1
Arrearage (in USD)	17369	18.50	0.00	237.96	0	0	0	111.20	0	1803.05
Delay (in Days)	14170	8.06	0	229.86	0	0	0	65.77	0	1358

Notes: Obvious special cases have been omitted from the analyses. In addition, the influence of outliers has been checked by re-running. The 5th and 95th percentiles of Loan Amount have been trimmed. The 3rd and 98th percentiles of Age have been trimmed. The 1st and 95th percentiles of Maturity have been trimmed. The 99th percentiles of non-zero Arrearage have been trimmed. The 95th percentiles of non-zero Delay have been trimmed. Q1, Q2 and Q3 are the first, second, and third quartiles, respectively.

Table 3.2

Correlations Analysis

'Married' is a dummy that is 1 if the client is married, and 0 otherwise; 'Cohabitation' is a dummy that is 1 if the client is living with his/her partner, and 0 otherwise; 'Divorced' is 1 if the client is divorced, and 0 otherwise; 'Single' is 1 if the client is single, and 0 otherwise; 'Widowed' is 1 if the client is a widow, and 0 otherwise; 'Gender' is 1 if the client is female, and 0 otherwise; 'Illiterate' is 1 if the client has never received any formal education, and 0 otherwise; 'Primary Completed' is 1 if the client has completed primary school, and 0 otherwise; 'Secondary Completed' is 1 if the client has completed secondary school, and 0 otherwise; 'Tertiary Completed' is 1 if the client has completed university, college or trade school education, and 0 otherwise; 'Loan Amount' is the loan outstanding per client measured in dollars; 'Maturity' is the number of days before maturity; 'Consumption' is 1 if the client uses the loan on consumption, and 0 otherwise; 'Buy Fixed Asset' is 1 if the client uses the loan to purchase fixed asset, and 0 otherwise; 'Agriculture' is 1 if the client uses the loan on agricultural production, and 0 otherwise; 'Commerce' is 1 if the client uses the loan on commercial activity, and 0 otherwise; 'Manufacture' is 1 if the client uses the loan on manufacturing, and 0 otherwise; 'Service' is 1 if the client uses the loan to provide services; 'Financing' is 1 if the client uses the loan to finance his/her business.

	Individual PaR30					Delayed Repayment (in Days)				
	(Sample: CACIL, INSOTEC & MICROCRED)					(Sample: CACIL & FINCA)				
	> 0%	≥ 5%	≥ 30%	≥ 50%	≥ 75%	> 0	≥ 30	≥ 90	≥ 180	≥ 360
Sociodemographic										
Marital Status										
Married	-0.07***	-0.06***	-0.07***	-0.07***	-0.05***	-0.003	-0.02**	-0.02**	-0.02***	-0.003
Cohabitation	0.04***	0.02**	0.03***	0.04***	0.04***	0.02*	-0.01	-0.006	-0.01	-0.005
Divorced	0.01	0.02***	0.02**	0.008	-0.009	0.007	0.0007	0.006	0.008	0.008
Single	0.05***	0.04***	0.06***	0.05***	0.04***	0.02**	0.03***	0.03***	0.03***	0.01
Widowed	0.009	0.01*	0.002	-0.004	-0.009	-0.008	-0.003	-0.002	0.001	-0.01*
Gender (Female -> 1)	0.003	0.007	-0.004	-0.007	-0.005	-0.10***	-0.09***	-0.06***	-0.05***	-0.04***
Age	-0.03***	-0.03***	-0.04***	-0.03***	-0.01*	-0.006	-0.02**	-0.01	-0.009	0.004
Education Levels										
Illiterate	0.16***	0.12***	0.10***	0.08***	0.02***	-0.02***	-0.03***	-0.02***	-0.02***	-0.01
Primary Completed	-0.02***	-0.02**	-0.006	-0.02**	-0.02***	-0.004	-0.009	-0.007	-0.02**	-0.02*
Secondary Completed	-0.03***	-0.03***	-0.02***	-0.006	0.02**	0.03***	0.04***	0.04***	0.04***	0.02**
Tertiary Completed	-0.01	-0.01	-0.02**	-0.02**	-0.009	-0.008	-0.009	-0.01	-0.007	0.002
Loan Status										
Loan Amount (in USD)	-0.008	-0.02**	-0.02***	-0.02**	0.009	0.15***	0.11***	0.05***	0.03***	0.03***
Maturity (in Days)	0.006	-0.04***	-0.05***	-0.04***	-0.01*	0.18***	0.12***	0.05***	0.02***	0.02**
Loan Purposes										
Consumption	0.01*	-0.03***	-0.02**	-0.01	-0.007	0.06***	0.04***	0.003	-0.01	0.001
Buy Fixed Asset	0.02***	-0.02***	-0.01*	-0.01	-0.006	0.07***	0.02**	-0.003	-0.006	-0.002
Agriculture	-0.03***	-0.02***	-0.005	-0.009	-0.02**	0.08***	0.09***	0.08***	0.06***	0.03***
Commerce	0.03***	0.04***	0.03***	0.03***	0.03***	-0.08***	-0.07***	-0.05***	-0.03***	-0.01
Manufacture	-0.01*	-0.003	-0.01	-0.02**	-0.02**	-0.01	-0.01	-0.01	-0.01*	-0.01*
Service	-0.007	-0.004	-0.01*	-0.009	0.008	0.01	0.02*	0.01	0.009	0.004
Financing	0.006	-0.005	-0.003	-0.002	-0.001	0.01*	0.02**	0.02**	-0.003	-0.002
Default Indicators										
Arrearage (in USD)	0.57***	0.63***	0.74***	0.68***	0.57***	0.37***	0.45***	0.56***	0.60***	0.51***
Delay (in Days)	0.59***	0.66***	-	-	-	0.64***	0.73***	0.81***	0.83***	0.80***

Table 3.3 Panel A

MLEs for Four Models (Dependent Variable: Amount in Arrears (in USD)).

'Probit' indicates that Probit regression was used as the estimation method; 'Tobit' indicates that Tobit regression was used as the estimation method; '2PM' indicates that Two-Part model was used as the estimation method; and 'DH' indicates that Double Hurdle model was used as the estimation method.

	Probit	Tobit	2PM	2PM	DH	DH
			Hurdle 1	Hurdle 2	Hurdle 1	Hurdle 2
Marital Status						
Married	-0.29***	-7.19***	-0.29***	-0.54***	-0.29***	-1.24***
	(0.03)	(0.85)	(0.03)	(0.13)	(0.03)	(0.16)
Cohabitation	-0.08	-2.03	-0.08	0.50**	-0.08	0.27
	(0.06)	(1.46)	(0.06)	(0.21)	(0.06)	(0.26)
Gender (Female -> 1)	-0.04	-0.94	-0.04	-0.40***	-0.04	-0.49***
	(0.03)	(0.74)	(0.03)	(0.12)	(0.03)	(0.14)
Age	0.006	0.09	0.006	0.02	0.006	0.04
	(0.01)	(0.28)	(0.01)	(0.04)	(0.01)	(0.05)
Age-squared	-0.0002	-0.004	-0.0002	-0.0005	-0.0002	-0.0009
	(0.0001)	(0.003)	(0.0001)	(0.0005)	(0.0001)	(0.0006)
Education Levels						
Primary Completed	-0.82***	-19.46***	-0.82***	-0.27	-0.82***	-2.37***
	(0.06)	(1.50)	(0.06)	(0.19)	(0.06)	(0.30)
Secondary Completed	-0.95***	-22.67***	-0.95***	-0.36**	-0.95***	-2.82***
	(0.06)	(1.51)	(0.06)	(0.17)	(0.06)	(0.30)
Tertiary Completed	-0.93***	-22.00***	-0.93***	-0.50**	-0.93***	-2.87***
	(0.07)	(1.79)	(0.07)	(0.23)	(0.07)	(0.35)
Loan Amount (in USD)	0.00001	0.0004	0.00001	0.001***	0.00001	0.001***
	(0.00002)	(0.0004)	(0.00002)	(0.0001)	(0.00002)	(0.00007)
Maturity (in Days)	-0.00006	-0.003	-0.00006	-0.004***	-0.00006	-0.004***
	(0.0001)	(0.003)	(0.0001)	(0.0005)	(0.0001)	(0.0006)
Loan Purposes						
Consumption	-0.23*	-5.37*	-0.23*	-0.75	-0.23*	-1.33**
	(0.12)	(3.16)	(0.12)	(0.46)	(0.12)	(0.56)
Buy Fixed Asset	-0.13	-3.83	-0.13	-1.77***	-0.13	-2.37***
	(0.14)	(3.46)	(0.14)	(0.49)	(0.14)	(0.61)
Agriculture	-0.18***	-4.69***	-0.18***	0.24	-0.18***	-0.32
	(0.06)	(1.44)	(0.06)	(0.23)	(0.06)	(0.27)
Commerce	0.07	1.42	0.07	-0.003	0.07	0.12
	(0.04)	(1.08)	(0.04)	(0.17)	(0.04)	(0.20)
Manufacture	-0.04	-1.00	-0.04	-0.12	-0.04	-0.21
	(0.06)	(1.45)	(0.06)	(0.23)	(0.06)	(0.27)
Financing	0.12	3.27	0.12	-0.11	0.12	0.51
	(0.60)	(15.09)	(0.60)	(2.06)	(0.60)	(2.56)
Service (Benchmark)	-	-	-	-	-	-
Constant	-0.25	-6.57	-0.25	14.90***	-0.25	11.84***
	(0.23)	(5.77)	(0.23)	(0.88)	(0.23)	(1.09)
MFI Controls	Yes	Yes	Yes	Yes	Yes	Yes
n	17369	17369	17369	17369	17369	17369
σ	-	25.553	-	-	-	3.568
λ	0.141	0.141	0.141	0.141	0.141	0.141
Log-L (last)	-4502.7	-9278.3	-4502.7	-7371.7	-4502.7	-7359.7
K	20	21	20	39	20	40
AIC = 2*(-LogL+K)/n	0.521	1.071	0.521	0.853	0.521	0.852

*Notes: Panel A includes CACIL, INSOTEC and MICROCRED only. For comparison, the statistics in the last six rows of columns 4 and 6 are calculated for the entire models instead of the 2nd hurdles alone. * Denote statistical significance at the 10% level; ** Denote statistical significance at the 5% level; and *** Denote statistical significance at the 1% level.*

Table 3.3 Panel B

MLEs for Four Models (Dependent Variable: Delayed Repayment (in Days))

'Probit' indicates that Probit regression was used as the estimation method; 'Tobit' indicates that Tobit regression was used as the estimation method; '2PM' indicates that Two-Part model was used as the estimation method; and 'DH' indicates that Double Hurdle model was used as the estimation method.

	Probit	Tobit	2PM	2PM	DH	DH
			Hurdle 1	Hurdle 2	Hurdle 1	Hurdle 2
Marital Status						
Married	-0.17***	-4.99***	-0.17***	-0.94*	-0.17***	-1.001
	(0.05)	(1.44)	(0.05)	(0.50)	(0.05)	(0.68)
Cohabitation	-0.14**	-4.16***	-0.14**	0.15	-0.14**	0.14
	(0.06)	(1.61)	(0.06)	(0.57)	(0.06)	(0.69)
Gender (Female -> 1)	-0.15**	-4.12**	-0.15**	0.12	-0.15**	0.06
	(0.06)	(1.74)	(0.06)	(0.55)	(0.06)	(0.68)
Age	0.03*	0.86**	0.03*	0.51***	0.03*	0.55***
	(0.02)	(0.42)	(0.01)	(0.15)	(0.02)	(0.17)
Age-squared	-0.0003*	-0.01**	-0.0003*	-0.006***	-0.0003*	-0.01***
	(0.0002)	(0.005)	(0.0002)	(0.002)	(0.0002)	(0.002)
Education Levels						
Primary Completed	0.03	0.89	0.03	0.29	0.03	0.27
	(0.07)	(1.86)	(0.07)	(0.64)	(0.07)	(0.66)
Secondary Completed	0.19***	6.06***	0.20***	2.55***	0.19***	2.74***
	(0.06)	(1.78)	(0.06)	(0.63)	(0.06)	(0.66)
Tertiary Completed	0.10	3.12	0.10	1.10	0.10	1.13
	(0.08)	(2.16)	(0.08)	(0.77)	(0.08)	(0.83)
Loan Amount (in USD)	0.001**	0.002**	0.001**	-0.0004*	0.001**	-0.0005
	(0.00002)	(0.0008)	(0.00002)	(0.001)	(0.00002)	(0.002)
Maturity (in Days)	0.001***	0.02***	0.001***	-0.004***	0.001***	-0.003
	(0.0002)	(0.004)	(0.0002)	(0.001)	(0.0002)	(0.002)
Loan Purposes						
Consumption	-0.25*	-7.92**	-0.26*	-3.23***	-0.25*	-3.48***
	(0.14)	(3.83)	(0.14)	(1.13)	(0.14)	(1.36)
Buy Fixed Asset	-0.38**	-12.02***	-0.40***	-2.87**	-0.38**	-3.13*
	(0.15)	(4.15)	(0.15)	(1.21)	(0.15)	(1.62)
Agriculture	0.02	1.05	0.03	1.42	0.02	1.64
	(0.11)	(3.12)	(0.11)	(1.001)	(0.11)	(1.04)
Commerce	0.04	0.61	0.03	-0.70	0.04	-0.66
	(0.07)	(2.04)	(0.07)	(0.68)	(0.07)	(0.71)
Manufacture	0.006	-0.44	0.003	-2.43*	0.006	-2.46*
	(0.13)	(3.56)	(0.13)	(1.27)	(0.13)	(1.32)
Financing	-0.22	-6.38	-0.21	0.55	-0.22	0.82
	(0.64)	(17.52)	(0.63)	(4.69)	(0.64)	(4.70)
Service (Benchmark)	-	-	-	-	-	-
Constant	-2.57***	-74.64***	-2.59***	5.11*	-2.57***	2.45
	(0.32)	(9.34)	(0.32)	(3.07)	(0.32)	(9.28)
MFI Controls	Yes	Yes	Yes	Yes	Yes	Yes
n	14170	14170	14170	14170	14170	14170
σ	-	28.529	-	-	-	4.577
λ	0.253	0.253	0.253	0.253	0.253	0.253
Log-L (last)	-1923.4	-3692.4	-1923.4	-3367.8	-1923.4	-3362.4
K	19	20	19	37	19	38
AIC = 2*(-LogL+K)/n	0.274	0.524	0.274	0.481	0.274	0.480

*Notes: Panel B includes CACIL and FINCA only. For comparison, the statistics in the last six rows of columns 4 and 6 are calculated for the entire models instead of the 2nd hurdles alone. * Denote statistical significance at the 10% level; ** Denote statistical significance at the 5% level; *** Denote statistical significance at the 1% level.*

Table 3.4

Subsample Analysis between MFIs

The Two-Part model was used as the estimation method in this table.

	Arrearage (in USD)						Delayed Repayment (in Days)			
	CACIL		INSOTEC		MICROCRED		CACIL		FINCA	
	Hurdle 1	Hurdle 2	Hurdle 1	Hurdle 2	Hurdle 1	Hurdle 2	Hurdle 1	Hurdle 2	Hurdle 1	Hurdle 2
Married	-0.16	-0.02	-0.20***	-0.22	-0.43***	-1.06***	-0.15	-1.13	-0.21***	-0.30
	(0.11)	(0.45)	(0.06)	(0.18)	(0.07)	(0.22)	(0.11)	(0.78)	(0.06)	(0.68)
Cohabitation	-0.03	0.22	-0.23	-0.23	-0.18*	0.15	-0.02	-0.68	-0.19***	0.47
	(0.13)	(0.51)	(0.15)	(0.48)	(0.09)	(0.30)	(0.13)	(0.89)	(0.07)	(0.73)
Divorced	0.08	-1.70	0.08	0.08	-0.23*	-0.86**	0.07	-6.16	0.13	1.61
	(0.67)	(2.53)	(0.08)	(0.25)	(0.12)	(0.42)	(0.67)	(4.46)	(0.20)	(1.93)
Widowed	-0.04	-1.14	0.22	-0.58	-0.23*	-0.80*	-0.05	-0.92	-0.09	0.35
	(0.44)	(1.75)	(0.15)	(0.42)	(0.12)	(0.45)	(0.45)	(3.08)	(0.16)	(1.66)
Single (Benchmark)	-	-	-	-	-		-	-	-	-
Gender (Female -> 1)	-0.08	0.07	-0.07	-0.23	0.04	-0.33**	-0.11	-0.69	-0.18**	2.17**
	(0.10)	(0.40)	(0.05)	(0.15)	(0.04)	(0.16)	(0.10)	(0.70)	(0.09)	(0.85)
Age	0.06*	0.26*	0.006	0.02	-0.02	-0.10	0.07**	0.64***	0.003	0.50***
	(0.03)	(0.13)	(0.02)	(0.05)	(0.02)	(0.07)	(0.03)	(0.23)	(0.02)	(0.18)
Age-squared	-0.0009**	-0.003**	-0.00008	-0.0005	0.00009	0.001	-0.001**	-0.008***	0.00001	-0.006***
	(0.0004)	(0.002)	(0.0002)	(0.0006)	(0.0002)	(0.0009)	(0.0004)	(0.003)	(0.0002)	(0.002)
Primary Completed	-0.04	-0.62	-0.04	0.50	-2.42***	-0.67**	-0.05	-2.57***	0.03	2.88***
	(0.12)	(0.48)	(0.12)	(0.37)	(0.13)	(0.30)	(0.12)	(0.82)	(0.09)	(0.93)
Secondary Completed	-0.05	-0.42	-0.08	0.38	-2.48***	-0.57***	-0.08	-1.70*	0.30***	5.07***
	(0.14)	(0.55)	(0.13)	(0.40)	(0.12)	(0.20)	(0.14)	(0.94)	(0.08)	(0.85)
Tertiary Completed	-0.25	-0.70	-0.12	-0.13	-2.45***	-0.78***	-0.31	-2.24*	0.20**	3.06***
	(0.19)	(0.78)	(0.17)	(0.54)	(0.13)	(0.27)	(0.19)	(1.36)	(0.09)	(0.97)
Loan Amount (in USD)	0.00003	0.002***	0.0001	0.003***	0.001***	0.003***	0.00005	0.002*	0.001***	-0.003***
	(0.0001)	(0.0005)	(0.0001)	(0.0004)	(0.00006)	(0.0002)	(0.0001)	(0.0009)	(0.0001)	(0.0009)
Loan Amount-squared	-0.00001	-0.001***	-0.00001	-0.001***	-0.001***	-0.001***	-0.00001	-0.00001*	-0.001***	-0.001***
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
Maturity (in Days)	0.0003	-0.003***	-0.003***	-0.007***	0.00007	-0.003***	0.0003	-0.003*	0.002***	-0.005*
	(0.0002)	(0.001)	(0.0005)	(0.002)	(0.0002)	(0.0008)	(0.0002)	(0.002)	(0.0003)	(0.003)
Consumption	-0.12	0.51					-0.04	0.32		
	(0.20)	(0.81)					(0.19)	(1.39)		
Buy Fixed Asset	-0.05	-0.34					-0.02	-1.73		
	(0.19)	(0.81)					(0.19)	(1.41)		
Agriculture	0.19	2.95***	-0.20**	-0.42			0.18	3.38***	-1.19***	0.06
	(0.19)	(0.75)	(0.09)	(0.29)			(0.19)	(1.31)	(0.40)	(4.61)
Commerce	0.27	1.30*	0.03	-0.40	0.10*	-0.04	0.30*	2.19*	-0.06	-1.79**
	(0.18)	(0.72)	(0.10)	(0.31)	(0.06)	(0.21)	(0.18)	(1.25)	(0.08)	(0.82)
Manufacture	0.02	2.22	-0.06	-0.74**	-0.0007	-0.09	0.04	4.36	-0.14	-3.26**
	(0.46)	(1.82)	(0.12)	(0.35)	(0.08)	(0.29)	(0.46)	(3.21)	(0.14)	(1.39)
Financing	0.11	0.88					0.12	1.67		
	(0.65)	(2.44)					(0.65)	(4.31)		
Service (Benchmark)	-	-	-	-	-	-	-	-	-	-
Constant	-2.02**	1.29	-0.59	14.48***	1.43***	15.75***	-2.14***	0.10	-2.59***	3.32
	(0.69)	(2.74)	(0.37)	(1.15)	(0.40)	(1.42)	(0.69)	(4.79)	(0.37)	(3.86)
n	1246	191	6854	442	9269	717	1246	198	12924	299
λ	0.145	0.145	0.145	0.145	0.145	0.145	0.255	0.255	0.255	0.255
Log-L (last)	-515	-422	-1584	-817	-2166	-1517	-526	-550	-1304	-859
Pseudo R-squared	0.035		0.033		0.142		0.035		0.083	

*Notes: The first hurdle is probit regression and the second hurdle is GLM. * Denote statistical significance at the 10% level; ** Denote statistical significance at the 5% level; *** Denote statistical significance at the 1% level.*

Table 3.5 Panel A

Interactions Analysis (Dependent Variable: Amount of Arrears (in USD))

The Two-Part model was used as the estimation method in this table.

		Age			Education			
	Gender	22-29	30-39	40-49	Primary	Secondary	Tertiary	All
Hurdle 1:								
Married	-0.26***	-0.32***	-0.23***	-0.24***	-0.28***	-0.19***	-0.27***	-0.01
	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.13)
*(Married * Gender)*	-0.0005							0.03
	(0.06)							(0.06)
*(Married * Age22-29)*		0.23***						0.26**
		(0.07)						(0.10)
*(Married * Age30-39)*			-0.09					0.02
			(0.06)					(0.10)
*(Married * Age40-49)*				-0.12*				-0.02
				(0.07)				(0.10)
*(Married * Edu Lv.1)*					0.05			-0.30***
					(0.06)			(0.12)
*(Married * Edu Lv.2)*						-0.17***		-0.43***
						(0.06)		(0.11)
*(Married * Edu Lv.3)*							0.07	-0.27*
							(0.10)	(0.14)
Other controls	Added	Added	Added	Added	Added	Added	Added	Added
n	17369	17369	17369	17369	17369	17369	17369	17369
λ	0.144	0.144	0.144	0.144	0.144	0.144	0.144	0.144
Log-L (last)	-4489	-4483	-4488	-4488	-4489	-4485	-4489	-4475
Pseudo R-squared	0.054	0.055	0.054	0.054	0.054	0.055	0.054	0.057
Hurdle 2:								
Married	-0.47***	-0.51***	-0.53***	-0.61***	-0.67***	-0.33**	-0.57***	-0.42
	(0.16)	(0.13)	(0.14)	(0.13)	(0.14)	(0.15)	(0.12)	(0.43)
*(Married * Gender)*	-0.14							-0.09
	(0.22)							(0.22)
*(Married * Age22-29)*		-0.16						0.05
		(0.25)						(0.39)
*(Married * Age30-39)*			-0.04					0.12
			(0.22)					(0.37)
*(Married * Age40-49)*				0.24				0.32
				(0.25)				(0.39)
*(Married * Edu Lv.1)*					0.37			0.04
					(0.23)			(0.34)
*(Married * Edu Lv.2)*						-0.52**		-0.51
						(0.22)		(0.33)
*(Married * Edu Lv.3)*							0.17	-0.08
							(0.36)	(0.45)
Other controls	Added	Added	Added	Added	Added	Added	Added	Added
n	1350	1350	1350	1350	1350	1350	1350	1350
λ	0.144	0.144	0.144	0.144	0.144	0.144	0.144	0.144
Log-L (last)	-2819	-2819	-2820	-2819	-2818	-2817	-2819	-2816

*Notes: Panel A includes CACIL, INSOTEC and MICROCRED only. *, ** and *** Denote statistical significance at the 10% level, the 5% level, and the 1% level respectively.*

Table 3.5 Panel B

Interactions Analysis (Dependent Variable: Delayed Repayment (in Days))

The Two-Part model was used as the estimation method in this table.

		Age			Education			
	Gender	22-29	30-39	40-49	Primary	Secondary	Tertiary	All
Hurdle 1:								
Married	-0.02	-0.12**	-0.15***	-0.06	-0.13**	-0.10*	-0.12**	0.16
	(0.10)	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)	(0.05)	(0.15)
*(Married * Gender)*	-0.14							-0.14
	(0.11)							(0.11)
*(Married * Age22-29)*		-0.02						-0.11
		(0.17)						(0.19)
*(Married * Age30-39)*			0.08					-0.05
			(0.10)					(0.13)
*(Married * Age40-49)*				-0.21**				-0.24*
				(0.10)				(0.13)
*(Married * Edu Lv.1)*					0.007			-0.08
					(0.10)			(0.13)
*(Married * Edu Lv.2)*						-0.05		-0.11
						(0.09)		(0.13)
*(Married * Edu Lv.3)*							-0.03	-0.10
							(0.13)	(0.16)
Other controls	Added	Added	Added	Added	Added	Added	Added	Added
n	14170	14170	14170	14170	14170	14170	14170	14170
λ	0.257	0.257	0.257	0.257	0.257	0.257	0.257	0.257
Log-L (last)	-1903	-1904	-1904	-1902	-1904	-1904	-1904	-1900
Pseudo R-squared	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120
Hurdle 2:								
Married	-0.98	-1.14**	-1.44**	-0.99*	-1.43***	-0.93	-1.29***	0.16
	(0.78)	(0.49)	(0.59)	(0.55)	(0.54)	(0.57)	(0.49)	(0.15)
*(Married * Gender)*	-0.25							-0.14
	(0.94)							(0.11)
*(Married * Age22-29)*		-0.09						-0.11
		(1.58)						(0.19)
*(Married * Age30-39)*			0.77					-0.05
			(0.93)					(0.13)
*(Married * Age40-49)*				-0.51				-0.24*
				(0.99)				(0.13)
*(Married * Edu Lv.1)*					0.99			-0.08
					(0.96)			(0.13)
*(Married * Edu Lv.2)*						-0.56		-0.11
						(0.89)		(0.13)
*(Married * Edu Lv.3)*							1.19	-0.10
							(1.26)	(0.16)
Other controls	Added	Added	Added	Added	Added	Added	Added	Added
n	497	497	497	497	497	497	497	497
λ	0.257	0.257	0.257	0.257	0.257	0.257	0.257	0.257
Log-L (last)	-1454	-1454	-1454	-1454	-1453	-1454	-1453	-1452

*Notes: Panel B includes CACIL and FINCA only. *, ** and *** Denote statistical significance at the 10% level, the 5% level, and the 1% level respectively.*

Table 3.6

Logistic Regression Analysis Based on Credit Collection Process

	Individual PaR30					Delayed Repayment (in Days)				
	(Sample: CACIL, INSOTEC & MICROCRED)					(Sample: CACIL & FINCA)				
Screening	> 0%	≥ 5%	≥ 30%	≥ 50%	≥ 75%	> 0	≥ 30	≥ 90	≥ 180	≥ 360
Censored Clients	1350	1101	313	171	48	497	376	284	248	99
Censored Rate	7.77%	6.34%	1.80%	0.98%	0.28%	3.51%	2.65%	2.00%	1.75%	0.70%
Marital Status										
Married	0.56***	0.57***	0.33***	0.23***	0.09***	0.69***	0.56***	0.53***	0.50***	0.68
	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.08)	(0.07)	(0.08)	(0.08)	(0.16)
Cohabitation	0.85	0.89	1.03	1.05	1.23	0.74**	0.67***	0.67**	0.58***	0.79
	(0.10)	(0.12)	(0.21)	(0.26)	(0.45)	(0.10)	(0.10)	(0.11)	(0.10)	(0.21)
Gender (Female -> 1)	0.92	0.92	0.75**	0.64***	0.52**	0.74**	0.81	0.84	0.72*	0.71
	(0.06)	(0.06)	(0.09)	(0.10)	(0.16)	(0.10)	(0.12)	(0.15)	(0.14)	(0.21)
Age	1.02	1.002	0.96	0.93	0.96	1.07**	1.12***	1.13***	1.10**	1.28***
	(0.02)	(0.03)	(0.04)	(0.06)	(0.12)	(0.04)	(0.04)	(0.05)	(0.05)	(0.11)
Age-squared	0.9995	0.9998	1.0002	1.0005	1.0002	0.999**	0.999***	0.997***	0.999**	0.997***
	(0.0003)	(0.0003)	(0.0006)	(0.0008)	(0.002)	(0.0004)	(0.0005)	(0.0005)	(0.0006)	(0.001)
Education Levels										
Primary Completed	0.22***	0.16***	0.14***	0.13***	0.17***	1.05	1.15	1.22	1.04	0.84
	(0.02)	(0.02)	(0.03)	(0.03)	(0.10)	(0.16)	(0.20)	(0.24)	(0.23)	(0.28)
Secondary Completed	0.16***	0.12***	0.09***	0.10***	0.23***	1.48***	1.95***	2.14***	2.28***	2.01**
	(0.02)	(0.014)	(0.02)	(0.02)	(0.11)	(0.21)	(0.33)	(0.41)	(0.46)	(0.60)
Tertiary Completed	0.17***	0.12***	0.08***	0.06***	0.06***	1.18	1.35	1.30	1.40	1.62
	(0.02)	(0.02)	(0.02)	(0.02)	(0.05)	(0.21)	(0.29)	(0.31)	(0.35)	(0.59)
Loan Amount (in USD)	1.0002	1.0001	1.0001*	1.0002**	1.001***	1.0001	1.0001	1.0001	1.0001	1.0001
	(0.00003)	(0.00004)	(0.00007)	(0.00009)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Maturity (in Days)	0.9999	0.998***	0.994***	0.993***	0.99***	1.001***	1.0006	1.0001	0.9997	0.9991
	(0.0002)	(0.0004)	(0.0009)	(0.001)	(0.002)	(0.0003)	(0.0003)	(0.0005)	(0.0006)	(0.0008)
Loan Purposes										
Consumption	0.65*	0.20				0.66	0.47**	0.25***	0.12***	0.20**
	(0.16)	(0.21)				(0.17)	(0.15)	(0.12)	(0.09)	(0.16)
Buy Fixed Asset	0.74	0.35				0.56**	0.29***	0.18***	0.19**	0.17
	(0.19)	(0.37)				(0.15)	(0.11)	(0.11)	(0.15)	(0.19)
Agriculture	0.69***	0.63***	0.60**	0.77	1.53	1.15	1.35	1.70*	1.77*	0.84
	(0.08)	(0.08)	(0.14)	(0.26)	(1.70)	(0.26)	(0.33)	(0.48)	(0.56)	(0.42)
Commerce	1.14	1.04	1.09	1.09	0.73	1.13	0.97	0.90	1.05	1.22
	(0.10)	(0.10)	(0.20)	(0.27)	(0.27)	(0.19)	(0.17)	(0.18)	(0.23)	(0.42)
Manufacture	0.95	0.87	0.74	0.57		1.07	0.87	0.68	0.62	0.30
	(0.11)	(0.11)	(0.19)	(0.22)		(0.31)	(0.29)	(0.27)	(0.28)	(0.32)
Financing	1.27					0.78	0.91	1.71		
	(1.41)					(0.88)	(1.03)	(1.95)		
Service (Benchmark)	-	-	-	-	-	-	-	-	-	-
Constant	0.71	2.02	11.15**	25.09**	7.01	0.006***	0.002***	0.002***	0.003***	0.001***
	(0.34)	(1.07)	(10.7)	(32.08)	(-17.4)	(0.004)	(0.002)	(0.002)	(0.003)	(0.0001)
MFI Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
n	17369	17363	16123	16123	14194	14170	14170	14170	14164	14164
Log-L (last)	-4499	-3863	-1386	-825	-260	-1933	-1574	-1303	-1181	-554
Pseudo R-squared	0.052	0.058	0.102	0.130	0.192	0.102	0.093	0.064	0.055	0.060

*Notes: *, ** and *** Denote statistical significance at the 10% level, the 5% level, and the 1% level respectively.*

Chapter 4:

What Drives Financial Awareness in Microfinance?

A Cross-Section Analysis

4.1 Introduction

As new products and services become more widespread, financial markets around the world have become increasingly accessible to smaller investors. People with credit cards, subprime mortgages and etc. were in the historically new position of being able to decide how much they wanted to borrow. The customers of financial services have much greater power and responsibility for decision making than before. On the other hand, financially complex products have proven to be difficult for financially unsophisticated investors to master. They impose a heavy burden on the individuals and households to achieve a higher level of financial literacy, which is the prerequisite of sound financial decisions.

The importance of financial literacy has been mentioned in many studies. For example, Utkus & Young (2010) and Mottola (2013) indicate that the least financial savvy incurs higher transaction costs, high fees and using high-cost of borrowings. Numerical ability, which is part of financial literacy, shows strong predictive power for mortgage defaults even after controlling for cognitive ability and general knowledge (Gerardi et al., 2013). Both self-assessed and actual financial literacy are found to affect the clients' credit card behaviour, such as the probability of defaults, over the life cycle (Allgood & Walstad, 2013). Last but not the least, financial literacy also impacts the on retirement, and it is associated with greater retirement planning and retirement wealth accumulation (Ameriks et al., 2003; Van Rooij et al., 2012).

The price of being financial illiterate can be very high. For instance, investors are estimated to have foregone substantial equity returns due to the ignorance of fees, expenses, and active investment trading costs, in an attempt to beat the market (Cocco et al., 2005; Calvet et al., 2007). Costs of financial ignorance arise not only in the saving and investment areas, but also influence how consumers manage their liabilities (Campbell, 2006; Meier et al., 2013).

In fact, nearly all respondents in the U.S. consider that well understanding finance is very important (Markow, 2005). However, the current average level of financial literacy is very low in general. Many investors (including college employees, local construction works, local tourists, and parents of students) lack basic financial literacy, such as knowing the differences between stocks and bonds and the fundamentals of mutual funds (Hancock, 2002). This low level of financial literacy is prevalent everywhere. In Germany, Australia and France, those who could answer all basic financial questions in the quasi-experiments are just 53%, 43% and 31% of respondents respectively (Bucher-Korenen & Lusardi, 2011; Agnew et al., 2012; and Arrondel et al., 2013). In the recent OECD survey of adult financial literacy competencies, Atkinson et al. (2016) show that only 56% of adults across the 29 participating countries and economies achieved their minimum target score, only 42% of adults are aware of the benefits of interest compounding on savings, and only 58% could calculate the simple interest on savings (See Appendix. B). Similar statistics can be found in the S&P Global FinLit Survey as well. In this survey, Klapper et al. (2015) indicate that only 33% of adults in 140 countries across the world are financially literate, which means around 3.5 billion adults, most of them in developing countries, lack an understanding of basic financial concepts. In addition, a great number of mortgage defaults during the financial crisis has suggested that debt management is another fertile area for mistakes. Many borrowers do not even know what interest rates are charged on their borrowings (Moore, 2003). It is not an issue of financial knowledge but an issue of financial awareness.

This paper sets out to be the first rigorous global study of the relation between a client's individual/household level characteristics and financial awareness of interest repayment, using a data set including 9,053 clients of 51 microfinance institutions (MFIs) from 27 underdeveloped or developing countries. Financial awareness is the most important component of financial literacy, in which there can be all kinds of issues that are necessary to make a rational financial decision, such as financial knowledge and skills. The reason why we prefer financial awareness above other issues will be discussed in the next section. The main contribution of this paper is providing a potential method to estimate the financial awareness of borrowers, and bridging the gap in the current measurement framework of financial literacy. The rest of the paper proceeds as follows: Section 2 reviews the literature on financial literacy and presents the hypotheses to be tested. Section 3 describes the data, the imputation methods, and the statistical methods employed. Section 4 reports empirical findings and discussion. Section 5 presents the study's conclusions and limitations.

4.2 Literature Review

4.2.1 Financial awareness and the framework of financial literacy

According to the current literature, there are no universally accepted definitions of financial literacy and financial awareness. Researchers of the prior studies have to establish their framework (and questionnaires) for financial literacy or awareness measurement. In recent years, some simple but very effective frameworks have arisen. A large number of empirical studies have measured financial literacy based on a framework with three basic financial questions designed by Lusardi and Mitchell (2008). They are as follows:

- Interest Rate Question - Suppose you had $100 in a savings account and the interest rate was 2 percent per year, after 5 years, how much you would have if you left the money to grow?
- Inflation Question - Imagine that the interest rate on your savings account was 1 percent per year and inflation was 2 percent per year. Would you be able to buy the same as today's money in this account after 1 year?
- Diversification Question - Do you think that the statement ''Buying a single company stock usually provides a safer return than a stock mutual fund'' is true or false?

However, as defined by INFE (2011), financial literacy is:

'A combination of awareness, knowledge, skill, attitude and behaviour necessary to make sound financial decisions and ultimately achieve individual financial well-being.'

It considers financial awareness as an essential component of financial literacy. Based on this definition, it is clear that financial awareness is missing from the framework presented above. In line with INFE's (2011) definition, Carpena et al. (2011) claim that financial literacy programmes can affect financial decision-making through other channels, by making individuals or households more aware of their financial conditions and available product choices, and thus reshaping their attitudes (e.g., confidence in ability and risk preference) towards financial activities.

Regarding the prior literature, most empirical studies are experimental and based on Lusardi and Mitchell's (2008) framework or similar research designs to measure financial literacy. The influences of financial awareness are excluded from those studies, in which they focused on assessing the potential effects of certain financial education programmes provided by various institutions. Therefore, it is acceptable to ignore financial awareness in their studies.

However, if we want to answer the question whether financial literacy can ultimately improve individuals or households financial decisions, or reduce loan defaults, then financial awareness is as essential as financial knowledge.

At last, it is important to clarify that the boundary between awareness and knowledge sometimes is vague in practice, especially when the learning cost of a knowledge point is close to nothing. By looking at the descriptive statistics (**Table 4.1**), we found that over 60% of clients cannot remember the interest rates or the interest amounts of their micro loans. This finding motivated us to use these variables as indicators of financial literacy. However, whether 'accurately-remember-interest' is a type of awareness or knowledge is controversial. Therefore, in the literature review and the rest of the paper, we consider financial literacy, financial knowledge, and financial awareness as the same thing.

4.2.2 Main findings in the previous studies

The majority of previous financial literacy studies focus on one or some of the independent variables such as gender, age, education background, and living location, while other factors (race and ethnicity, nationality, religion, family background, employment status, etc.) were seldom mentioned.

In terms of the findings related to age, lots of evidence shows that young respondents are generally more financially knowledgeable than older respondents. However, instead of a low level of financial literacy, some studies indicate that overestimation of financial literacy might be the real cause of poor performance and defaults. Older people usually score poorly on basic financial literacy questions in surveys. Nevertheless, older people also give themselves very high scores on financial literacy self-assessments (Lusardi & Tufano, 2009a; Lusardi & Mitchell, 2011a, 2011b). Similarly, Finke et al. (2011) indicate that people's confidence in their financial decision-making abilities increases with age while financial literacy falls with age. Such a mismatch between the actual and perceived financial knowledge might potentially explain why financial frauds are often perpetrated against the elderly (Deevy et al., 2012).

There are also a great number of experimental studies arguing that for both the elderly and young people, men are generally more financially knowledgeable than women (Hung et al., 2009; Lusardi et al., 2009; Lusardi & Mitchell, 2009; Lusardi & Tufano, 2009a, 2009b). Some studies have examined this result in samples with highly educated respondents exclusively.

Regarding high schools and colleges specifically, men are still more financially knowledgeable than women (Chen & Volpe, 2002; Mandell, 2008). Even for well-educated women, financial literacy was found to be very low (Mahdavi & Horton, 2014).

A number of studies have tried to explain the reasons for such a phenomenon. Brown and Graf (2013) claim that the differential interest in finance and financial matters between women and men is not one of the reasons. In fact, some sex differences may be rational due to specialisation of labour within households. Married women usually only build up financial knowledge late in their lives for unavoidable reasons, such as when their husbands pass away (Hsu, 2016).

However, why single women also have lower financial literacy than men has not been answered yet. Fonseca et al. (2012) suggest that women may acquire financial literacy differently from men while they have the approximately equal educational achievement. In addition, the different self-confidence of financial literacy might be a potential explanation of the weaker literacy score of female respondents in the research (Bucher-Koenen et al., 2012).

Besides the major findings illustrated above, there are also some interesting findings on education background and living location. Both of them have been proven to be related to financial literacy in most studies. Those without a college education are much less likely to be knowledgeable about basic financial literacy compared to college graduates (Lusardi & Mitchell, 2007, 2011b). Such a positive correlation between educational background and financial literacy might be driven by cognitive ability (McArdle et al., 2009). Numeracy is especially poor for those with low educational attainment. But we should notice that the cognitive factor does not fully account for the variance in financial literacy, though it always has a significant and much higher coefficient than most variables (Lusardi et al., 2009). On the other hand, regarding the factor of living location, it is found that those living in rural areas generally score worse in financial literacy than their city counterparts. A possible reason is that financial literacy is more easily acquired via interaction with others in the same areas (Klapper et al., 2011). Formal financial education is probably not the primary route for clients to absorb relevant knowledge and necessary information.

In terms of the associations between access to credit and financial literacy, most empirical studies use financial literacy as a potential determinant of access to credit, instead of as a dependent variable. Regarding the influence of access to credit on financial literacy, the only

relevant paper is conducted by Guiso and Jappelli (2005). They analyze the 'attention to information barriers' for limited financial market participation and found that financial awareness to household characteristics is positively associated with socioeconomic variables (education, wealth, income, and age), the intensity of social interaction, and most importantly, long-term bank relations. They also indicate that the most plausible interpretation of the significant relation between long-term bank relations and awareness is that: the banks have a greater incentive to inform that clients that they have superior information. It inspires us to examine whether previous access to microfinance services is associated with financial awareness as well.

4.2.3 Limitations of the previous studies

There is no lack of financial literacy studies that use experimental methods in the current literature. However, there are obvious limitations of the experimental studies. Very few of them have undertaken a rigorous evaluation of the impact of financial education based on a suitable theoretical model which shows how financial literacy is accumulated, or a carefully-designed empirical quasi-experiment approach. Consequently, the impact of education programmes has been greatly underestimated. Fernandes et al. (2014) point out that financial education explained only 0.44% of the variance in financial knowledge, and such effect is much weaker than the other domains of education. In addition, while we have strong evidence that financial literacy is generally low, some evidence that financial illiteracy has a negative influence in the prior literature, there is no evidence that financial literacy can be increased by education interventions in a cost-effective way until now (Karlan et al., 2014).

While the experimental and quasi-experimental methods are more popular in the financial literacy studies, we consider that the non-experimental method might be more appropriate in our study. The major criticism of non-experimental methods is that there are no answers given regarding why particular groups have lower levels of financial literacy in the studies. In other words, the high correlation between an indicator and financial literacy does not necessarily reflect a causal relationship between them. On the other hand, exclusive of selection bias is the main advantage of non-experimental studies comparing to experimental ones. In terms of experimental methods, there are a couple of disadvantages. Firstly of all, it is extremely difficult to conduct an experimental or quasi-experimental study in a large number of countries and generate cross-sectional data. It is difficult to control all potential unobservable influencers in our case. Secondly, any test of the effect of experimental education programmes on subsequent economic behaviours is designed for particular courses. Without

further testing in different circumstances, the applicability of the findings to more general occasions is unclear. The third issue of experimental study is the difficulty created by potential selection bias. Changes in observed economic behaviours may just reflect the motivation or aptitude of participants rather than a programme's effectiveness. Finally, impacts of the experimental programmes may need a long time to take effect. It is tough to measure the long-term outcomes and behavioural changes. In contrast, significant associations are much easier to be detected in the non-experimental studies.

4.2.4 Motivations and research questions

Considering to the characteristics of microfinance (comparably higher interest rate, shorter repayment cycle, and lower level of collaterals), financial awareness might be even more crucial than financial knowledge to a borrower's probability of default. On the other hand, a generally low level of financial literacy, and very strong associations between socio-demographic characteristics and financial literacy have been found in a number of empirical studies. Therefore, it would be interesting to know if these findings still hold for financial awareness solely, and what features of the clients may relate to higher financial awareness. The hypotheses tested in this paper can be described as follows:

H1.1 Women have a lower probability of being aware of their interest rate

H1.2 Older borrowers have a lower probability of being aware of their interest rate

H1.3 Less educated borrowers have a lower probability of being aware of their interest rate

H1.4 Borrowers living in rural areas have a lower probability of being aware of their interest rate

In addition, as most of the prior financial literacy studies focus on U.S. citizens, there is insufficient evidence in the context of underdeveloped or developing countries, and no cross-country analyses available now. Therefore, we set out to fill this gap by using a cross-country survey data covering 27 countries and conduct split-sample examinations for different regions and religions as well to see if **H1.1** to **H1.4** are persistent.

We also propose the following main hypotheses as we expect that experience of financial services may help to improve the financial awareness of the microfinance participants:

H2. Clients who have held savings accounts before (or previously accessed to moneylenders, previously accessed to MFIs, previously accessed to formal banks) have a higher probability of being aware of their interest rate

H3. Borrowers who have no education, but have held savings accounts before (or previously accessed to the moneylenders, or previously accessed to MFIs, or previously accessed to formal banks) have a higher probability of being aware of their interest rate.

4.3 Data and Methodologies

4.3.1 Descriptive statistics

The individual level survey data is provided by Micro Finanza Rating, which is a leading private and independent international rating agency specialized in microfinance. It consists of 9,053 clients of 51 MFIs from 27 underdeveloped or developing countries (See Appendix. C). 180 clients were randomly selected in each MFI. All surveys included in this paper were conducted in the period from 2007 to 2012. A major advantage of using this survey data set is that it covers a wide range of unique client characteristics that have been ignored in the former microfinance literature. For example, the data contains information on the clients' financial awareness of interest repayment, and their previous access to different sources of credit, such as moneylenders, MFIs, formal banks and etc. Besides, the survey data also included those essential variables have been widely studied before, such as age, gender, and education background.

Previous studies have designed various qualitative questionnaires to measure the clients' awareness of the interest repayment. For instance, the INFE (2011) has developed the OECD financial literacy questionnaire. It considered awareness as an indispensable component of financial literacy. However, the designed questions are subjective as shown below (scale of 1 to 5, completely agree 1, completely disagree 5):

- Before I buy something, I carefully consider whether I can afford it.
- I tend to live for today and let tomorrow take care of itself.
- I find it more satisfying to spend money than to save it for the long term.
- I pay my bills on time.
- I am prepared to risk some of my own money when saving or making.
- I keep a close personal watch on my financial affairs.
- I set long-term financial goals and strive to achieve them.
- Money is there to be spent.

In order to quantify financial awareness and cope with the disadvantages brought by the unstandardized measurements as presented above, the two proxies for literacy used in this paper are completely irrelevant to the clients' honesty, numeracy, and accessibility of product information, but only memory and awareness of interest repayment. Hence, they are unbiased and more reliable compared to the other measurement frameworks that have been widely used. Micro Finanza Rating has constructed two indicators as follows:

$$KR_i = \begin{cases} 1, & for\ \frac{|R_i - \widetilde{R_i}|}{R_i} \leq 0.25 \\ 0, & for\ \frac{|R_i - \widetilde{R_i}|}{R_i} > 0.25 \end{cases} \quad (19)$$

$$KA_i = \begin{cases} 1, & for\ \frac{|A_i - \widetilde{A_i}|}{A_i} \leq 0.25 \\ 0, & for\ \frac{|A_i - \widetilde{A_i}|}{A_i} > 0.25 \end{cases} \quad (20)$$

where KR_i and KA_i are a pair of binary variables that indicate whether client i can accurately remember his/her interest rate and total interest payment; R_i and A_i are the interest rate and total interest payment of client i actually recorded on the MFIs' administrative loan book; $\widetilde{R_i}$ and $\widetilde{A_i}$ are the interest rate and total interest payment reported by client i during the surveys; KR_i and KA_i equal to 1 when the absolute difference between the actual and reported values is no greater than one-fourth of the actual value.

At first glance, there is no fundamental difference between KR_i and KA_i. Because the clients can easily calculate one another with the knowledge of total loan amounts. Nevertheless, regarding to the most financially knowledgeable respondents, only 36% of them can accurately report both numbers. It means that a large proportion of the rest (64%) only pay attention to either interest rate or total interest payment for unidentified reasons. Therefore, KR_i and KA_i are not interchangeable, and they need to be treated individually. In this paper, we have used KR_i as the major proxy for financial awareness to conduct regression analyses, and used KA_i in robustness tests only. It is because the size of the KA_i sample is three times smaller than that of the KR_i sample. Considering the issues of error and missing data, the results based on the KA_i sample might be less convincing.

Table 4.1 reports the sample size, percentage of missing data, minimum, maximum, standard deviation, skewness and kurtosis for all variables in our sample. We see that there are 38% of respondents can remember the interest rates, and 34% of them can remember the total interest payment. In terms of who can remember either interest rates or total interest payments, there are still only 48% of the entire sample. In general, more than one-half of

the microfinance participants in our sample were not knowledgeable of their interest repayments. The widespread financial unawareness may be more intimidating than financial incapability in the microfinance sector and lead to lower repayment rate.

As far as the access to credit variables are concerned, the means of previous access to different sources of credit are not very high. Only 17% of participants in our sample have experience of borrowing from MFIs. 21% of them have accessed formal banking before, while 10% of them have tried to borrow from relatives, friends or moneylenders. Moreover, we can see that a noticeable number of clients have accessed more than one credit source at the same time. 8% of clients have borrowed from two different MFIs. Besides, 3.6% of them even have debts with moneylenders. These clients were probably using new loans from the MFIs to pay off old debts, and suffering from over-indebtedness.

For the gender variables, we see that, on average, MFIs have 59% female clients. In terms of the ordinal variable of women's control of loans, 1 means partial control and 2 means completed control. As the mean of women's control of loans is slightly less than 1 and the mean of gender is closed to 0.6, it indicates that more than 17% of the loans, where borrowers were female, were actually controlled by their husbands. Regarding the educational background, 1.55 means the majority of clients were graduates from primary and secondary schools. About 27% of the client have studied at tertiary schools, while 16% of them have never engaged in any formal education. In brief, the average level of education is relatively low, especially for the users of financial services.

4.3.2 Missing data imputation methods

Until recently, listwise deletion is still the most popular way of dealing with missing data. It simply eliminates any cases with missing data or errors on one or more of the variables. In reality, the percentage of cases missing have to be carefully examined if listwise deletion is implemented. As a rule of thumb, datasets in which more than roughly 20% are excluded by deletion might lead to substantial bias in estimations. As shown in **Table 4.1**, there are four variables (KA_i, age, employment status, and income per capita) that have over 20% missing data. By applying listwise deletion to all variables in our data set, the sample size will decrease by more than 60%. Besides, as the four variables stated above have been claimed to be related to financial literacy in the previous studies, we cannot simply drop them out either. Therefore, listwise deletion is infeasible in this situation, and a missing data imputation method is needed.

Based on the characteristics of missing values and our research objectives, the Multiple Imputation (MI) method is the optimal solution in this case due to three reasons: 1. the numbers of missing data on some variables are substantial; 2. the correlations between the variables with missing data and the other variables can be well estimated; 3. the real relationships between variables are much easier to be detected as the MI method will strengthen the correlations and preserves the distributions. An in-depth explanation and discussion of how to choose the best approach to handle missing data can be referred to chapter 5.

In this paper, we follow Little and Rubin's (1987) framework for missing data, which was specified by Schafer and Graham (2002), assume the incomplete data of variables are missing at random (MAR), and then apply the MI method to all variables with missing data, except for the dependents KR_i and KA_i. In terms to the iterations of imputation, based on the recommendations from Graham et al. (2007), Bodner (2008) and White et al. (2011), we consider that 50 times will be sufficient to yield more than 95% of efficiency of missing data estimation.

However, the MI method is far from perfect. The major downside of it is reducing generalisation of the sample and over-stating the actual correlations. Because the method predicts the incomplete variables by stochastic regression with estimations of the means and the covariances which may not persist in the actual missing data. Hence, we also estimate the missing data conservatively by the traditional mean imputation method, rerun the regressions, and finally compared the results to see if there are potential false significant relationships. The mean imputation will significantly attenuate the overall correlations estimated (Baraldi and Enders 2010). On the other hand, it might damage the distributions of the incomplete variables and over-estimate the correlations for those complete variables. As results, we should conclude that the estimated correlations are unbiased (neither over-fitted nor under-fitted) only when they are consistent in both data sets with different imputation methods. By comparing the descriptive statistics of the raw data (**Table 4.1**) and the pooled data of 50 multiple imputations (**Table 4.2**), we can see no substantial differences on the means and distributions of these two data sets. The imputation is statistically reliable and valid.

4.3.3 Estimation methods

The research design of this study involves logit regression analyses with cross-sectional data. Logit regression is used because the dependent variable KR_i is dichotomous. Our estimation is based on a choice-based sample in which 38% of the clients have financial awareness of

interest rates and 62% of them have no financial awareness. These percentages are determined by the indicator's evaluation standard (less than 25% difference from the true value) that designed by Micro Finanza Rating. If we raise the barrier from to 25% to 5%, there will be almost no respondents can be classified as financially knowledgeable. Therefore, the process used here slightly differs from a pure random sampling approach.

The unequal sampling for the two groups will finally lead to the bias in the constant term, which is a compulsory component to build a predictive model. Fortunately, model development is not the purpose of this study. Maddala (1991) claims that the unequal sampling rates do not affect the coefficients of the explanatory variables but the constant term. The weighting procedure is unnecessary if we just perform a logit analysis. Inspired by the empirical designs in Guiso and Jappelli (2005) and Lusardi & Tufano, (2009a), we test the hypotheses between the clients' previous access to credits and financial awareness of interest rate described in H1 and H2, and regress KR_i with controls as follows:

$$KR_i = \alpha + \beta_1 PSAVE_i + \beta_2 PMONEY_i + \beta_3 PMFI_i + \beta_4 PBANK_i + \beta_5 \boldsymbol{L}_i + \beta_6 \boldsymbol{F}_i + \beta_7 \boldsymbol{S}_i + \beta_8 COUNTRY_i + \varepsilon_i \quad (21)$$

where $PSAVE_i$, $PMONEY_i$, $PMFI_i$, and $PBANK_i$ are dummy variables that indicate whether client i has opened saving account, borrowed from moneylenders, MFIs and formal banks before he/she accessed to the current MFI respectively; $\boldsymbol{L}_i$ is matrix of loan-specific controls, such as annual interest rate, loan size, and extra loans from other moneylenders or institutions; $\boldsymbol{F}_i$ is a matrix of individual level financial-specific variables that capture employment status, number of fixed income sources, income per capita, and ownership of properties; matrix $\boldsymbol{S}_i$ consists of a set of socio-demographic characteristics, such as gender, age, education background, and living location, because empirical studies have shown that the level of financial literacy is related to these individual or household features; vector $COUNTRY_i$ controls the country in which an MFI is active.

The effects of previous access to credits exerted on the financial awareness of clients can be more prevalent under certain conditions or apply more for certain categories of socio-demographic characteristics. In order to examine the heterogeneous effects, interaction terms in the regression equations are therefore included as follows:

$$KR_i = \alpha + \beta_1 PSAVE_i + \beta_2 PMONEY_i + \beta_3 PMFI_i + \beta_4 PBANK_i + \beta_1' PSAVE_i * INT_i + \beta_2' PMONEY_i * INT_i + \beta_3' PMFI_i * INT_i + \beta_4' PBANK_i * INT_i + \beta_5 \boldsymbol{L}_i + \beta_6 \boldsymbol{F}_i + \beta_7 \boldsymbol{S}_i + \beta_8 COUNTRY_i + \varepsilon_i \quad (22)$$

where $PSAVE_i * INT_i$, $PMONEY_i * INT_i$, $PMFI_i * INT_i$ and $PBANK_i * INT_i$ are the interaction terms that measure whether the effects of previous access to saving service, moneylenders, MFIs and formal banks differ with the interaction variables INT_i respectively. The major interaction variables we include in this analysis are education background, living location, and whether a client is the head of household. Since prior empirical studies have found evidence that people who live in rural areas and with lower education levels are much less likely to be knowledgeable about basic financial literacy, as discussed in the literature review section, it is interesting to see if providing financial services will have stronger influence on these particular clients. In addition, as the heads of households usually take more control over financial issues and decisions than their counterparts, it is reasonable to suppose that the influence on them will also be more noticeable.

4.4 Results and Discussion

4.4.1 Relation between access to credits and financial awareness

Table 4.3 reports the impacts of the clients' previous access to credit on their financial awareness of their interest rate. The different columns correspond to the different missing data imputation methods (multiple imputation vs mean imputation) and different control variables. Columns (1)-(4) present the results of regressions that only include the socio-demographic factors which have been widely studied before. We see that women have a higher awareness of their interest rate. But this relationship becomes insignificant when we control for the countries. Alternatively, when we replace gender with the women's actual control power on loans in columns (9)-(12), and there are no significant results either. On the hand, the clients living in the rural area are more cautious of the interest rate at the 1% level.

These two findings are different from most of the present literacy studies, such as Chen et al. (2002), Hung et al. (2009), and Lusardi et al. (2009), who demonstrate that men are generally more financially knowledgeable than women, and people living in rural areas generally score worse in financial literacy. The different results can be simply caused by sample difference, as these surveys are conducted in the U.S. Otherwise, it is probably because the former literature has excluded the awareness of interest from the measurement of financial literacy. Alternatively, it could suggest that the difference between genders is absorbed by the control variables of the countries. In most cases, there is only one MFI in a country, and

some MFIs provide services to female clients exclusively. Regarding the impact of gender, regression results without country controls might be more reliable in this case. We will further examine it in the later subsection of split-sample analysis.

In terms of the other two variables, age and education background, the results are consistent with prior results in the financial literacy studies. Significant results are found for these two variables. A possible interpretation of the results is that the respondents who are older and less educated are more likely to forget their interest rate of repayment.

In columns (5)-(8) of **Table 4.3**, we examine the impacts of previous access to financial services with variables of loan status, financial status, and the socio-demographic variables introduced above. As can be seen, previously borrowing from moneylenders, friends, relatives and family members is highly significant at the 1% level. Coefficients are at least 0.36, regardless of the estimation model. In other words, the clients' odds to remember the interest rate is 43% higher than who never accessed to money lenders and etc. In contrast, the coefficients of previous access to saving service are negative at the 1% level. It means that the clients who have had saving accounts before are less likely to know the interest rate. In addition, regarding previous access to MFIs, and formal banks, there are no significant results. According to our data, the average interest rate of MFIs is 27%, which is much higher than the general deposit rate (2%) and borrowing rate (5%) of formal banks, but much lower than the general interest rate of moneylenders (from 90% to 180%). In brief, it seems that the extremely high borrowing rate, and maybe unpleasant experience, will strongly improve the clients' financial awareness, while the extremely low saving rate and satisfactory banking service could weaken their attention to interest rates.

Further examining the other variables of household characteristics, loan status and financial situation, we see that mainly annual interest rate, loan size, income per capita, extra loans from other MFIs, and whether client is a household head are related to KR_i. In particular, a higher KR_i is associated with a lower interest rate, a larger size of loan, a higher income, as well as borrowing from more than one MFIs. Finding that the interest rate is negatively related to financial awareness is surprising, since greater financial burden should force the debtors pay more attention on their loans. The marginal effect of a 1% increase in interest rate is a 90% decrease in the odds of KR_i. Considering the coefficients between KR_i and previous access to moneylenders (0.36), previous access to MFIs (0.04), and previous access to formal banks (-0.05), it is reasonable to infer that there might be a nonlinear (convex) relationship between interest rate and KR_i. Further examination is needed but this is beyond

the purpose of our study. Moreover, it is interesting to find that the clients who borrowed from multiple MFIs actually have lower odds to remember the interest rates. Such finding could indicate that the ignorance or underestimation of interest repayment might be one of the reasons to excess borrowing and even over-indebtedness.

4.4.2 Split-sample Analysis

Table 4.4 presents the results of the split-sample analysis, where the main regression is repeated for different regions in which the MFIs are active and different religions of the respondents. We analysis the microfinance borrowers classified into four groups (Africa, Catholic Europe, Latin America, and the Middle East) based on two dimensions of cross-cultural variation in the world[2]: 1. from Traditional to Secular-Rational; and 2. from Survival to Self-Expression. For instance, the clients in the Middle East usually emphasize the importance of traditional value, economic and physical security, while the clients in Europe are more secular-rational and have stronger motivations to pursue self-expression. We also classify the clients into three categories regarding the dominant religion of where they belong to. If the dominant religion takes up less than 40% of the population, the related areas will be identified as a mixture (a great diversity of beliefs). Columns (1)-(2) and (11)-(12) present the regressions with variables of regions or religions. It is shown that they are all significant at the 1% level. Respondents who are Islamist or live in the Middle East have much highest odds to correctly remember the interest rates. In comparison, the odds for those from Africa or the countries with multiple religions are lowest.

Columns (3)-(10) and (13)-(18) of **Table 4.4**, reconfirm that, in Latin America, Middle East, and Christian countries, KR_i is positively associated to previous access to moneylenders, and negatively associated to previous access to saving service. However, we have found inconsistent results in other areas. Previous access to moneylenders has no significant influence on KR_i in the countries with multiple religions. In Africa, previous access to saving service is insignificant. In Europe, previous access to all sources of credit (moneylenders, MFIs and banks) has negative impacts on KR_i and the impact of previous access to saving service has become positive. One potential explanation for this result is that Europe is a more regulated

[2] This is a relative scoring method instead of a qualitative description for cultural values. For example, the people in Catholic Europe is more secular than those in Middle East, meanwhile, 72% of respondents in the Eurobarometer Survey (2012) described themselves as Christianity. Further details of the definitions and scoring method can be referred to the World Values Survey (http://www.worldvaluessurvey.org/WVSContents.jsp).

financial market than the other regions. Considering the potential convex relationship between interest rate and KR_i discussed above, if the upper limits of interest rate of all moneylenders, MFIs and banks are restricted to very low levels (at the left-hand side of the vertex), then the coefficients of previous access to credits will become negative.

Examining the socio-demographic variables once again, we found different results compared to those introduced in subsection 4.4.1. From columns (3)-(10) of **Table 4.4**, we see that gender becomes insignificant in Europe, Middle East, and the countries which are Islamic or have mixed religious. On the other hand, we found significant positive relations between women and the awareness of interest rate in Latin America and Christian countries. In these areas, while men are generally more financially knowledgeable than women, women are more financially cautious than men. In fact, the majority of Christians concentrate in Latin America nowadays. Location and religion are highly correlated. Both Christianity and Islam well recognise marriage and consider a woman's primary responsibility is to fulfil her role as a wife. But Islam also stresses that a woman's responsibilities to nurture, educate, and protect her children have taken priority over working and financially support, where women still have the right and are free to work. It may be the reason why women in Christian countries have higher financial awareness than men.

Moreover, age becomes insignificant across all split-samples with multiple imputations after releasing the control of countries. This result is inconsistent with the prior results presented in **Table 4.3** as well. With mean imputations, significant results for age are found in columns (4), (8) and (16). Considering the extremely high percentage of missing values in age (40%, see **Table 4.1**), it is reasonable to guess that the results with multiple imputations are more accurate than those with mean imputation. However, we leave the association between age and financial literacy as unknown in the paper for robustness. Because the accuracy of multiple imputations for a discrete and censored variable with very high missing rate is still unclear. In fact, this issue motivates us to evaluate the imputation performances of different missing data techniques. It will be discussed in details in the next chapter.

What is more, in terms of educational background and living location, the results are also consistent with those shown in Table 3, except for those estimated with the split-sample of Europe, in which both variables have no significant relationship with KR_i. The results with multiple and mean imputations are the same. Therefore, we can conclude that hypotheses 1.3 and 1.4 are accepted.

4.4.3 Interaction effects

Regression outputs with respect to the interaction terms are shown in **Table 4.5**. The different columns represent the different interaction terms (education background, living location, and whether a client is household head) that are added subsequently. In order to capture interaction effects of specific levels of education, we transform the ordinal variable of education to three dichotomous variables before analysis. Note that the coefficient for clients with previous access to different types of credit, now represents the relation between previous access to credits and financial awareness of interest rate in the reference category (Part 3 and Part 4 of **Table 4.5**), whereas the sum of the reference coefficient and the coefficient for the interaction term is the one actually indicate the true relation to the dependent variable.

Based on the prior finding that previous access to MFIs and formal banks has no relation to KR_i, we are curious as to whether it has an influence on the particular clients with comparably lower levels of education. As can be seen from columns (1)-(6) at Part 3 of **Table 4.5**, almost all interaction terms between education and previous access to credits return insignificant coefficients, which indicates that the relation between previous access to credit and KR_i does not differ with education background. The only exception is the interaction term between the uneducated clients and previous access to saving service. It is significant at the 1% level and the coefficient is negative 0.48. Along with the reference coefficient of previous access to savings (-0.07), we may conclude that providing an access to saving service to the clients with at least primary education may potentially strengthen their financial awareness of interest repayment. However, access to saving service is not good for uneducated clients.

On the other hand, it is surprising that previous access to MFIs and formal banks are both significant and positive for the clients living in an urban area. Meanwhile, previous access to saving service significant and positive for the clients living in rural area. These two findings are in line with Klapper et al. (2011), who claim that financial literacy is usually acquired via interaction with others instead of education. As the main objective of microfinance is to reduce poverty by providing small loans and savings facilities to the rural poor who are excluded from commercial financial services, in terms of enhancing the clients' consciousness of finance, microcredit is more effective in the urban area, while microsaving is more effective in the rural area. Finally, we also found that access to formal banks might have a significant influence on the financial awareness of interest rate for the clients who are not the household heads. All other interaction terms in this group are insignificant, including the one

with MFIs. In summary, these findings suggest that it is tough to develop greater financial awareness of the microfinance participants by simply providing credit and saving services. Hence, education of basic financial regulations and proper supervision are indispensable.

4.4.4 Further Analysis

An extra test has been carried out to test the robustness of the results and analyse the findings more in detail. Specifically, since sometimes clients only remember the total amount of their interest repayment but not the interest rates, we want to see whether the results with KR_i hold for KA_i as well. **Table 4.6** presents the regressions on both dependents with a new sample which is generated by simple listwise deletion based on the missing data of KA_i. Note that it is normal that we have slightly different results in columns (5)-(8) of **Table 4.6** compared to the results in **Table 4.3** because different samples have been applied. Further discussion about the potential bias caused by sample selection is beyond the purpose of this test. The main objective in this section is to examine whether KR_i and KA_i can be replaced by each other and generate similar results. In fact, the correlation between the two financial awareness proxies is just 0.27, which is much lower than expected. The reason why the clients prefer one proxy over another is unclear.

As illustrated in the table, we see that the results from columns (1)-(4) are inconsistent with the results from columns (5)-(8), especially for our key variables of previous access to various sources of credit. Previous access to moneylenders and MFIs are significant at the 5% level and the 1% respectively by applying KA_i, while previous access to credit are all insignificant by applying and KR_i. One possible explanation for the inconsistency is that when the borrowing amount is small, a client is likely to prefer the amount of interest over the interest rate. As presented in **Table 4.1**, the average loan size of microfinance is just 1,518 USD, and it is heavily skewed to the right (11.66). It may be easier to remember the loan amount approximately in this case. By looking at the results of KA_i along, it seems that the improvement in financial awareness led by previous access to MFIs is noticeable. Hence, someone may argue that both of KR_i and KA_i are biased indicators for financial awareness, and further analysis with a new proxy (which indicates that one of KR_i and KA_i is non-zero) could be conducted. Unfortunately, this is infeasible in practice. 25% difference from the interest rate (KR_i's definition) and 25% difference from the loan size (KA_i's definition) are clearly not comparable, as the interest rate ranges from 13.0% to 48.5% while the loan size range from 11 USD to 136,224 USD according to our data (**Table 4.1**). Therefore, how to merge the

information of KR_i and KA_i and generate a more reliable proxy of financial awareness still waits to be solved.

4.5 Conclusions and limitations

This paper uses a large global data set covering 51 MFIs in 27 countries to test for individual/household effects on the clients' financial awareness of interest rate. This is important, given the documented popular belief that the financial awareness is very low in general, and strengthening the financial awareness through education programmes and supervision would greatly increase the operating cost of MFIs. Hence, a cost-effective screening method for financial awareness is necessary. As far as we know, no rigorous worldwide empirical study has been devoted to this issue. Financial awareness is studied through the proxies designed by Micro Finanza Rating. They are a pair of dummies which indicate whether a client can accurately (less than 25% different from the actual values) remember his/her interest rate and total interest payment. To test our hypotheses, we have applied multiple imputations and mean imputation methods on missing data, and logistic regression on the cross-sectional data. In addition, a test has been carried out to test the robustness of the results.

The descriptive statistics do confirm that the financial literacy of interest rate and total interest payment is very low for microfinance participants in general. Our findings indicate that previous access to moneylenders improved the awareness of interest. Clients who have had saving accounts before were less knowledgeable about the interests. But previous access to saving service has a positive effect on the clients with at least primary education. Previous access to microfinance has positive relation to the financial awareness of the clients who lived in urban areas.

The overall findings regarding the socio-demographic variables suggest that in our sample the association between gender and financial literacy of interest rate only exists in Latin America and Christian countries. Women may be more financially cautious than men in these areas. The results for education background and living location are all significant. They show that a more educated client who lives in the rural area has a much higher probability to be financially cautious. In addition, there are no results for age. Because the missing rate is too high, and the results with multiple imputations and the result with mean imputation are inconsistent.

In terms of the limitations of this study, there is only one single indicator used as the proxy for financial literacy. We only capture the respondents' awareness of their current financial conditions but their financial knowledge, skills, attitudes and etc. Hence, this study does not tell if a respondent was financially literate in every aspect as defined by INFE (2011). Researchers should be careful when trying to use these findings. On the other hand, as a general issue of non-experimental study, any high correlations between indicators and financial literacy may not reflect causal relationships. Hence, we can only confirm there are certain associations between clients' characteristics and their financial awareness. Finally, the biggest limitation of this study is that the two different indicators of financial awareness (KR_i and KA_i) cannot be simply combined at this stage, and the regressions with different dependent variables have generated inconsistent results. How to weight and merge the information of KR_i and KA_i into a single reliable indicator of financial awareness is needed in further studies.

Table 4.1

Descriptive Statistics of the Data before Appling Multiple-Imputation

'Know Interest Rate' is a dummy that is 1 if the client knows his/her interest rate, and 0 otherwise; 'Know Interest Amount' is a dummy that is 1 if the client knows his/her interest amount, and 0 otherwise; 'Gender' is 1 if the client is female, and 0 otherwise; 'Women's Control on Loan' is an ordinal variable that is 2 if the household finance is fully controlled by women, 1 if the household finance is partially controlled by women, and 0 otherwise; 'Education Below Primary School' is 0 if the client has completed primary school, and 1 otherwise; 'Education Below Secondary School' is 0 if the client has completed secondary school, and 1 otherwise; 'Education Below Tertiary School' is 0 if the client has completed university, college or trade school education, and 1 otherwise; 'Living at Rural Area' is 1 if the client is living outside towns and cities, and 0 otherwise; 'Household Size' is the number of family members living in the client's household; 'Client is a House Head' is 1 if the client pay more than half the cost of supporting and housing a qualifying person, and 0 otherwise; 'Employment Status' is 1 if the client is now employed, and 0 otherwise; 'Number of Fixed Income Sources' is the number of payments of a fixed amount on a fixed schedule that received by the client; 'Income per capita' is the average income per person in the household; 'Have Dwellings' is 1 if the client owns a house, flat, or other place of residence, and 0 otherwise; 'Have Land' is 1 if the client owns a piece of land, and 0 otherwise; 'Annual Interest Rate' is the annual rate charged for borrowing from the MFIs; 'Loan Size' is the loan outstanding per client measured in dollars; 'Other Loans from Moneylenders' is 1 if the client is borrowing from moneylenders at the same time, and 0 otherwise; 'Other Loans from MFIs and etc.' is 1 if the client is borrowing from other MFIs at the same time, and 0 otherwise; 'Other Loans from Banks and etc.' is 1 if the client is borrowing from formal banks at the same time, and 0 otherwise; 'Have Saving Account Before' is 1 if the client has opened saving account before, and 0 otherwise; 'Accessed to Moneylenders' is 1 if the client has borrowed from moneylenders before, and 0 otherwise; 'Accessed to MFIs and etc.' is 1 if the client has used any services provided by MFIs before, and 0 otherwise; 'Accessed to Banks and etc.' is 1 if the client has used any services provided by formal banks before, and 0 otherwise;

		N	Missing	Min	Max	Mean	Std. Dev	Skewness		Kurtosis	
Data and Variables		Stat	Stat	Stat	Stat	Stat	Stat	Stat	Std. Err	Stat	Std. Err
Pooled data of multiple imputation (50 iterations)	Know Interest Rate	9053	4%	0	1	0.380	0.485	0.496	0.026	-1.754	0.051
	Know Interest Amount	5845	38%	0	1	0.336	0.472	0.696	0.032	-1.516	0.064
	Gender	9465	0%	0	1	0.594	0.491	-0.383	0.025	-1.854	0.050
	Women's Control on Loan	9043	5%	0	2	0.908	0.869	0.179	0.026	-1.652	0.052
	Age	3801	60%	17	90	39.681	11.258	0.408	0.040	-0.281	0.079
	Education Level	9267	2%	0	3	1.545	0.955	-0.117	0.025	-0.921	0.051
	Education: Below Primary School	9267	2%	0	1	0.164	0.371	1.812	0.025	1.283	0.051
	Education: Below Secondary School	9267	2%	0	1	0.458	0.498	0.167	0.025	-1.973	0.051
	Education: Below Tertiary School	9267	2%	0	1	0.832	0.374	-1.777	0.025	1.159	0.051
	Living at Rural Area	9272	2%	0	1	0.430	0.495	0.283	0.025	-1.920	0.051
	Living at Urban Area	9272	2%	0	1	0.570	0.495	-0.238	0.025	-1.920	0.051
	Household Size	9468	0%	1	128	5.604	4.558	7.136	0.025	101.309	0.050
	Client is a Household Head	9432	0%	0	1	0.575	0.494	-0.302	0.025	-1.909	0.050
	Client is not a Household Head	9432	0%	0	1	0.425	0.494	0.302	0.025	-1.909	0.050
	Employment Status	4889	48%	0	1	0.932	0.252	-3.423	0.035	9.722	0.070
	Number of Fixed Income Sources	9244	2%	0	23	2.082	1.504	2.785	0.025	21.756	0.051
	Income per captita	4409	53%	0	27134	614.457	1280.037	6.841	0.037	81.803	0.074
	Have Dwellings	9275	2%	0	1	0.737	0.440	-1.079	0.025	-0.836	0.051
	Have Land	8904	6%	0	1	0.534	0.499	-0.138	0.026	-1.981	0.052
	Annual Interest Rate	8190	14%	0.13	0.485	0.270	0.115	1.445	0.027	5.130	0.054
	Loan Size	9259	2%	11	136224	1517.793	3611.080	11.655	0.025	280.763	0.051
	Other Loans from Moneylenders	8195	13%	0	1	0.036	0.185	5.011	0.027	23.117	0.054
	Other Loans from MFIs and etc.	8195	13%	0	1	0.081	0.272	3.080	0.027	7.491	0.054
	Other Loans from Banks and etc.	8195	13%	0	1	0.117	0.321	2.391	0.027	3.716	0.054
	Have Saving Account Before	8033	15%	0	1	0.442	0.497	0.232	0.027	-1.947	0.055
	Accessed to Moneylenders	9425	0%	0	1	0.096	0.295	2.743	0.025	5.524	0.050
	Accessed to MFIs and etc.	9425	0%	0	1	0.172	0.377	1.742	0.025	1.034	0.050
	Accessed to Banks and etc.	9425	0%	0	1	0.211	0.408	1.417	0.025	0.007	0.050

Table 4.2

Descriptive Statistics of the Data after Appling Multiple-Imputation

'Know Interest Rate' is a dummy that is 1 if the client knows his/her interest rate, and 0 otherwise; 'Know Interest Amount' is a dummy that is 1 if the client knows his/her interest amount, and 0 otherwise; 'Gender' is 1 if the client is female, and 0 otherwise; 'Women's Control on Loan' is an ordinal variable that is 2 if the household finance is fully controlled by women, 1 if the household finance is partially controlled by women, and 0 otherwise; 'Education Below Primary School' is 0 if the client has completed primary school, and 1 otherwise; 'Education Below Secondary School' is 0 if the client has completed secondary school, and 1 otherwise; 'Education Below Tertiary School' is 0 if the client has completed university, college or trade school education, and 1 otherwise; 'Living at Rural Area' is 1 if the client is living outside towns and cities, and 0 otherwise; 'Household Size' is the number of family members living in the client's household; 'Client is a House Head' is 1 if the client pay more than half the cost of supporting and housing a qualifying person, and 0 otherwise; 'Employment Status' is 1 if the client is now employed, and 0 otherwise; 'Number of Fixed Income Sources' is the number of payments of a fixed amount on a fixed schedule that received by the client; 'Income per capita' is the average income per person in the household; 'Have Dwellings' is 1 if the client owns a house, flat, or other place of residence, and 0 otherwise; 'Have Land' is 1 if the client owns a piece of land, and 0 otherwise; 'Annual Interest Rate' is the annual rate charged for borrowing from the MFIs; 'Loan Size' is the loan outstanding per client measured in dollars; 'Other Loans from Moneylenders' is 1 if the client is borrowing from moneylenders at the same time, and 0 otherwise; 'Other Loans from MFIs and etc.' is 1 if the client is borrowing from other MFIs at the same time, and 0 otherwise; 'Other Loans from Banks and etc.' is 1 if the client is borrowing from formal banks at the same time, and 0 otherwise; 'Have Saving Account Before' is 1 if the client has opened saving account before, and 0 otherwise; 'Accessed to Moneylenders' is 1 if the client has borrowed from moneylenders before, and 0 otherwise; 'Accessed to MFIs and etc.' is 1 if the client has used any services provided by MFIs before, and 0 otherwise; 'Accessed to Banks and etc.' is 1 if the client has used any services provided by formal banks before, and 0 otherwise;

		N	Missing	Min	Max	Mean	Std. Dev	Skewness		Kurtosis	
Data and Variables		Stat	Stat	Stat	Stat	Stat	Stat	Stat	Std. Err	Stat	0.050
Pooled data of multiple imputation (50 iterations)	Know Interest Rate	9053	4%	0	1	0.380	0.485	0.496	0.026	-1.754	0.051
	Know Interest Amount	5845	38%	0	1	0.336	0.472	0.696	0.032	-1.516	0.064
	Gender	9471	0%	0	1	0.594	0.491	-0.382	0.025	-1.854	0.050
	Women's Control on Loan	9471	0%	0	2	0.899	0.869	0.196	0.025	-1.649	0.050
	Age	9471	0%	17	90	40.004	10.910	0.293	0.025	-0.248	0.050
	Education Level	9471	0%	0	3	1.542	0.955	-0.117	0.025	-0.919	0.050
	Education: Below Primary School	9471	0%	0	1	0.165	0.371	1.804	0.025	1.254	0.050
	Education: Below Secondary School	9471	0%	0	1	0.459	0.498	0.165	0.025	-1.973	0.050
	Education: Below Tertiary School	9471	0%	0	1	0.834	0.372	-1.792	0.025	1.213	0.050
	Living at Rural Area	9471	0%	0	1	0.436	0.496	0.259	0.025	-1.933	0.050
	Living at Urban Area	9471	0%	0	1	0.564	0.496	-0.259	0.025	-1.933	0.050
	Household Size	9471	0%	1	128	5.604	4.558	7.135	0.025	101.306	0.050
	Client is a Household Head	9471	0%	0	1	0.575	0.494	-0.302	0.025	-1.909	0.050
	Client is not a Household Head	9471	0%	0	1	0.425	0.494	0.302	0.025	-1.909	0.050
	Employment Status	9471	0%	0	1	0.935	0.247	-3.522	0.025	10.414	0.050
	Number of Fixed Income Sources	9471	0%	0	23	2.090	1.503	2.727	0.025	21.218	0.050
	Income per captita	9471	0%	0	27134	683.441	1159.159	4.386	0.025	53.140	0.050
	Have Dwellings	9471	0%	0	1	0.740	0.439	-1.096	0.025	-0.799	0.050
	Have Land	9471	0%	0	1	0.530	0.499	-0.122	0.025	-1.986	0.050
	Annual Interest Rate	9471	0%	0.13	0.485	0.270	0.114	1.284	0.025	4.502	0.050
	Loan Size	9471	0%	11	136224	1559.451	3598.837	11.500	0.025	277.713	0.050
	Other Loans from Moneylenders	9471	0%	0	1	0.036	0.187	4.965	0.025	22.663	0.050
	Other Loans from MFIs and etc.	9471	0%	0	1	0.080	0.272	3.089	0.025	7.545	0.050
	Other Loans from Banks and etc.	9471	0%	0	1	0.118	0.322	2.375	0.025	3.640	0.050
	Have Saving Account Before	9471	0%	0	1	0.435	0.496	0.262	0.025	-1.932	0.055
	Accessed to Moneylenders	9471	0%	0	1	0.096	0.295	2.740	0.025	5.511	0.050
	Accessed to MFIs and etc.	9471	0%	0	1	0.172	0.377	1.741	0.025	1.031	0.050
	Accessed to Banks and etc.	9471	0%	0	1	0.211	0.408	1.417	0.025	0.007	0.050

Table 4.3

Previous Access to Credits and Financial Awareness of Interest Rate (Part 1)

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Imputation Method	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i
Country Control			Yes	Yes			Yes	Yes			Yes	Yes			Yes	Yes
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent Variable	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR
Gender	0.090*	0.097*	-0.030	-0.025	0.138***	0.143***	-0.016	-0.017								
	[109%]	[110%]	[97%]	[98%]	[115%]	[115%]	[98%]	[98%]								
Women's Control on Loan									0.002	-0.029	-0.005	-0.040	0.018	-0.008	0.000	-0.032
									[100%]	[97%]	[100%]	[96%]	[102%]	[99%]	[100%]	[97%]
Age	0.000	0.002	-0.006**	-0.017***	-0.001	0.001	-0.006*	-0.017***	0.000	0.003	-0.006**	-0.016***	-0.001	0.001	-0.006*	-0.017***
	[100%]	[100%]	[99%]	[98%]	[100%]	[100%]	[99%]	[98%]	[100%]	[100%]	[99%]	[98%]	[100%]	[100%]	[99%]	[98%]
Education Level	0.362***	0.375***	0.237***	0.239***	0.401***	0.415***	0.263***	0.262***	0.356***	0.366***	0.238***	0.239***	0.394***	0.406***	0.264***	0.262***
	[144%]	[145%]	[127%]	[127%]	[149%]	[151%]	[130%]	[130%]	[143%]	[144%]	[127%]	[127%]	[148%]	[150%]	[130%]	[130%]
Living at Rural Area	0.864***	0.878***	0.420***	0.396***	0.731***	0.765***	0.417***	0.384***	0.857***	0.863***	0.421***	0.391***	0.725***	0.754***	0.418***	0.380***
	[237%]	[241%]	[152%]	[149%]	[208%]	[215%]	[152%]	[147%]	[236%]	[237%]	[152%]	[148%]	[206%]	[213%]	[152%]	[146%]
Household Size	0.002	0.005	0.013*	0.014**	-0.010*	-0.005	0.011	0.014*	0.002	0.004	0.013*	0.014**	-0.012**	-0.007	0.012	0.014*
	[100%]	[101%]	[101%]	[101%]	[99%]	[100%]	[101%]	[101%]	[100%]	[100%]	[101%]	[101%]	[99%]	[99%]	[101%]	[101%]
Client is a Household Head	0.093*	0.090*	0.187***	0.187***	0.140***	0.130**	0.182***	0.185***	0.057	0.038	0.197***	0.183***	0.091*	0.069	0.188***	0.181***
	[110%]	[109%]	[121%]	[121%]	[115%]	[114%]	[120%]	[120%]	[106%]	[104%]	[122%]	[120%]	[110%]	[107%]	[121%]	[120%]
Employment Status					0.104	0.283**	0.081	0.268*					0.092	0.270*	0.082	0.268*
					[111%]	[133%]	[108%]	[131%]					[110%]	[131%]	[109%]	[131%]
Number of Fixed Income Sources					0.084***	0.086***	0.008	0.007					0.084***	0.086***	0.007	0.007
					[109%]	[109%]	[101%]	[101%]					[109%]	[109%]	[101%]	[101%]
Income per capita					0.000	0.000	0.000***	0.000***					0.000	0.000	0.000***	0.000***
					[100%]	[100%]	[100%]	[100%]					[100%]	[100%]	[100%]	[100%]
Have Dwellings					0.255***	0.227***	0.071	0.068					0.256***	0.228***	0.070	0.068
					[129%]	[125%]	[107%]	[107%]					[129%]	[126%]	[107%]	[107%]
Have Land					-0.075	-0.094*	0.013	0.006					-0.086*	-0.110**	0.014	0.003
					[93%]	[91%]	[101%]	[101%]					[92%]	[90%]	[101%]	[100%]
Annul Interest Rate					-2.003***	-2.366***	-2.302***	-2.735***					-1.994***	-2.354***	-2.301***	-2.732***
					[13%]	[9%]	[10%]	[6%]					[14%]	[9%]	[10%]	[7%]
Loan Size					0.000***	0.000***	0.000**	0.000***					0.000***	0.000***	0.000**	0.000***
					[100%]	[100%]	[100%]	[100%]					[100%]	[100%]	[100%]	[100%]

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level.*

Table 4.3

Previous Access to Credits and Financial Awareness of Interest Rate (Part 2)

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Imputation Method	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i
Country Control			Yes	Yes			Yes	Yes			Yes	Yes			Yes	Yes
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent Variable	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR
Other Loans from Moneylenders					0.027	0.035	-0.024	-0.012					0.025	0.034	-0.023	-0.012
					[103%]	[104%]	[98%]	[99%]					[103%]	[103%]	[98%]	[99%]
Other Loans from MFIs and etc.					-0.233**	-0.278***	-0.235**	-0.280**					-0.234**	-0.280***	-0.235**	-0.280**
					[79%]	[76%]	[79%]	[76%]					[79%]	[76%]	[79%]	[76%]
Other Loans from Banks and etc.					-0.193**	-0.23***	-0.120	-0.164*					-0.191**	-0.231***	-0.120	-0.164*
					[82%]	[79%]	[89%]	[85%]					[83%]	[79%]	[89%]	[85%]
Have Saving Account Before					-0.444***	-0.531***	-0.172***	-0.193***					-0.439***	-0.524***	-0.172***	-0.192***
					[64%]	[59%]	[84%]	[82%]					[64%]	[59%]	[84%]	[83%]
Accessed to Moneylenders					0.394***	0.402***	0.360***	0.369***					0.398***	0.407***	0.360***	0.369***
					[148%]	[149%]	[143%]	[145%]					[149%]	[150%]	[143%]	[145%]
Accessed to MFIs and etc.					-0.073	-0.070	0.043	0.041					-0.067	-0.062	0.042	0.043
					[93%]	[93%]	[104%]	[104%]					[94%]	[94%]	[104%]	[104%]
Accessed to Banks and etc.					-0.022	-0.020	-0.051	-0.049					-0.024	-0.023	-0.050	-0.050
					[98%]	[98%]	[95%]	[95%]					[98%]	[98%]	[95%]	[95%]
Constant	-0.669**	-0.805***	-1.157***	-0.774**	-0.450	-0.692**	-1.045***	-0.774*	-0.585**	-0.690**	-1.182***	-0.765**	-0.323**	-0.541*	-1.064***	-0.764*
	[51%]	[45%]	[31%]	[46%]	[64%]	[50%]	[35%]	[46%]	[56%]	[50%]	[31%]	[47%]	[72%]	[58%]	[35%]	[47%]
Observations	9053	9053	9053	9053	9053	9053	9053	9053	9053	9053	9053	9053	9053	9053	9053	9053
Nagelkerke R^2	0.078	0.081	0.281	0.282	0.119	0.127	0.298	0.303	0.078	0.081	0.281	0.282	0.119	0.126	0.298	0.303

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level.*

Table 4.4

Previous Access to Credits Regional and Religious Split-Sample Analysis (Part 1)

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Imputation Method	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i
Imputation Iterations	50		50		43		50		50		50		50		50		50	
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sub-Sample	All	All	Africa	Africa	Europe	Europe	Lat.Am.	Lat.Am.	Mid.East	Mid.East	All	All	Christian	Christian	Mixture	Mixture	Islam	Islam
Dependent Variable	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR
Gender	0.230***	0.233***	-0.257*	-0.168	-0.111	-0.096	0.499***	0.493***	0.214	0.213	0.168***	0.175***	0.361***	0.362***	-0.118	-0.121	-0.014	-0.022
	[126%]	[126%]	[77%]	[85%]	[89%]	[91%]	[165%]	[164%]	[124%]	[124%]	[118%]	[119%]	[143%]	[144%]	[89%]	[89%]	[99%]	[98%]
Age	-0.003	-0.005	-0.010	-0.014*	-0.011	-0.028	-0.004	-0.017***	-0.002	-0.007	-0.003	-0.002	-0.004	-0.006	-0.010	-0.017*	-0.003	-0.005
	[100%]	[100%]	[99%]	[99%]	[99%]	[97%]	[100%]	[98%]	[100%]	[99%]	[100%]	[100%]	[100%]	[99%]	[99%]	[98%]	[100%]	[100%]
Education Level	0.337***	0.345***	0.416***	0.392***	-0.044	-0.070	0.192***	0.187***	0.201***	0.233***	0.399***	0.415***	0.261***	0.261***	0.500***	0.497***	0.397***	0.413***
	[140%]	[141%]	[152%]	[148%]	[96%]	[93%]	[121%]	[121%]	[122%]	[126%]	[149%]	[151%]	[130%]	[130%]	[165%]	[164%]	[149%]	[151%]
Living at Rural Area	0.712***	0.717***	0.644***	0.662***	-0.023	0.031	0.511***	0.489***	0.868***	0.880***	0.802***	0.852***	0.695***	0.681***	-0.174	-0.171	1.058***	1.21***
	[204%]	[205%]	[190%]	[194%]	[98%]	[103%]	[167%]	[163%]	[238%]	[241%]	[223%]	[234%]	[200%]	[198%]	[84%]	[84%]	[288%]	[335%]
Household Size	0.009	0.013**	0.023***	0.023***	-0.193***	-0.178**	0.042**	0.041**	-0.059**	-0.065***	-0.022***	-0.016***	-0.003	-0.002	0.023	0.030	-0.023***	-0.008
	[101%]	[101%]	[102%]	[102%]	[82%]	[84%]	[104%]	[104%]	[94%]	[94%]	[98%]	[98%]	[100%]	[100%]	[102%]	[103%]	[98%]	[99%]
Client is a Household Head	0.207***	0.199***	0.302**	0.321**	0.021	0.038	0.046	0.038	0.340**	0.335**	0.155***	0.141***	0.051	0.036	0.278	0.290	0.301**	0.281**
	[123%]	[122%]	[135%]	[138%]	[102%]	[104%]	[105%]	[104%]	[140%]	[140%]	[117%]	[115%]	[105%]	[104%]	[132%]	[134%]	[135%]	[132%]
Employment Status	0.149	0.389***	-0.099	-0.355	-0.221	-0.325	0.080	0.452	-0.008	-0.131	0.014	0.116	0.025	0.305	0.085	0.081	-0.362*	-0.462**
	[116%]	[148%]	[91%]	[70%]	[80%]	[72%]	[108%]	[157%]	[99%]	[88%]	[101%]	[112%]	[103%]	[136%]	[109%]	[108%]	[70%]	[63%]
Number of Fixed Income Sources	0.048***	0.050***	0.020	0.028	0.418***	0.425***	-0.078***	-0.080***	0.079**	0.079**	0.075***	0.080***	0.062***	0.059**	0.078	0.080*	0.036	0.029
	[105%]	[105%]	[102%]	[103%]	[152%]	[153%]	[92%]	[92%]	[108%]	[108%]	[108%]	[108%]	[106%]	[106%]	[108%]	[108%]	[104%]	[103%]
Income per capita	0.000	0.000	0.000	0.000	0.000	0.000	0.000***	0.000***	0.000***	-0.001***	0.000**	0.000	0.000*	0.000**	0.000	0.000	0.000***	-0.001***
	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]
Have Dwellings	0.104*	0.094	0.149	0.114	0.086	0.124	0.051	0.043	0.364**	0.426***	0.266***	0.247***	0.204***	0.188***	0.081	0.075	0.486***	0.466***
	[111%]	[110%]	[116%]	[112%]	[109%]	[113%]	[105%]	[104%]	[144%]	[153%]	[130%]	[128%]	[123%]	[121%]	[108%]	[108%]	[163%]	[159%]
Have Land	-0.012	-0.032	0.106	0.099	-0.272	-0.379	0.006	-0.006	-0.056	-0.222	-0.052	-0.077	0.054	0.044	-0.052	-0.083	-0.305**	-0.491***
	[99%]	[97%]	[111%]	[110%]	[76%]	[68%]	[101%]	[99%]	[95%]	[80%]	[95%]	[93%]	[106%]	[104%]	[95%]	[92%]	[74%]	[61%]
Annul Interest Rate	-2.486***	-2.891***	-4.107***	-8.710***	-6.485**	-5.666**	-2.150***	-2.356***	-1.543**	-1.902***	-2.193***	-2.541***	-2.645***	-2.913***	0.760	1.178	-0.43	-0.824
	[8%]	[6%]	[2%]	[0%]	[0%]	[0%]	[12%]	[9%]	[21%]	[15%]	[11%]	[8%]	[7%]	[5%]	[214%]	[325%]	[65%]	[44%]
Loan Size	0.000***	0.000***	0.000	0.000	0.000	0.000	0.000**	0.000***	0.000***	0.000***	0.000***	0.000***	0.000	0.000**	0.000***	0.000***	0.000***	0.000***
	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]
Other Loans from Moneylenders	0.016	0.042	-0.070	-0.025			-0.112	-0.144	0.462	0.638	0.013	0.02	0.041	0.056	-0.238	-0.246	0.067	0.099
	[102%]	[104%]	[93%]	[98%]			[89%]	[87%]	[159%]	[189%]	[101%]	[102%]	[104%]	[106%]	[79%]	[78%]	[107%]	[110%]
Other Loans from MFIs and etc.	-0.347***	-0.413***	-0.253	-0.333	-0.429	-0.449	-0.194	-0.265**	-0.646***	-0.699***	-0.252***	-0.297***	-0.099	-0.134	-0.409	-0.398	-0.648***	-0.698***
	[71%]	[66%]	[78%]	[72%]	[65%]	[64%]	[82%]	[77%]	[52%]	[50%]	[78%]	[74%]	[91%]	[87%]	[66%]	[67%]	[52%]	[50%]

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level.*

Table 4.4

Previous Access to Credits Regional and Religious Split-Sample Analysis (Part 2)

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Imputation Method	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i
Imputation Iterations	50		50		43		50		50		50		50		50		50	
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sub-Sample	All	All	Africa	Africa	Europe	Europe	Lat.Am.	Lat.Am.	Mid.East	Mid.East	All	All	Christian	Christian	Mixture	Mixture	Islam	Islam
Dependent Variable	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR
Other Loans from Banks and etc.	-0.189**	-0.250***	0.037	0.017	0.188	0.212	-0.181	-0.287**	-0.590***	-0.674***	-0.137	-0.167*	-0.197*	-0.276***	0.394	0.403	-0.347	-0.403*
	[83%]	[78%]	[104%]	[102%]	[121%]	[124%]	[83%]	[75%]	[55%]	[51%]	[87%]	[85%]	[82%]	[76%]	[148%]	[150%]	[71%]	[67%]
Have Saving Account Before	-0.174***	-0.208***	0.087	0.143	0.402*	0.377*	-0.185**	-0.201***	-0.521***	-0.611***	-0.342***	-0.418***	-0.233***	-0.255***	-0.381**	-0.396**	-0.649***	-1.073***
	[84%]	[81%]	[109%]	[115%]	[149%]	[146%]	[83%]	[82%]	[59%]	[54%]	[71%]	[66%]	[79%]	[77%]	[68%]	[67%]	[52%]	[34%]
Accessed to Moneylenders	0.434***	0.437***	0.573***	0.592***	-1.129**	-1.145**	0.418***	0.422***	0.587***	0.567***	0.423***	0.429***	0.487***	0.49***	0.186	0.186	0.306*	0.33**
	[154%]	[155%]	[177%]	[181%]	[32%]	[32%]	[152%]	[153%]	[180%]	[176%]	[153%]	[154%]	[163%]	[163%]	[120%]	[120%]	[136%]	[139%]
Accessed to MFIs and etc.	-0.068	-0.065	0.228	0.221	-0.632*	-0.690*	-0.037	-0.040	-0.223	-0.210	-0.047	-0.046	0.020	0.012	-0.163	-0.171	-0.179	-0.147
	[93%]	[94%]	[126%]	[125%]	[53%]	[50%]	[96%]	[96%]	[80%]	[81%]	[95%]	[96%]	[102%]	[101%]	[85%]	[84%]	[84%]	[86%]
Accessed to Banks and etc.	-0.003	0.003	0.255	0.220	-0.468*	-0.514**	-0.139	-0.125	0.253	0.261	-0.004	0.000	-0.129	-0.125	0.264	0.245	0.064	0.069
	[100%]	[100%]	[129%]	[125%]	[63%]	[60%]	[87%]	[88%]	[129%]	[130%]	[100%]	[100%]	[88%]	[88%]	[130%]	[128%]	[107%]	[107%]
Region: Africa	-1.530***	-1.540***																
	[22%]	[21%]																
Region: Europe	-1.245***	-1.250***																
	[29%]	[29%]																
Region: Latin America	-0.699***	-0.714***																
	[50%]	[49%]																
Region: Middle East	B.M.	0.164																
		[118%]																
Region: South Asia		B.M.																
Religion: Christianity											-0.282***	-0.304***						
											[75%]	[74%]						
Religion: Mixture											-1.201***	-1.182***						
											[30%]	[31%]						
Religion: Islam											B.M.	B.M.						
Constant	1.035***	0.936**	0.097	1.299*	-0.591	-0.129	-1.270***	-0.954**	0.136	0.743	-0.226	-0.412	-0.587**	-0.657*	-3.879***	-3.634***	0.279	0.803
	[282%]	[255%]	[110%]	[367%]	[55%]	[88%]	[28%]	[39%]	[115%]	[210%]	[80%]	[66%]	[56%]	[52%]	[2%]	[3%]	[132%]	[223%]
Observations	9053	9053	1986	1986	682	682	4361	4361	1846	1846	9053	9053	5402	5402	1232	1232	2419	2419
Nagelkerke R^2	0.188	0.192	0.144	0.164	0.367	0.372	0.12	0.123	0.276	0.283	0.149	0.155	0.146	0.149	0.272	0.272	0.263	0.293

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level; B.M. is the benchmark.*

Table 4.5

Interactions on the Relation between Previous Access to Credits and Socio-demographic Characteristics (Part 1)

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8	9	10
Imputation Method	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i
Country Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent Variable	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR
Gender	-0.026	-0.028	-0.016	-0.016	-0.042	-0.043	-0.018	-0.017	-0.019	-0.02
	0.000	[97%]	[98%]	[98%]	[96%]	[96%]	[98%]	[98%]	[98%]	[98%]
Age	-0.006**	-0.017***	-0.007**	-0.017***	-0.008**	-0.019***	-0.006*	-0.017***	-0.006*	-0.017***
	0.638	[98%]	[99%]	[98%]	[99%]	[98%]	[99%]	[98%]	[99%]	[98%]
Education Level							0.266***	0.265***	0.263***	0.262***
							[130%]	[130%]	[130%]	[130%]
Education: Below Primary School	-0.397***	-0.339***								
	[67%]	[71%]								
Education: Below Secondary School			-0.449***	-0.449***						
			[64%]	[64%]						
Education: Below Tertiary School					-0.122	-0.103				
					[89%]	[90%]				
Living at Rural Area	0.385***	0.356***	0.395***	0.363***	0.351***	0.318***				
	[147%]	[143%]	[148%]	[144%]	[142%]	[137%]				
Living at Urban Area							-0.446***	-0.404***	-0.419***	0.387***
							[64%]	[67%]	[66%]	[147%]
Household Size	0.008	0.011	0.010	0.013*	0.008	0.011	0.012*	0.014**	0.012	0.014**
	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]
Client is a Household Head	0.177***	0.179***	0.184***	0.184***	0.171***	0.171***	0.183***	0.186***		
	[119%]	[120%]	[120%]	[120%]	[119%]	[119%]	[120%]	[120%]		
Client is not a Household Head									-0.275***	-0.263***
									[76%]	[77%]
Employment Status	0.101	0.287*	0.093	0.269*	0.126	0.291*	0.080	0.265	0.078	0.268*
	[111%]	[133%]	[110%]	[131%]	[113%]	[134%]	[108%]	[130%]	[108%]	[131%]
Number of Fixed Income Sources	0.012	0.011	0.009	0.009	0.011	0.011	0.006	0.006	0.008	0.008
	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]	[101%]
Income per capita	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]
Have Dwellings	0.068	0.065	0.074	0.069	0.06	0.053	0.062	0.058	0.069	0.066
	[107%]	[107%]	[108%]	[107%]	[106%]	[105%]	[106%]	[106%]	[107%]	[107%]

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level.*

Table 4.5

Interactions on the Relation between Previous Access to Credits and Socio-demographic Characteristics (Part 2)

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8	9	10
Imputation Method	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i
Country Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent Variable	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR
Have Land	0.018	0.012	0.014	0.007	0.007	0.001	0.016	0.008	0.011	0.004
	[102%]	[101%]	[101%]	[101%]	[101%]	[100%]	[102%]	[101%]	[101%]	[100%]
Annul Interest Rate	-2.26***	-2.675***	-2.263***	-2.707***	-2.203***	-2.654***	-2.302***	-2.73***	-2.309***	-2.742***
	[10%]	[7%]	[10%]	[7%]	[11%]	[7%]	[10%]	[7%]	[10%]	[6%]
Loan Size	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000**	0.000***	0.000***	0.000***
	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]
Other Loans from Moneylenders	-0.007	0.005	-0.034	-0.023	-0.037	-0.021	-0.027	-0.013	-0.024	-0.011
	[99%]	[101%]	[97%]	[98%]	[96%]	[98%]	[97%]	[99%]	[98%]	[99%]
Other Loans from MFIs and etc.	-0.245**	-0.292***	-0.231**	-0.275**	-0.245**	-0.294***	-0.239**	-0.280**	-0.234**	-0.280**
	[78%]	[75%]	[79%]	[76%]	[78%]	[75%]	[79%]	[76%]	[79%]	[76%]
Other Loans from Banks and etc.	-0.106	-0.152	-0.114	-0.156	-0.087	-0.133	-0.126	-0.163*	-0.117	-0.160*
	[90%]	[86%]	[89%]	[86%]	[92%]	[88%]	[88%]	[85%]	[89%]	[85%]
Have Saving Account Before	-0.071	-0.075	-0.141*	-0.162**	-0.018	-0.005	-0.034	-0.025	-0.164**	-0.173**
	[93%]	[93%]	[87%]	[85%]	[98%]	[100%]	[97%]	[98%]	[85%]	[84%]
Accessed to Moneylenders	0.420***	0.420***	0.379***	0.368***	0.474*	0.476**	0.316**	0.322**	0.315***	0.322***
	[152%]	[152%]	[146%]	[144%]	[161%]	[161%]	[137%]	[138%]	[137%]	[138%]
Accessed to MFIs and etc.	0.004	-0.001	-0.04	-0.046	0.110	0.102	-0.106	-0.140	-0.030	-0.030
	[100%]	[100%]	[96%]	[96%]	[112%]	[111%]	[90%]	[87%]	[97%]	[97%]
Accessed to Banks and etc.	-0.042	-0.040	-0.074	-0.070	0.046	0.051	-0.269**	-0.269**	-0.174*	-0.171*
	[96%]	[96%]	[93%]	[93%]	[105%]	[105%]	[76%]	[76%]	[84%]	[84%]

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level.*

Table 4.5

Interactions on the Relation between Previous Access to Credits and Socio-demographic Characteristics (Part 3)

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8	9	10
Imputation Method	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i
Country Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent Variable	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR
Below Primary School * Have Saving Account before	-0.479***	-0.650***								
	[62%]	[52%]								
Below Primary School * Accessed to Moneylenders	-0.360*	-0.306								
	[70%]	[74%]								
Below Primary School * Accessed to MFIs and etc.	0.170	0.188								
	[119%]	[121%]								
Below Primary School * Accessed to Banks and etc.	0.259	0.250								
	[130%]	[128%]								
Below Secondary School * Have Saving Account before			-0.064	-0.070						
			[94%]	[93%]						
Below Secondary School * Accessed to Moneylenders			-0.035	0.001						
			[97%]	[100%]						
Below Secondary School * Accessed to MFIs and etc.			0.173	0.181						
			[119%]	[120%]						
Below Secondary School * Accessed to Banks and etc.			0.083	0.070						
			[109%]	[107%]						
Below Tertiary School * Have Saving Account before					-0.148	-0.183				
					[86%]	[83%]				
Below Tertiary School * Accessed to Moneylenders					-0.150	-0.141				
					[86%]	[87%]				
Below Tertiary School * Accessed to MFIs and etc.					-0.088	-0.082				
					[92%]	[92%]				
Below Tertiary School * Accessed to Banks and etc.					-0.098	-0.105				
					[91%]	[90%]				

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level.*

Table 4.5

Interactions on the Relation between Previous Access to Credits and Socio-demographic Characteristics (Part 4)

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8	9	10
Imputation Method	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i	Multi-i	Mean-i
Country Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent Variable	KR	KR	KR	KR	KR	KR	KR	KR	KR	KR
Living at Urban Area * Have Saving Account before							-0.239**	-0.280**		
							[79%]	[76%]		
Living at Urban Area * Accessed to Moneylenders							0.082	0.087		
							[109%]	[109%]		
Living at Urban Area * Accessed to MFIs and etc.							0.277**	0.331**		
							[132%]	[139%]		
Living at Urban Area * Accessed to Banks and etc.							0.371***	0.371***		
							[145%]	[145%]		
Not a Household Head * Have Saving Account before									-0.018	-0.047
									[98%]	[95%]
Not a Household Head * Accessed to Moneylenders									0.096	0.100
									[110%]	[111%]
Not a Household Head * Accessed to MFIs and etc.									0.158	0.154
									[117%]	[117%]
Not a Household Head * Accessed to Banks and etc.									0.319**	0.313**
									[138%]	[137%]
Constant	-0.698*	-0.512	-0.318	-0.041	-0.315	-0.008	-0.617	-0.362	-0.405	-0.561
	[50%]	[60%]	[73%]	[96%]	[73%]	[99%]	[54%]	[70%]	[67%]	[57%]
Observations	9053	9053	9053	9053	9053	9053	9053	9053	9053	9053
Nagelkerke R^2	0.297	0.302	0.297	0.302	0.292	0.297	0.300	0.305	0.299	0.304

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level.*

Table 4.6

Robustness Check with Alternative Proxy of Interest Repayment Awareness

The logistic model was used as the estimation method in this table.

	1	2	3	4	5	6	7	8
Imputation Method	Multi-i	Multi-i	Multi-i	Multi-i	Multi-i	Multi-i	Multi-i	Multi-i
Country Control		Yes		Yes		Yes		Yes
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent:	KA	KA	KA	KA	KR	KR	KR	KR
Gender	0.041	-0.226***	-0.002	-0.235***	0.008	0.014	0.041	0.018
	[104%]	[80%]	[100%]	[79%]	[101%]	[101%]	[104%]	[102%]
Age	-0.005	0.001	-0.001	0.001	0.001	-0.003	0.000	-0.003
	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]	[100%]
Education Level	-0.154***	0.118***	-0.060*	0.140***	0.324***	0.269***	0.379***	0.289***
	[86%]	[113%]	[94%]	[115%]	[138%]	[131%]	[146%]	[134%]
Living at Rural Area	0.581***	0.487***	0.436***	0.456***	0.727***	0.478***	0.613***	0.462***
	[179%]	[163%]	[155%]	[158%]	[207%]	[161%]	[185%]	[159%]
Household Size	0.032***	-0.005	0.018***	-0.011	-0.001	0.018**	-0.006	0.015**
	[103%]	[100%]	[102%]	[99%]	[100%]	[102%]	[99%]	[102%]
Client is a Household Head	-0.110	-0.078	-0.063	-0.048	0.119*	0.209***	0.133*	0.206***
	[90%]	[92%]	[94%]	[95%]	[113%]	[123%]	[114%]	[123%]
Employment Status			-0.323**	-0.024			-0.216	-0.060
			[72%]	[98%]			[81%]	[94%]
Number of Fixed Income Sources			0.047**	0.054**			0.036	0.013
			[105%]	[106%]			[104%]	[101%]
Income per capita			0.000***	0.000			0.000	0.000
			[100%]	[100%]			[100%]	[100%]
Have Dwellings			-0.060	-0.024			0.244***	0.134
			[94%]	[98%]			[128%]	[114%]
Have Land			0.105	0.130*			-0.096	0.066
			[111%]	[114%]			[91%]	[107%]
Annul Interest Rate			-2.372***	-0.813**			-1.951***	-2.227***
			[9%]	[44%]			[14%]	[11%]
Loan Size			0.000***	0.000***			0.000***	0.000**
			[100%]	[100%]			[100%]	[100%]

	1	2	3	4	5	6	7	8
Imputation Method	Multi-i	Multi-i	Multi-i	Multi-i	Multi-i	Multi-i	Multi-i	Multi-i
Country Control		Yes		Yes		Yes		Yes
Survey Year Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent:	KA	KA	KA	KA	KR	KR	KR	KR
Other Loans from Moneylenders			-0.034	-0.140			-0.122	-0.137
			[97%]	[87%]			[89%]	[87%]
Other Loans from MFIs and etc.			-0.150	-0.081			-0.225*	-0.177
			[86%]	[92%]			[80%]	[84%]
Other Loans from Banks and etc.			-0.063	-0.069			-0.265**	-0.251
			[94%]	[93%]			[77%]	[78%]
Have Saving Account Before			-0.082	-0.055			-0.387***	-0.060
			[92%]	[95%]			[68%]	[94%]
Accessed to Moneylenders			0.244**	0.224**			0.242**	0.323
			[128%]	[125%]			[127%]	[138%]
Accessed to MFIs and etc.			0.152*	0.296***			0.001	0.183
			[116%]	[134%]			[100%]	[120%]
Accessed to Banks and etc.			-0.111	0.061			-0.098	-0.062
			[89%]	[106%]			[91%]	[94%]
Constant	-0.753***	-0.314	0.242	-0.319	-1.771***	-2.521***	-1.141***	-2.222***
	[47%]	[73%]	[127%]	[73%]	[17%]	[8%]	[32%]	[11%]
Observations	5650	5650	5650	5650	5650	5650	5650	5650
Nagelkerke R^2	0.056	0.213	0.099	0.222	0.088	0.225	0.124	0.244

*Notes: Odds ratios are shown in the brackets; * denote statistical significance at the 10% level; ** denote statistical significance at the 5% level; *** denote statistical significance at the 1% level.*

Chapter 5:

Multiple Imputation, Maximum Likelihood and Predictive Mean Matching for Semi-continuous Missing Data:

A Study of A Microfinance Administrative Loan Book

5.1 Introduction

Incomplete or missing data exist in almost all areas of empirical research. They are especially common in social and behavioral studies. Many statistical procedures have been developed for analyzing missing data. Two notable ones are and Multiple Imputation (MI), Maximum Likelihood (ML) estimation. Under the assumption of a correctly specified parametric model and that data are Missing at Random, both MI and ML generate consistent parameter estimates and consistent standard errors (e.g., Little and Rubin, 2014; Schafer, 1997).

All these missing data techniques are established based on the assumption that the actual data without missing values are normally distributed. However, such assumption is impractical in most cases. Regarding the administrative loan books of banks, the data of their clients' arrears or the delinquency amounts are usually semi-continuous. Semi-continuous variables consist of a usually fairly large proportion of responses with point masses that are fixed at some value and a continuous distribution among the remaining responses. Variables of this type are often collected in economic applications but can also be found in medical applications. Examples of semi-continuous variables with point masses at zero are income from employment, number of employees, or bacterial counts. Semi-continuous variables differ from censored and truncated variables in that the data represented by the zeros are real and valid, as opposed to the data being proxies for negative values or missing responses (Schafer and Olsen, 1999).

Recent developments indicate that the normal-distribution-based ML can still generate consistent parameter estimates and consistent standard errors even when the population distribution is unknown (Yuan, 2009). Although no analytical results exist for MI to generate consistent parameter estimates when the parametric model is misspecified, it has been stated in the literature that the normal-distribution-based MI generates reasonable parameter estimates and standard errors with distribution violations (e.g., Schafer, 1997:136; Schafer and Graham, 2002; Schafer and Olsen, 1998).

On the other hand, some studies indicated that Predictive Mean Matching (PMM) is better than procedures that assumed normal distributions such as MI and ML in terms of semi-continuous data (e.g., Yu et al., 2007). In fact, PMM has been proposed for a long time (Rubin 1986, Little 1988). However, it becomes available and practical to use only recently. Previously, it could only be utilised in circumstances where a single variable had missing data or when the missing data pattern was monotonic. However, the PMM method is now embedded in many software packages that implement an approach to multiple imputations variously known as Multiple Imputation by Chained Equations (MICE), Sequential Generalized Regression, or the Fully Conditional Specification. It is available in many statistical packages, including SAS, STATA, and R, all of which allow us to use PMM for virtually any missing data pattern. PMM is an attractive method to conduct multiple imputations for missing data, especially for the quantitative variables that are not normally distributed. But it is also easy to do it the wrong way.

5.2 Goals of this research

As data sets in social sciences are seldom normally distributed (Micceri, 1989), it is important to know how MI and ML behave relative to each other under the condition of distribution violations as well. Since both MI and ML are available in various statistical programmes, with typical samples in social sciences coming from populations whose distributions are unknown, further empirical studies with real data will give the needed information for applied researchers to choose a more appropriate MDT.

On the other hand, only a handful of studies have evaluated the performance of PMM, so it is not clear how well it compares with alternative methods such as MI and ML. Little is known

about the practical applicability of PMM on different types of data and how the method compares to other techniques that might be suitable to handle these types of data. Certain characteristics, such as sample size, missing rate, missing mechanism, as well as the types of data, may play a vital role in the performance of PMM.

We will investigate how PMM compares to MI and ML for imputing semi-continuous data, binary data, and ordinal data. We will also investigate how performance is affected by sample size and missing rate in the data, and look into the effects of the missing data mechanism on imputation methods for imputing different types of data. In addition, we will investigate the aforementioned methods in the presence of univariate and multivariate missingness. Finally, we wonder: which is the appropriate method when imputing actual semi-continuous loan book data?

The main contribution of this paper is to provide a systematic evaluation for the imputation performances of MI, ML and PMM methods with actual administrative loan book data, as there are so few performance comparison studies of different missing data techniques (MDT) available in the current literature. The rest of the paper proceeds as follows: Section 3 reviews the literature of missing data mechanisms and the MDT evaluated in this paper. Section 4 describes the data, and how to simulate different types of missingness. Section 5 presents the details of different imputation methods and the criterion for performance evaluation. Section 6 reports the empirical results of all MDT. Section 7 presents the study's conclusions and limitations.

5.3 Literature Review

5.3.1 The Distribution of Missing Data

Missingness is considered to be a probabilistic phenomenon in modern missing-data procedures (Rubin, 1976). For any data sets, we usually define a matrix R of indicators to identify whether a variable is known or missing and refer to R as Missingness. R is a set of random variables having a joint probability distribution which may not be specified by us. It describes the patterns of missing values and to capture roughly possible associations between the missingness and the values of the missing items. As a result, the distribution of R is classified according to the nature of its relationship to the data.

Rubin (1976) developed a widely cited typology for these distributions. A Missing Completely at Random (MCAR) mechanism is present if the missing values of datasets are a random sub-sample of the complete data set. The distribution of R is independent of other variables in the dataset including the target variable. An example of MCAR may be a client who is on vacation during the household survey conducted by a microlender.

Assuming a Missing at Random (MAR) mechanism, the distribution of R depends on the variables of the dataset, but not on the values of the target variable.

In case of Missing Not at Random (MNAR), the distribution of R depends on both observed and unobserved variables. For instance, an MNAR is present if a borrower will not report about his over-indebtedness since he fears the consequences of doing so.

It is usually hard to distinguish between MCAR, MAR and MNAR. Sound knowledge of substantial coherences in the dataset is necessary. The mechanisms offer a mathematical approach to model the distribution of R in association with other variables in the dataset. Nevertheless, it should be pointed out that Rubin's (1976) definitions do not describe a causal relationship between the data and missingness.

By adopting the generic notations, we denote the complete data as Y_{com} and partition it as $Y_{com} = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} indicate the observed and missing variables respectively. Rubin (1976) defined missing to be MAR if the distribution of R does not depend on Y_{mis},

$$P(R|Y_{com}) = P(R|Y_{obs}). \quad (23)$$

In other words, MAR allows the probabilities of missingness to depend on observed data but not on missing data. In fact, MCAR is just a special case of MAR, and it occurs when the distribution does not depend on Y_{obs} either,

$$P(R|Y_{com}) = P(R). \quad (24)$$

When Equation 23 is violated, and the distribution depends on Y_{mis}, the missing data are said to be MNAR. MCAR and MAR are also called ignorable nonresponse, while MNAR is called non-ignorable nonresponse (Allison, 2002). Applying a complete case analysis (CCA) with a potential MNAR dataset might lead to biased or invalid results (Little & Rubin, 2002).

Most misunderstandings of MCAR, MAR, and MNAR arise from common notions about the meaning of random. To a statistician, random suggests a process that is probabilistic rather than deterministic. In that sense, the three mechanisms are all random, because they all posit probability distributions for missingness. However, random may also suggest a process that is unpredictable and extraneous to variables, a notion that agrees more closely with MCAR than with MAR. In this sense, MAR has nothing to do with random at all.

In some research, MAR is known to hold. These include planned missingness in which the missing data were never intended to be collected. Planned missingness values are usually MCAR, but MAR situations sometimes arise. For instance, microfinance participants are included in a follow-up measure only if their applications are approved. The latent variables are missing with probability one and therefore also known to be MAR.

However, in most cases, missingness is beyond the control of researchers. The distribution of R is unknown, and MAR is just an assumption. In general, there is no way to examine whether MAR holds in a dataset. The potential impacts of the departures from MAR-based methods are controversial (Graham et al., 1997). Collins Schafer and Kam (2001) demonstrated that an erroneous assumption of MAR might often have only a minor impact on estimates and standard errors. In contrast, Baraldi and Enders (2010) claimed that the MCAR and MAR based MDT performs poorly with MNAR data. Therefore, the first goal of this paper is to measure the accuracy of different missing data imputation methods under MCAR, MAR and MNAR.

5.3.2 Multiple Imputation and Maximum Likelihood Estimation

Most developments for MI and ML with the different missing mechanisms are based on correctly specified distributions. With complete data, we can use existing procedures to check the distributional properties of the sample before choosing a parametric model (e.g., D'Agostino et al., 1990). With missing data, especially when missing values are MAR, the observed data can be skewed and possess excess kurtosis even when the underlying population is normally distributed. Then most procedures for testing univariate or multivariate normality are not applicable (see e.g., Yuan et al., 2004). Thus, we have to rely on the robust properties of MI or ML in data analysis with missing values.

For the robustness of MI, Graham and Schafer (1999) performed a simulation study by treating a real data set as the population. They found that the absolute values of the biases are small while

most of their population values of the regression parameters are also small. Actually, several biases of their estimates are greater than the population values of the regression parameters. It is not clear whether the small biases are due to the small values of the population parameters. The simulation reported in section 6.4 of Schafer (1997) is also based on a real dataset. But the study does not include an evaluation of the impacts of population skewness and kurtosis on parameter estimates by MI.

Demirtas et al. (2008) conducted a more comprehensive simulation study on MI with two variables, one is complete, and one contains missing values. They found that estimates of variance parameters by MI can suffer from serious bias when the proportion of missing data is large, and the sample size is small, especially when the population is non-normally distributed.

Enders (2001) evaluated biases in ML estimations in the context of distribution violation when missing values are MAR. It is not clear why the bias decrease as the proportion of missing values increases for a population with heavy tails.

None of the above literature compared MI against ML, and none systematically evaluated the influences of different combinations of sample size, missing rate, missing mechanism, data type, and the number of missing variables.

5.3.3 Predictive mean matching

The method PMM is a partially parametric approach, and it predicts the values for the missing data based on a linear prediction model. Firstly, for each missing value, a number of observed values (k) that are closest to the predicted means of the missing values are selected. Secondly, if k=1, the missing observation will be replaced by this observed value; if k>1, an observed value will be randomly selected from the k nearest candidates (Little, 1988). The main attraction of this approach is that the distribution and range of the data are well preserved, and plausible imputed values are guaranteed because only observed values are used here.

Comparing to other standard methods based on linear regression and the normal distribution, the values imputed by PMM are much more realistic. For instance, if the original values of a variable are skewed, the imputed values will be skewed as well; if the original values of a variable are bounded by 1 and 10, the imputed values will be bounded by 1 and 10; and if the real values are discrete (such as the ages of clients), the imputed values will be discrete. This is because all imputed values real values that are “copied” from individuals with real data.

An issue that usually arises when imputing missing values is how to impute variables with non-normal distribution. One common option used in practice to deal with such variables to make the normality assumption more plausible is to conduct a transformation (such as the log or zero-skewness log transformation) to de-skew the data prior to imputation (Lee et al., 2010). A noticeable issue that arises with using de-skewing transformation such as the log transformation for skewed data is that the imputed data may have very large outlying values when the imputed data are transformed back to the original scale. On the other hand, Von Hippel (2013) have compared the methods of rounding, truncating, and transformation when imputing non-normal variables with a lower bound. He suggested that missing data imputations should be conducted on the raw scale with no transformation or post-imputation rounding, even when the data are not normally distributed. However, his focus was restrictive as his study only considered data from an exponential distribution with the lower range restricted.

On the other hand, Yu et al. (2007) investigated general purpose imputation software packages for multiple imputing semi-continuous data. Among the software investigated were routines and packages for SAS [PROC MI, PROC MIANALYZE, and IVEware (Raghunathan, Solenberger and Van Hoewyk, 2002)], R [MICE (Van Buuren and Groothuis-Oudshoorn, 2011)], and STATA [ICE (Royston, 2005)]. They concluded that procedures involving PMM performed very similar to each other and better than the procedures that assumed normal distributions. PMM not only yielded acceptable estimates but also managed to maintain the underlying distributions of the data (Yu et al., 2007).

Although the research by Yu et al. (2007) is useful, it yields only limited insight regarding the reasons why PMM works for semi-continuous data. Yu et al. (2007) focused on readily available software implementations, setting aside methods specifically designed for semi-continuously distributed data (Schafer and Olsen, 1999; Olsen and Schafer, 2001; Javaras and Van Dyk, 2003). Even the procedures implementing PMM had different performances, indicating that a distinction must be made between methods and software implementations.

5.4 Data and Missingness Simulation

In order to compare the performance of the imputations and demonstrate their validity on a real microfinance dataset, we use the 2010 administrative loan book data of Cooperativa de Ahorro y Credito Ceibeña Ltda. (COAC Ceibeña). It is a subset of the dataset which was used in chapter 3 previously. COAC Ceibeña was founded in 1974 on the initiative of Father Donaldo McMillan, and a group of women gathered to the local Catholic Church at La Ceiba Honduras. It is a credit union offering safe and transparent microfinance financial products and services to the local community. The raw data of 2010 COAC Ceibeña has 24 variables and 8,063 cases. The 11 explanation variables (**Table 5.1**) selected in this paper is based on previous studies and expert advice from the microlender staffs, as there is no universally accepted approach to select the explanatory variables for credit scoring (Dinh & Kleimeier, 2007).

5.4.1 Modifying the population

As the size of 8,063 is relatively large for most microfinance institutions in developing countries, we would like to scale down the population to generalize the administrative loan book data. We begin with handling outliers. The occurrence of outliers in our data is very limited, and there are no signs of correlated outliers. Therefore, the simple winsorizing and trimming of Wainer (1976) are adopted here. All observations of Outstanding Balance under $50 or over $10,000 are replaced by the limits. Arrears is trimmed at $2,000. Age is restricted to the range from 20 to 80. Next, we separate the raw data on the level of the point mass and generate two populations. Both populations have size N=3,200, but the populations differ in the size of the point mass: 83.50% and 85.34% point masses at zero for the data with Arrears and the data with Credit Risk respectively. These two populations will be used as the benchmark datasets. It should be noticed that the estimates such as the mean, median, and variances will also change as the size of the point mass changes. The summary statistics of the 11 selected variables with modified population are presented in **Table 5.1**.

5.4.2 Sampling benchmark datasets and skewness preservation

To investigate the performance of MDT under different sample sizes (N), we sample 1,000, 1,700, 2,200, 2,700, and 3,200 cases from the populations. As shown in **Table 5.1**, the variables of interest (y), Arrears, dichotomized Arrears, and Credit Risk, are heavily skewed in practice. In

theory, it may severely impair a method's imputation performance. In order to evaluate the imputation methods at the same distributions across different simulations, we have generated 30 benchmark datasets in this stage (5 sample sizes * 3 variables of interest * 2 models with different number of missing variables). Each table of empirical results in this chapter has 5 sections and each section corresponds to a specific benchmark.

The samples of univariate missing data imputation in this paper are generated following the process as below:

1. We start by separating the data into two sections S_0 and S_+ based on the zeros and non-zeros of y. S_y represents a section in which $y \in [0, +]$
2. Subsample Sub_0 is generated by random sampling a certain percentage (Pct) in S_0.
3. In terms of generation of Sub_+, it depends on the data types of y:
    a. If y is binary, then Sub_+ is generated by random sampling Pct in S_+.
    b. If the y is ordinal categorical (3 levels), we divide S_+ into S_{+_1} and S_{+_2} based on the values of y, then randomly sample Pct in S_{+_1} and S_{+_2}, and finally merge S_{+_1} and S_{+_2} to generate Sub_+.
    c. If y is semi-continuous, we sort all non-zero cases based on the values of y, next, divide S_+ into N sections S_{+_1}, S_{+_2},…, S_{+_n} with equal number of cases, then random sample Pct in each S_{+_k}, and finally merge $S_{+_k}$to generate Sub_+.
4. At last, combing Sub_0 and Sub_+ into a sample ready for missingness simulation.

On the other hand, sample generation for multivariate missing data is similar to that for univariate missing data. The only difference is that the samples should be divided into more subsamples as we need to preserve the skewness of a new continuous variable (y^*), such as Loan Maturity, with missing data as well. The details of the generation process is as follows:

1. We start by separating the data into two groups of sections $S_{0,1}$, $S_{0,2}$,…, $S_{0,m}$ and $S_{+,1}$, $S_{+,2}$,…, $S_{+,m}$ based on the zeros and non-zeros of y and y^*. S_{y,y^*}represents a section in which $y \in [0, +]$, and $y^* \in [1, 2, \dots, m]$.
2. Randomly sampling Pct in each S_{0,y^*}, and then merge all S_{0,y^*} to generate Sub_0.
3. In terms of generation of Sub_+, it depends on the data types of y:
    a. If y is binary, we divide S_+ into $S_{+,1}$, $S_{+,2}$,…, $S_{+,m}$ based on the values of y^*, then randomly sample Pct in each S_{+,y^*}, finally merge all S_{+,y^*} to generate Sub_+.

b. If the y is ordinal categorical (3 levels), we divide S_+ into $S_{+_1,1}$, $S_{+_1,2}$,…, $S_{+_1,m}$, $S_{+_2,1}$, $S_{+_2,2}$,…, $S_{+_2,m}$ based on the values of y and y^*, then random sample Pct in each S_{+_k,y^*}, and finally merge all S_{+_k,y^*} to generate Sub_+.

c. If y is semi-continuous, we divide S_+ into $S_{+,1}$, $S_{+,2}$,…, $S_{+,m}$ based on the values of y^*, then we sort all non-zero cases based on the values of y in each $S_{+,j}$, next, we divide each S_{+,y^*} into N sections S_{+_1,y^*}, S_{+_2,y^*},…, S_{+_n,y^*} with equal number of cases, then we random sample Pct in each S_{+_k,y^*}, and finally merge S_{+_k,y^*} to generate Sub_+.

4. At last, combing Sub_0 and Sub_+ into a sample ready for missingness simulation.

5.4.3 Generating missingness

In terms of the missing data mechanisms, MCAR, MAR, and MNAR are used in our simulations. To investigate the performance of the methods under different missing rates (R), the details of the three mechanisms are adjusted to yield an overall rate of missingness at five different levels (10%, 20%, 30%, 40%, and 50%) in this paper. Regarding to the functions in missing data imputation, all variables can be classified into three types: X, which always observed; Y, which is partly observed; and Z, which may be observed and is a potential cause of missingness for Y. X and Y represent variables that will automatically appear in an imputation because they are of research interest. Z represents variables that is not of direct interest but might be included in the model if the researchers consider it is beneficial.

To model MCAR, missing values are randomly imposed on Y independently of X, Y, and Z at different missing rates stated above. It is straightforward.

To model MAR, the only requirement is that the missingness of Y associates with Z. The potential relations between Y and Z are countless, and it is impossible to model all of them. Common conditions for MAR in the previous literature include: Linear, in which the missingness of Y is linearly related to Z; Quadratic, in which the missingness of Y at the extremes of Z is different from that in the middle; and Sinister, in which the missingness of Y is a function of the correlation between X and Z. The study of Collins et al. (2001) shows that the selection of MAR conditions has little effect on the biases of correlation estimation between X and Y. Therefore, we only focus on linear MAR in this paper for simplicity. In terms of administrative loan books and surveys of MFIs, a typical scenario of MAR would be that males (Z) have higher probability of

nonresponse on the question of Arrears (Y) than females. To simulate such a scenario, we impose a linear MAR missing mechanism on $Gender$ (Z) following a semi-random sampling process designed as follows:

1. We start by separating the initial sample ($S_{complete}$) into four sections $C_{+,f}$, $C_{+,m}$, $C_{0,f}$, and $C_{0,m}$ based on the values of y and z. $C_{y,z}$ represents the sample size of a section in which $y \in [0,+]$, $z \in [female, male]$.

To simulate MAR missingness, we impose different weights ($W_z \in [0,1], z \in [female, male]$) on the missing rates based on $Gender$. The gap between W_f and W_m indicates the strength of association between $Gender$ and missingness of $Arrears$. When $W_f = W_m$, the missing data is MCAR. $W_f \leq W_m$ simulates the scenario that women have lower probability of being missing from the datasets. For simplicity, we setup $W_f = 1$ and $W_m = 0.9$ in this study. In next stages, we impose different missing rates on the four sections as follows:

2. Randomly drop $R_{+,f}$ percent of cases in the section with $C_{+,f}$:
   $R_{+,f} = R * C \,/\, (4 * C_{+,f}) * W_f * 100$
3. Randomly drop $R_{+,m}$ percent of cases in the section with $C_{+,m}$:
   $R_{+,m} = R * C \,/\, (4 * C_{+,m}) * W_m * 100$
4. For the section with $C_{0,f}$, we sort the data based on the values of y and z in ascending order, divide the section into 20 subsections with equal amount of data, randomly drop $R_{0,f}$ percent of cases in each subsection, and then merge all subsection back to $C_{0,f}$:
   $R_{0,f} = R * C \,/\, (4 * C_{0,f}) * W_f * 100$
5. For the section with $C_{0,m}$, we sort the data based on the values of y and z in ascending order, divide the section into 20 subsections with equal amount of data, randomly drop $R_{0,m}$ percent of cases in each subsection, and then merge all subsection back to $C_{0,m}$:
   $R_{0,m} = R * C \,/\, (4 * C_{0,m}) * W_m * 100$

After merging the processed $C_{+,f}$, $C_{+,m}$, $C_{0,f}$, and $C_{0,m}$ back to a single sample ($S_{incomplete}$), we use C^* to denote the sample size of $S_{incomplete}$, and calculate C^* as:

$$C^* = C_{+,f} * (1 - R_{+,f}) + C_{+,m} * (1 - R_{+,m}) + C_{0,f} * (1 - R_{0,f}) + C_{0,m} * (1 - R_{0,m}).$$

When $W_f = 1$ and $W_m = 0.9$, we can infer that $C^* < \mathrm{C} * (1 - R)$. In order to preserve the joint distributions embedded in $S_{incomplete}$ and increase its sample size to $\mathrm{C} * (1 - R)$, we refill the

incomplete sample and generate the final sample with MAR missingness for further imputation evaluations as follows:

6. Randomly select $N_{+,f}$ cases from the abandoned data generated in step 2 and merge them back to $S_{incomplete}$. $N_{+,f}$ is calculated as:
   $N_{+,f} = (C - C^*) * C_{+,f} * \left(1 - R_{+,f}\right) / C^*$
7. Randomly select $N_{+,m}$ cases from the abandoned data generated in step 3 and merge them back to $S_{incomplete}$. $N_{+,m}$ is calculated as:
   $N_{+,m} = (C - C^*) * C_{+,m} * \left(1 - R_{+,m}\right) / C^*$
8. Randomly select $N_{0,f}$ cases from the abandoned data generated in step 4 and merge them back to $S_{incomplete}$. $N_{0,f}$ is calculated as:
   $N_{0,f} = (C - C^*) * C_{0,f} * \left(1 - R_{0,f}\right) / C^*$
9. Randomly select $N_{0,m}$ cases from the abandoned data generated in step 5 and merge them back to $S_{incomplete}$. $N_{0,m}$ is calculated as:
   $N_{0,m} = (C - C^*) * C_{0,m} * \left(1 - R_{0,m}\right) / C^*$

Note that in actuality, the mechanism shown above will be MAR only if Z (e.g., $Gender$) appears in the procedure . If Z is omitted, then the mechanism is actually MNAR and procedures based on an assumption of ignorability may be biased. Again, the potential unobservable variables associated to Y are countless and we cannot model all of them. Therefore, we only consider the most common form of MNAR in this paper. For the microfinance loan books, one example of MNAR would be that clients with non-zero Arrears (Y) have higher probability of nonresponse to a question of Arrears (Y) than clients with zero Arrears. To simulate such scenario, we can simply allow Y (e.g., Arrears in this case) to take the place of Z (e.g., $Gender$) in mechanism MAR above, forcing it to be MNAR. In addition, the generation process of MNAR can refer to the process of MAR.

5.5 Missing Data Imputation Methods

5.5.1 Multiple Imputation

$Y = (Y_{obs}, Y_{mis})$ is an incomplete variable with n sample units, where Y_{obs} and Y_{mis} denote the observed values and the missing values in Y respectively. Besides, let $X = (X_1, \dots, X_j)$ be a set of j fully observed covariates, where X_{obs} and X_{mis} correspond to the observed missing parts in in Y. The number of sample units with observed values of Y and the number of sample units with missing values are denoted by n_{obs} and n_{mis} respectively. Finally, let R be a response indicator. It equals to 1 when Y is observed and 0 when Y is missing. In this study, we consider both univariate and multivariate cases to maintain generality.

The Multiple Imputation (MI) method can be described by a Bayesian approach. In terms of a parametric model for the variable to be imputed, the parameters of the model can be viewed as random variables to which a prior distribution is assigned. In this context, an uninformative prior is used commonly. The information on the parameters is then updated by taking the observed data into account. It leads to the posterior predictive distribution for the parameter vector. To obtain the imputations for the missing values, we can draw a value from the posterior distribution of the parameter vector, and then draw a value from the distribution of the missing data given the drawn value of the parameter vector and the observed data. When this procedure is repeated for m times, m imputations will be obtained for each missing value that are draws from the posterior distribution of missing data.

According to Gelman (2007), MI creates several imputed values for each missing value from similar but different methods, and each method spits out a complete dataset. Using these datasets, we can draw a combined inference across all datasets. Gelman also gives us an example of details in his book. If we use regression and we want to make inference about the coefficient $\hat{\beta}$, we can simply take the average across all datasets $\hat{\beta}_m$. The mean and variance can be expressed as:

$$\hat{\beta} = \frac{1}{m}\sum_{m=1}^{M}\hat{\beta}_m$$

$$V_\beta = \frac{1}{m}\sum_{m=1}^{M} s^2{}_m + \left(1 + \frac{1}{m}\right)\frac{1}{m-1}\left(\hat{\beta}_m - \hat{\beta}\right)^2 \quad (25)$$

where m is number of methods we are using and β_m and s_m as the estimates from every individual method.

The missing data imputation methods introduced in the remainder of this subsection make use of two parametric models: the linear regression model and the logistic regression model. The linear regression model for a target variable Y can be expressed as:

$$Y_i = X_i^T \beta + \epsilon_i \quad (26)$$

where X_i is the vector of values from the *j* covariates for unit *i*, β is the corresponding regression coefficient vector, and ε_i is a normally distributed random error term with expectation zero and variance σ^2. The parameter estimates $\hat{\beta}$, $\hat{\epsilon}$, and $\hat{\sigma}^2$ in this model can be obtained by OLS using the units for which both *Y* and *X* are observed. Using uninformative priors for β and σ^2, the posterior distribution for β will be $N\left(\hat{\beta}, V\left(\hat{\beta}\right)\right)$. It is normal with mean $\hat{\beta}$ and covariance matrix $\mathrm{V}\left(\hat{\beta}\right) = \sigma^2\left(X_{obs}^T X_{obs}\right)^{-1}$. On the other hand, the posterior distribution for σ^2 is given by $\hat{\epsilon}^T\,\hat{\epsilon}/A$, where is *A* is a chi-square variate with $(n_{obs} - r)$ degrees of freedom. A draw from the posterior predictive distribution for a missing value for unit *i* can be obtained by drawing values σ^{2*} and β^* from their posterior distributions, and then drawing a value for $Y_{mis,i}$ from the distribution $N\left(X_i^T \beta^*, \sigma^{2*}\right)$.

The logistic regression model for a binary (0,1) target variable *W* can be written as

$$log\,\frac{\pi_i}{1-\pi_i} = X_i^T \gamma \quad (27)$$

where *γ* is the corresponding regression coefficient vector, and π_i is the probability of observing $W_i = 1$, or equivalently, $\pi_i = E[W_i]$. An expression for π_i in terms of the linear predictor $X_i^T\gamma$ is obtained from the inverse log transformation: $\pi_i = expit\left(X_i^T\gamma\right) = \exp\left(X_i^T\gamma\right)/\left[1 + exp\left(X_i^T\gamma\right)\right]$. Using an uninformative prior for *γ*, the corresponding posterior distribution is approximately $\mathrm{N}\left(\hat{\gamma}, \hat{V}(\hat{\gamma})\right)$ with $\hat{\gamma}$ the maximum likelihood estimator for *γ* and $\hat{V}(\hat{\gamma})$ the associated covariance matrix. A draw from the posterior predictive distribution of a missing value $W_{mis,i}$ can be obtained by drawing a value ***γ**** from the posterior distribution for *γ* and then drawing a value W_i^* from a Bernoulli distribution with parameter $\pi^* = expit\left(X_i^T\gamma^*\right)$.

5.5.2 Predictive Mean Matching

The algorithm of multiple imputing Y_{mis} by means with Predictive Mean Matching (PMM) approach can be expressed as follows:

1. Use linear regression of Y_{obs} given X_{obs} to estimate $\hat{\beta}$, $\hat{\sigma}$, and $\hat{\varepsilon}$ by means with OLS.
2. Draw $\sigma^{2*} = \hat{\varepsilon}^T \hat{\varepsilon}/A$, where *A* is a X^2 variate matrix with $(n_{obs} - r)$ degrees of freedom.
3. Draw β^* from a multivariate normal distribution centered at estimate $\hat{\beta}$ with covariance matrix $\sigma^{2*}\left(X_{obs}^T X_{obs}\right)^{-1}$.
4. Generate $\hat{Y}_{obs} = X_{obs}\hat{\beta}$ and $\hat{Y}_{mis} = X_{mis}\beta^*$.
5. Find $\Delta = \left|\hat{Y}_{obs} - \hat{Y}_{mis,i}\right|$ for each $\hat{Y}_{mis,i}$.
6. Sample values from ($\Delta^{(1)}$, $\Delta^{(2)}$, $\Delta^{(3)}$) randomly and take the corresponding $Y_{obs,i}$ as the imputation, where $\Delta^{(1)}$, $\Delta^{(2)}$ and $\Delta^{(3)}$ denote the three smallest Δ respectively.
7. Repeat steps 1 to 6 *m* times, and save the completed dataset in each repetition.

The default setup of the function 'mi impute pmm' in STATA conducts multiple imputations (m = 20 times) according to the description of the algorithm presented above. The function 'mi.pmm' in R also performs PMM imputation. But it calculates $\Delta = \min\left|\hat{Y}_{obs} - \hat{Y}_{mis,i}\right|$ and selects the corresponding $Y_{obs,i}$ as the imputation.

5.5.3 Maximum Likelihood Estimation

The principle of drawing inferences from a likelihood function has been widely accepted. By assuming that the data is MAR and the model for the complete data is realistic, the marginal distribution of the observed data can provide the correct likelihood for the unknown parameters. Hence, Little and Rubin (1987) referred to this function as the likelihood ignoring the missing-data mechanism. The logarithm of this function is presented as follows:

$$l(\theta; Y_{obs}) = \log L(\theta; Y_{obs}) \quad (28)$$

where θ and Y_{obs} indicate the unknown parameters and the observed data respectively. The Maximum Likelihood (ML) estimate of θ tends to be approximately unbiased when the sample size is large enough. The efficiency of imputation is also positively associated to the sample size,

and its variance approaches the theoretical lower boundary of what is achievable by all unbiased estimators.

Confidence intervals and regions are usually calculated by appealing to the fact that, in regular circumstances with large samples, the ML estimate of θ is approximately normally distributed about the true parameter θ with the approximate covariance matrix:

$$V(\hat{\theta}) \approx \left[-l''(\hat{\theta})\right]^{-1} \quad (29)$$

where $l''(\hat{\theta})$ is the matrix of second partial derivatives of Equation 28 with respect to the elements of θ. The matrix $\left[-l''(\hat{\theta})\right]$ is often called observed information. It describes how fast the log-likelihood function drops as we move away from the ML estimate. A steep decline indicates that the ML estimate is apparently precise, and a gradual decline implies there is considerable uncertainty about the actual location of the true parameter. Sometimes, this matrix is replaced by its expected value, which is call expected information, because the expected value is easier to compute in some cases. In complete-data problems, the approximation in Equation 29 is valid if we replace the observed information with the expected information. Nevertheless, Kenward and Molenberghs (1998) have pointed out that it is not necessarily true with missing data. Expected information implicitly uses Equation 28 as a sampling distribution for Y_{obs}, which is valid only when the data is MCAR. Therefore, for the missing-data problems in this chapter, we obtain standard errors and confidence intervals from the observed data under the MAR condition, and we obtain standard errors and confidence intervals from the expected information matrix under the MCAR condition.

Except for some rare cases, expressions for ML estimates should not be written down in closed form in general. Computing ML estimates require iteration. In an influential article on the Expectation Maximisation (EM) algorithm, Dempster et al. (1977) have described a general method to compute ML in missing-data problems. The EM algorithm is commonly used in data clustering and machine learning. The key idea of this algorithm is to solve a difficult incomplete-data estimation by iteratively solving an easier complete-data estimation. The EM algorithm consists of two steps: Expectation (E-step) and Maximisation (M-step). The E-step calculates the conditional expectation of the parameter on missing data, which are the objective values we are trying to

impute given every iteration on the corresponding parameters. The M-step estimates the parameters by maximizing the likelihood on complete data. Hence, the estimated mean of the EM algorithm can be expressed as:

$$Q\left(\theta\middle|\theta^{(t)}\right) = \int l(\theta|y)\, f\left(Y_{mis}\middle|Y_{obs},\ \theta = \theta^{(t)}\right) dY_{mis} \quad (30)$$

where $l(\theta|y)$ indicates the complete-data log-likelihood. The goal of the M-step is:

$$Q\left(\theta^{(t+1)}\middle|\theta^{(t)}\right) \geq\ Q\left(\theta\middle|\theta^{(t)}\right) \quad (31)$$

until this chain reaches a point where the estimate of θ converges.

In most cases, the data is generally assumed to be normally distributed as it is the most common distribution in nature. With this assumption, the EM algorithm can be formulated into a more concrete form. The two parameters that determine the difference between normal distributions are the mean μ and the variance σ^2. We assume the data is sorted. Observations 1 through r are observed, and observations r through n are missing for notation and explanation. The E-step estimates the values of the parameters as follows:

$$E\left(y_i\middle|\theta^{(t)}, Y_{obs}\right) = \sum_{i=1}^{r} y_i + (n-r)\mu^{(t)}$$

$$E\left({y_i}^2\middle|\theta^{(t)}, Y_{obs}\right) = E({y_{obs}}^2 + {y_{mis}}^2) = \sum_{i=1}^{r} {Y_i}^2 + E(n-r)(\mu^t)^2 \quad (32)$$

The estimates described above are called sufficient statistics. We could use these two values to sample the distribution. In this case, the first equation gives us the estimate of $(t+1)$ th step estimate μ, and the second equation gives us the expected square sum of all observations.

The M-step uses the same sufficient statistics using the estimated parameters from the E-step. The t-th step estimates of the mean μ and the variance σ^2 can be presented as follows:

$$\mu^{(t+1)} = \frac{E\left(\sum_{i=1}^{n} Y_i|\theta^{(t)}, Y_{obs}\right)}{n} = \frac{E\left(\sum_{i=1}^{r} Y_i + \sum_{obs=1}^{n} Y_{obs}\right)}{n}$$

$$\left(\sigma^{(t+1)}\right)^2 = \frac{E\left(\sum_{i=1}^{n} {Y_i}^2|\theta^{(t)}, Y_{obs}\right)}{n} - \left(\mu^{(t+1)}\right)^2 \quad (33)$$

The interpretation of $\mu^{(t+1)}$ uses the observed values and estimated parameters from the E-step to calculate the expected values regarding the data as there are no missing values in it. As

we know $\sigma^2 = E(X^2) - \big(E(X)\big)^2$, with the t-th step estimates, the expectation $E(X^2)$ is known so that the variance σ^2 can be computed.

5.5.4 Completed Case Analysis

Finally, we consider the Completed Case Analysis (CCA) method and compare it to the other imputation methods for its accuracy in returning the parameters of linear regressions after missing values have been removed. CCA is the oldest and most popular solution to deal with missing data. It simply discards units whose information is incomplete instead of imputing the missing values. One of the most widely known approaches based on CCA is the Listwise Deletion (LD). It is used by default in many statistical programmes. But details of its implementation may vary. CCA confines attention to units that have observed values for all variables under consideration. For instance, suppose we compute a sample covariance matrix for items $X_1, \dots, X_m$, CCA will omits from consideration any case that has a missing value on any of the variables $X_1, \dots, X_m$.

CCA can be motivated as a sampling distribution for observables and is generally valid only under MCAR condition. In a few circumstances, it produces inferences that are optimal when the data are MAR. For instance, under a univariate missingness pattern[3], the parameters of the regression on any subset of $X_1, \dots, X_m$ can be estimated from the complete cases and the estimates are both valid and efficient under MAR condition (e.g., Graham & Donaldson, 1993). Nevertheless, this result is not applicable to other measures of association between a dependent and independents, such as correlation coefficients, and it is not applicable to parameters of the marginal distribution of the independent either. The results from CCA might be biased when the missing mechanism is not MCAR. Because the complete cases can be unrepresentative of the full population. The impact of this bias might be ignorable if the departure from MCAR is not serious. However, it is hard to judge how large the bias might be in practice.

In fact, CCA can still be inefficient even when the assumption of MCAR holds. Suppose that a dependent variable is highly related to a dependent variable and the correlation coefficient is close to 1, the missing values of the dependent should be predicted with near certainty. CCA bases es-

[3] The form of the missingness depends on the complexity of the nonresponse patterns. The missingness can be: (a) univariate pattern; (b) monotone pattern, or (c) arbitrary pattern. In this paper, we focus on the univariate pattern only to simplify our estimations.

timates on the reduced sample of the dependent variable and ignores useful predictive information contained in the independents. However, simplicity is still the main advantage of CCA. If a missing data issue can be resolved by dropping only a tiny proportion of data, then CCA can be quite effective and worth to be considered. Therefore, CCA is mentioned as a benchmark for the performance evaluations of missing data imputations in this chapter.

5.5.5 Evaluation of Imputation Performance

We directly compare imputed values to the true values of the missing data points, and we run a linear model on the final imputed data that we compare to the same model run on the true data. Note that the parameters of the linear model are distinct and separately estimated from the parameters of the PMM, MI, and ML models. We use several measures to assess the accuracy of the PMM, MI and ML algorithms. In terms of MSE and BIAS, lower values indicate higher quality. Before imputations, we run a linear OLS regression on the benchmark datasets. Each imputation algorithm outputs 25 imputed datasets, and we combine estimates from the same model fit to each of these imputed datasets using Rubin's rules (Rubin, 1987). In this paper, we consider the following measures of model accuracy:

1. A useful measure of overall accuracy is the mean-square error (MSE), the average squared difference between the estimate and its target. This measure of accuracy combines the concepts of bias and efficiency because it can be shown that the MSE of an estimate is equal to its squared bias plus its variance. For easier interpretation, we report the square root of the MSE, to put it on the same scale as the parameter.
2. In our simulations, we also report the actual coverage of nominal 95% intervals. The actual coverage translates directly to an actual Type I error rate. If the coverage of a nominal 95% interval is actually 90%, then the actual Type I error rate for a testing procedure with a 0.05 level criterion is twice as high as it ought to be. We regard the performance of the interval procedure to be troublesome if its coverage drops below 90%. The 95% confidence interval coverage probability (CP) is calculated as:

   $\mathrm{CP} = \frac{1}{R}\sum_{r=1}^{R} CI_{95}\left(\hat{\theta}_{pr}\right).$ (34)

   If the parameter p falls in the 95% confidence interval at the r times of iteration, then

   $CI_{95}\left(\hat{\theta}_{pr}\right) = 1$; otherwise, $CI_{95}\left(\hat{\theta}_{pr}\right) = 0$ (35)

3. Allowing for comparability with previous research, bias was expressed as a percentage of sample bias relative to the true parameter value. This is calculated as:

$$\%\text{BIAS} = \left[\left(\hat{\theta}_{ij} - \theta_i\right) / \theta_i\right] * 100, \quad (36)$$

where θ_i is the true population parameter for the i th element of θ in each benchmark dataset, and $\hat{\theta}_{ij}$ is the corresponding parameter estimate taken from the j th iteration. The mean percentage bias was subsequently calculated across the 1,000 replications within each cell.

5.6 Empirical Results

5.6.1 Semi-continuous variable in univariate missing data

The purpose of the first study is to investigate the effects on results of omitting a semi-continuous variable on the standardized bias, RMSE, and coverage rate. Previous studies claim that PMM preserves data distribution and imputes only non-negative values when the data consist of non-negative values. In contrast, the log-transformation procedure of MI-LOGIT may lead to imputing non-negative values that are far outside the range of observed values and change the distribution. Therefore, it is expected that PMM tends to outperform MI-LOGIT and other methods in the condition of using semi-continuous data.

The evaluations for the estimated coefficients of the variable with missing data can be found in **Table 5.2**. In general, most of the biases are high and exceed the significant criterion of 40 (shaded areas). All MDT perform the worst under the assumption of MAR instead of MNAR. We also found that the biases of all techniques seem to positively associate with the missing rate, while the relations between them are not strictly linear. Another interesting finding is that the sensitivity of biases to the missing rate is affected by sample size. It seems that the biases are less sensitive to changing missing rates when the sample size is smaller.

When the model complexity is low, the sample size is very small, and collecting a bit more data may dramatically reduce the generalization error (GE) in the perspective of machine learning (which is indicated by BIAS in this chapter), we are likely to overfit the data. GE is the sum of MSE and variance, which associate with sample size and missing rate respectively. The relation-

ship between variance and GE is convex, and the sensitivity between them reaches the maximum level at the lowest point of the GE curve. The finding of the sensitivity between biases and changing missing rates stated above can be explained by the tradeoff between MSE and variance.

For most of the conditions with a significant bias, there is a correspondingly greater disruption in coverage. Hence, the coverage probability is also higher in the conditions where the sample size is large, and the missing rate is low. In most of the imputations, the coverage rates are very low and under 90% (**Table 5.2**).

All these results indicate that data quantity is in great demand in a univariate imputation of semi-continuous data with a high portion of zeros. When conducting imputation for semi-continuous data with less than 3,200 cases, it is recommended that the percentage of missing values should be no greater than 10%. The demand for data quantity for imputation would be slightly lower under the assumption of MCAR.

In **Table 5.2**, we can see that PMM has lower biases and higher coverage rates than MI-LOGIT and ML when the sample size is large enough. Besides, the RMSEs of PMM are lower than those of MI-LOGIT and ML when the missing rates are lower than a certain percentage.

We also see that MI-LOGIT has lower biases and RMSEs than ML, while their coverage rates are approximately the same. In fact, Collins et al. (2001) have indicated that a likelihood-based analysis (ML) and a Bayesian analysis (MI) produce very similar results when the sample size is large enough. With 16 variables in our regression function, the estimation of an unstructured 16 *16 covariance matrix should be relatively stable with more than 1,200 cases. This is why the performance differences between MI-LOGIT and ML showed here are so small.

On the other hand, we also notice that CCA outperforms all other MDT in many cases, especially when the missing rates are high (≥40%). These findings remind us that modern missing data imputation techniques are not always the best. Sometimes simpler is better.

As illustrated above, the advantage of PMM is to preserve the distributional shapes of the variables even for the most extremely skewed semi-continuous ones. Its main drawback is the information lost in the right tail of the distributions due to sampling. In contrast, MI-LOGIT and ML will preserve the continuous part of a semi-continuous variable which clearly shows from the

plots. Regarding semi-continuous data, point mass is the most influential factor affecting its distribution. The size of the point mass mentioned in prior studies related to semi-continuous data imputation is mostly around 20% to 60% (e.g., Yu et al., 2007). In practice, the point mass of administrative loan book data that provided by financial institutions is much higher. In this paper, the proportion of zeros in the raw data used in this paper is very high (83.5%). Therefore, the information lost in right tail caused by PMM may require a greater quantity of data to compensate. It might be the reason why CCA is better than PMM in most conditions here.

What is more, it is also found that the break-even points for performances between PMM and CCA shift downwards as we change the missing mechanism from MCAR to MAR and MNAR. For instance, by comparing the RMSEs of PMM and CCA with 10% missing data, we can see that PMM outperforms CCA when the sample size is 3,200 with MCAR pattern (**Table 5.2 Panel 1**), 2,700 with MAR pattern (**Panel 2**), or 1,700 with MNAR pattern (**Panel 3**).

One possible explanation of these results is that the point masses were slightly reduced to lower levels during the simulation of MAR and MNAR mechanisms. For instance, in this paper, MNAR on the variable of Arrears is designed to simulate a common situation that clients with delinquency have a higher probability of missing from reporting. As a result, more information is preserved by PMM, and it leads to better imputations. Nevertheless, we should notice that there are some other MAR and MNAR simulation methods which have no impacts on the distribution of semi-continuous data. In these cases, the performance of PMM might be consistent across different missingness mechanisms.

5.6.2 Semi-continuous variable in multivariate missing data

The second part of the first study is to investigate the effects on results of omitting two variables simultaneously. Previous simulation studies (e.g., Enders, 2001) show that when there are a lot of missing variables, CCA will have inappropriate standard errors and biased parameters. Hence, we should expect that the performance of CCA decreases dramatically as the new missing simulation is introduced to another variable (Loan Maturity). The values of Loan Maturity are positive and continuous. It is heavily skewed to the right as well. In order to minimize the impacts brought by missing mechanisms, the missingness of Loan Maturity is assumed to be MCAR, and that of Arrears is assumed to be MNAR. We would like to examine if the findings of PMM and

other imputation methods presented previously are consistent with multivariate missing data as well.

Table 5.3 presents evaluations for the estimated coefficients of the same variable as in the previous subsection. By comparing the three panels in **Table 5.3** to those in **Table 5.2**, we see that the biases of multivariate missing imputations are generally greater than the biases of univariate missing imputations. Most of the biases have exceeded the significant criterion of 40. However, with the MNAR missing mechanism, PMM, MI-LOGIT, and ML perform very well as the missing rate is low (≤20%), no matter what size the sample is.

Besides, the biases of imputation are much more sensitive to the changing missing rate with multivariate missing data. It is especially obvious when the missing mechanism is MAR. In **Table 5.3 Panel 2**, we also notice that biases dramatically arise as the percentages of missing values increased from 20% to 30%. It indicates that there might be a concave relation between missing rate and imputation bias.

In terms of RMSEs and biases, we found that PMM performs better than MI-LOGIT and ML when the missing mechanism is MAR (**Table 5.3 Panel 2**). This result is consistent with the findings in the prior literature (e.g., Vink et al., 2014). If we change the missingness to MCAR (**Table 5.3 Panel 1**), PMM is found to be more accurate in most of the simulations, when the sample sizes are greater than 1,200, and the missing rates are lower than 30%. If we change the missingness to MNAR (**Table 5.3 Panel 3**), MI-LOGIT and ML are still underperformed PMM as the sample sizes are large (≥2,700) and the missing rates are low (≤20%). What is more, we notice that PMM has much higher coverage rates across different missing mechanisms in **Table 5.3**. All these results indicate that PMM is still preferable under certain conditions when the missing mechanism is not MAR.

As we evaluate the performance of all four MDT in the MCAR data (**Table 5.3 Panel 1**), it is clear that the RMSEs and biases of CCA are much higher than those of PMM, MI-LOGIT and ML. The coverage rates of CCA are the lowest as well. However, it is surprising that CCA consistently outperforms other MDT when the missing rate is 10%, and the missing mechanism is MAR or MNAR (**Table 5.3 Panel 2** and **Panel 3**). It implies that CCA might still be a preferable method to deal with multivariate missing data when the proportion of cases lost to missingness is small.

5.6.3 Binary variable in univariate missing data

The second study aims to investigate the effects on the results of omitting a binary variable on the standardized bias, RMSE, and the coverage rate. In order to preserve the point mass, the binary variable for imputation is acquired by dichotomizing the semi-continuous form of the same variable used in former subsections. In practice, it is unusual for the formal banks to dichotomize their administrative loan book even when the point mass of Arrears is very high (≥95%). If the quantity of their data is enormous, then MDT such as PMM is still applicable. However, the loan books for most of the microfinance institutions contain fewer than 10,000 cases (e.g., CACIL HONDURAS, 2010, 4,171 clients; INSOTEC ECUADOR, 2010, 7,993 clients). Considering the high proportion of zeros in loan books, sometimes it is worth gaining extra precision for missing data imputation at the price of information lost caused by dichotomization.

Theoretically, PMM derives a distribution for each variable and then draws imputed categories by matching the conditional mean of each missing value to the observed values. Previous studies indicated that PMM avoids using rounding and probabilities to draw categories from continuous imputation, but it does not avoid treating categorical variables as continuous (Kropko et al., 2014). This may lead to inaccurate imputed values, and matching may be inaccurate as well. Therefore, in this subsection, we would expect that the MI-LOGIT and ML methods may outperform PMM in the condition with binary data.

The evaluations for the estimated coefficients of the binary variable with missing data are presented in **Table 5.4**. As can be seen, the coverage rates and biases of all four methods dramatically improved after the semi-continuous missing data is dichotomized to binary format. All biases are under the significant criterion of 40, and most of the coverage rates are higher than 90% now. In addition, the imputations with MCAR missing mechanism have the highest coverage rates and the lowest biases, followed by those with MAR. All these results indicate that the four MDT are robust and yields accurate inference across most simulations in this case.

In **Table 5.4**, we can see that MI-LOGIT and ML perform better than PMM in general as expected, especially with MCAR and MNAR missing mechanisms. PMM has lower imputation biases than MI-LOGIT and ML under the assumption of MAR only when the missing rates are low

(≤20%). By comparing the results presented in **Table 5.2** and **Table 5.4**, it is obvious that transforming a continuous variable into binary format actually impair the performance advantage of PMM against MI-LOGIT and ML.

When the missing mechanism is MCAR or MAR, MI-LOGIT has marginally smaller biases comparing to ML. This is consistent with the results of the previous test with semi-continuous data (**Table 5.2**). Nevertheless, the biases of ML are found to be slightly smaller than those of MI-LOGIT with MNAR. In terms of RMSEs and coverage rates, MI-LOGIT still performs better than ML in general. What is more, we also found that the gaps between the performances of MI-LOGIT and ML are narrower in MCAR. A possible explanation is that the point mass of the MCAR missing data (**Table 5.4 Panel 1**) is higher than those of MAR and MNAR data (**Table 5.4 Panel 1** and **Panel 2**).

At last, the results also show that PMM, MI-LOGIT and ML have lower biases than those of the benchmark method CCA when the missing data is MCAR with small sample size (1,200) or when the missingness is MAR. In addition, the coverage rates of CCA are lower than those of the other MDT across MCAR, MAR and MNAR.

5.6.4 Binary variable in multivariate missing data

The second part of the second study aims to investigate the impacts on the results of omitting two variables at the same time. Once again, we will simulate missingness on the continuous variable Loan Maturity and the dichotomized form of Arrears. The missing mechanisms of Loan Maturity and Arrears will be MCAR and MNAR respectively. Based on the results of the semi-continuous data in subsection 2, it is reasonable to infer that, in general, PMM, MI-LOGIT and ML will outperform CCA in this subsection as well. We would like to examine whether MI-LOGIT and ML will be more accurate than PMM after a heavily skewed variable with missing values is introduced into the imputation equation.

Table 5.5 shows that not all MDT are robust across the simulations with multivariate missing data. When the missing mechanism is MAR or MNAR, and the missing rates are high (≥40%), the biases of CCA and the coverage rates of all MDT still exceed the significant criterions. By comparing the results presented in **Table 5.3** and **Table 5.5**, we can see that the performances of all MDT have been improved. Hence, transformations of multiple semi-continuous variables are highly recommended when their missing rates are higher than 10%, and the mechanism is MAR.

The advantages of MI-LOGIT and ML against PMM have been noticeably weakened as a heavily skewed variable with missing data is introduced to the datasets. While MI-LOGIT and ML still outperform PMM when the missing mechanism is MCAR, the performance of PMM is consistently better than those of MI-LOGIT and ML in every aspect in the missing data with MAR (**Table 5.5 Panel 2**). In addition, PMM has lower biases and coverage rates when the sample size equals to 3,200, and the missing rates are smaller than 40%.

The results between MI-LOGIT and ML are a bit different from those in the previous test with univariate missing data (**Table 5.4**). When the missing mechanism is MCAR, ML performs slightly better than MI-LOGIT in terms of coverage rate and bias when the missing rates are very small (**Table 5.5 Panel 1**). On the hand, we also found that MI-LOGIT has smaller biases than ML when the sample size is very large, and the missing data is MNAR (**Table 5.4 Panel 3**) in this case. However, it is worth stressing that the gaps between the performances of MI-LOGIT and ML are still marginal.

By comparing the performances to the benchmark method, the results show that PMM, MI-LOGIT and ML perform better than CCA in every aspect when the missing mechanism is MAR or MNAR. With MCAR missing data, the biases of the three MDT are greater than those of CCA when the missing rates are lower than 40%. Nevertheless, we also notice there is a clear trend that CCA starts losing its dominance as the sample size decreases.

5.6.5 An Ordinal variable in univariate missing data

The purpose of the last study is to investigate the effects on results of omitting an ordinal categorical variable on the standardized bias, RMSE and coverage rate. In order to balance the need to preserve the continuous nature and the accuracy of estimated coefficients, discretizing the variable from semi-continuous to ordinal sounds to be a compromising solution. As there is very little economic significance for the direct discretisation of Arrears, we switch the variable for imputation to Credit Risk in this case.

The main limitation for discretisation is computational feasibility. Computation times increase dramatically as the number of categorical levels increases to 15 in statistical analysis software such as STATA and SAS. For the sake of minimizing computing time, we simply transform Credit Risk to an ordinal variable with 4 levels only. In this paper, the indicator Credit Risk is generated based on whether a client has Arrears and whether his/her Loan Maturity is longer than three

years. Hence, the skewness of Arrears might be mitigated in the distribution of Credit Risk in our datasets.

As White et al. (2011) claimed that MI-OLOGIT and PMM have similar performances when imputing continuous variable as ordinal, in this subsection, we will examine whether MI-OLOGIT and PMM have similar performances with the data of a discretized semi-continuous variable.

Table 5.6 has presented the evaluations for the estimated coefficients of the ordinal categorical variable with missing data. In general, the coverage rates are higher than those in the previous tests with semi-continuous or even binary missing data (**Table 5.2 and 5.4**). On the other hand, the biases of all MDT shown here are lower than those in the previous tests and under the significant criterion of 40.

To better understand the source of biases, we rerun the tests on the discretized Loan Maturity only. We found that the biases of imputations for Loan Maturity are much lower than those for Arrears when both variables are dichotomized. This result confirms that the biases presented in **Table 5.4** are caused by the extremely high mass-point of the dichotomized Arrears. It explains the abnormality that MDT seems to perform better with ordinal categorical variables than with binary variables in this paper.

In **Table 5.6**, we can see that PMM has the lowest biases when the missing rates are very high, and the missing mechanism is MAR or MNAR. In contrast, MI-OLOGIT and ML have lower biases than PMM in the missing data with MCAR. Regarding the coverage rate, PMM performs slightly better than MI-OLOGIT and ML when the missing rates are low (≤20%) across all mechanisms. But the RMSEs of PMM are higher than those of MI-OLOGIT and ML in general.

The gaps between the biases of MI-OLOGIT and ML are noticeably larger than those in **Table 5.2** and **Table 5.4**, though their degrees are still ignorable. When the missing mechanism is MAR, MI-OLOGIT outperforms ML in terms of coverage rate and bias in most simulations, while the RMSEs of MI-OLOGIT is slightly lower than those of ML. In contrast, ML has lower biases than MI-OLOGIT when missing rates are higher than 10%, and the missingness is MCAR or MNAR.

By comparing the performance of the four MDT, we can see that the benchmark method CCA have higher coverage rates and lower biases than PMM, MI-OLOGIT and ML across different mechanisms in general. The only exception is when the missing rates are very low (10%) in the missing data with MCAR or MNAR.

5.6.6 An Ordinal variable in multivariate missing data

In this subsection, we further examine the effects on results of introducing extra missingness to Loan Maturity in the same dataset. The specifications of Loan Maturity are the same as what we described in subsections 2 and 4. Again, the performance of CCA is expected to be inferior compared to the other techniques when imputing multivariate missing data. As the skewness is mitigated, we are curious if the differences between the performances of different MDT in this subsection will be the same as those presented in subsection 4.

In **Table 5.7**, most of the biases are under the significant criterion of 40, and a noticeable number of coverage rates reach 90%. By comparing the results shown in **Table 5.5** and **Table 5.7**, we can see that both coverage rates and biases are similar. There is no significant influence on the performances of MDT by changing the variable of interest from the dichotomized Arrears to the ordinal Credit Risk. It may imply that the excessive biases are dominated by the joint distribution of different variables with missing data instead of the distributions of each one.

Regarding the comparison of MDT, the three performance indicators sometimes lead to very different conclusions. In terms of bias, MI-OLOGIT and ML outperform PMM in all simulations except for the largest samples (3,200) with MAR missingness. On the other hand, the coverage rates of PMM are higher than MI-OLOGIT and ML when the missing mechanism is MCAR or MAR. When the sample sizes of the MNAR datasets are very large, PMM's coverage rates are the highest as well. Regarding the RMSE, PMM performs better in all sizes of MCAR datasets, the MAR datasets with more than 1,700 cases, and the MNAR datasets with more than 2,700 cases.

When the missing mechanism is MAR (or MNAR), the biases of MI-OLOGIT are found to be smaller than those of ML in the simulations with sample sizes no smaller than 1,700 (or 3,200 for MNAR respectively). In the simulations with MCAR, MI-OLOGIT has smaller biases than ML only when the missing rates are very low (10%). These results are different from those with the condition of univariate missing data presented in **Table 5.6**. In terms of RMSE and coverage rate, ML outperforms MI-OLOGIT across all simulations when the mechanism is MAR or MNAR. On the contrary, MI-OLOGIT has better RMSEs and coverage rates than ML in most simulations of the MCAR datasets.

By comparing the performances to that of the benchmark method, we can see that PMM, MI-OLOGIT and ML have lower biases, lower RMSEs, and higher coverage rates than CCA in most of the simulations when the missing mechanism is MCAR or MNAR. In addition, all three MDT perform consistently better than CCA when the missing data is MAR.

5.7 Conclusions and Discussion

Regarding the previous empirical studies of missing data imputation, readers usually doubt the values of MI, ML or even PMM when population distribution is unknown, the sample size is too small, and the missing rate is not trivial. Most of the studies are Monte Carlo based, and few of them use real data. When researchers design their Monte Carlo studies, they have to subjectively select a small range of sample sizes and missing rates, instead of considering sample size and missing rate as the variables of interest. The reason is simple. Multivariate missing data imputation is very time-consuming. Runtime grows dramatically when the sample size rises, and the missing rate is close to 50%. As a result, the sample sizes (and missing rates) used in many studies are unrealistically large (and low).

To make the empirical findings of Monte Carlo studies applicable to real data, we need at least two assumptions for sample size (or missing rate): 1. the relation between imputation quality and sample size (or missing rate) is strictly linear; 2. the sensitivities between imputation quality and sample size (or missing rate) are the same for the MDT in comparison. However, the findings in this paper suggest that these assumptions are too strong for administrative loan book data. For instance, we found that PMM usually outperforms MI and ML when the sample sizes are large, and the missing rates are lows when the missing mechanism is MAR. Compared to MI and ML, PMM is more sensitive to the changing sample sizes and missing rates. It reminds us that we should not overestimate the capabilities of MDT and neglect the size effects.

The missing values in this paper are created by removing the Arrears (Y) corresponding to Gender (Z). In practice, missing values may occur corresponding to all ranges of values of the observed variables. Therefore, the actual biases associated with estimates of coefficients by MI, ML and PMM should be as severe as shown in this paper. These MDT are still the most promis-

ing methods before we know the underlying population distribution. MDT based on the true underlying population is always preferred. If prior information is available and properly included, then MI may outperform ML and PMM in small samples as it allows choosing informative priors.

However, we know that it is almost impossible to check the underlying population distribution behind a sample with missing values. Hence, a desirable missing data method needs to be robust to distribution violations. Our studies only focus on the non-normal distribution which is embedded in the microfinance loan book data. The results should be applicable to the commercial banks, credit unions, and other financial institutions as well.

Generally speaking, all MDT have comparatively lowest biases and highest coverage rates when the missing data is ordinal categorical. Most of their biases and coverage rates have exceeded the significant criterion when the missing data is semi-continuous. On the other hand, the MDT perform better with univariate missing data than with multivariate missing data. For semi-continuous data, we also found that sample size will affect the relationship between bias and missing rate. The biases are less sensitive to the changes of missing rates in small samples.

When the missing data are semi-continuous, PMM outperforms MI and ML in most simulations. For binary or ordinal categorical data, MI and ML are generally better than PMM. But we also notice that PMM performance surpasses MI and ML when the sample sizes are very large, the missing rates are low, and the missing mechanism is MAR.

In terms of the comparison between MI and ML, we found that MI performs better than ML when the missing data are semi-continuous, or when the missingness is MAR. Consistent with the findings in the prior literature, ML outperforms MI in small samples in general. However, it should be stressed that the differences between the biases of MI and ML are still marginal.

At last, we found that the MI, ML and PMM underperform the benchmark CCA in many simulations. In univariate missing data, CCA provides more accurate coefficient estimations in most simulations across different data types and missing mechanisms. The only exception is when the missing data are binary with MAR missingness. In multivariate missing data, MI, ML and PMM perform better than CCA in most simulations when the missing data are MAR or MNAR. But CCA is still preferable when the missing data are MCAR, and the missing rates are very low.

Table 5.1
Summary Statistics of 11 Variables (N=3,200)

Variable	Type	Distribution	Description
Outstanding	Continuous	Mean - 6,375.19,	The unpaid, interest-bearing balance of a
Balance	(in USD)	SD - 12,855.89,	loan averaged from the date of loan
		Range - [50, 10,000]	approval to the date of loan book update
Arrears	Semi-	Zeroes:2,672;	The part of a debt that is overdue for more
(raw data)	Continuous	Mean - 5.24,	than 30 days after missing one or more
	(in USD)	SD - 62.65,	required payments
		Range - [0, 2,000]	
Arrears	Binary	Zeroes:2,672;	The Arrears dichotomized to 2 levels
(dichotomized)		Positives:528	
Loan Maturity	Continuous	Mean - 43.98	The length of the period before the date
	(in Months)	SD - 53.66	that the full amount on the loan must be
		Range - [1, 240]	paid back to the microfinance lenders
Loan Maturity	Binary	Under 3 Years:1,836;	The Loan Maturity dichotomized to 2 levels
(dichotomized)		Over 3 Years :1,364	
Credit Risk	Ordinal	Very High:234;	Very High: positive Arrears and over 3 yrs to Maturity
	Categorical	High:294;	High: positve Arrears and no more than 3 yrs to Maturity
		Low:1,130;	Low: No Arrears and over 3 yrs to Maturity
		Very Low:1,542	Very Low: No Arrears and no more than 3 yrs to Maturity
Loan Purpose	Unorderred	Consumption:449;	Consumption and buying fixed assets are
	Categorical	Buying Fixed Assets:997;	non-productive activities;
		Agriculture:716;	Productive fixed assets are included in
		Commerce:649;	Agriculture, Manufacture, Sevice and etc.
		Manufacture: 32;	
		Service: 339;	
		Financing:18	
Gender	Binary	Male:1,731;	Gender of the microfinance clients
		Female:1,469	
Age	Continuous	Mean - 41.68,	Age of the microfinance clients
		SD - 12.27,	
		Range - [20, 80]	
Education	Ordinal	No School:543;	Secondary education includes middle
	Categorical	Primary School:986;	schools and high schools;
		Seconary School:1,288;	Tertiary education includes universities,
		Tertiary School:383	colleges, and technical training institues
Marital Status	Unorderred	Married: 1,779;	Noncouple includes single, divorced,
	Categorical	Cohabiting:513	separated, and widowed
		Noncouple: 908	

Table 5.2 Panel 1

RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Univariate MCAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

			RMSE				CP				BIAS			
			cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	|	4.50	4.19	4.82	4.81	0.96	0.98	0.94	0.93	4.67	1.92	5.46	5.67
3200	20%	|	5.12	4.68	6.37	6.39	0.90	0.96	0.89	0.89	6.75	3.60	7.55	7.96
3200	30%	|	10.00	10.94	10.42	10.50	0.82	0.87	0.78	0.77	23.19	13.04	28.21	28.37
3200	40%	|	16.10	18.15	16.62	16.70	0.72	0.72	0.66	0.66	33.43	34.52	34.63	35.62
3200	50%	|	17.39	23.80	20.24	20.44	0.65	0.61	0.63	0.62	103.26	112.94	124.74	125.68
2700	10%	|	5.59	5.69	6.16	6.18	0.94	0.96	0.93	0.93	2.48	0.84	1.91	1.90
2700	20%	|	7.65	7.66	7.84	7.94	0.88	0.90	0.87	0.87	11.43	6.78	12.76	13.04
2700	30%	|	15.27	16.07	15.82	15.93	0.78	0.80	0.75	0.75	34.03	31.40	36.69	37.51
2700	40%	|	20.34	23.65	22.51	22.70	0.69	0.69	0.67	0.67	52.39	60.71	62.42	63.12
2700	50%	|	24.27	30.32	29.42	29.71	0.62	0.59	0.61	0.59	85.92	116.14	110.57	112.03
2200	10%	|	7.54	8.01	8.03	8.03	0.94	0.95	0.93	0.92	10.39	9.50	11.22	11.44
2200	20%	|	11.93	13.26	13.53	13.54	0.87	0.88	0.87	0.87	22.42	23.40	25.42	25.62
2200	30%	|	17.65	18.78	17.84	17.87	0.80	0.80	0.78	0.78	43.95	50.66	51.43	51.96
2200	40%	|	25.01	27.72	27.70	27.94	0.67	0.67	0.64	0.65	56.67	74.28	71.71	72.75
2200	50%	|	27.52	36.76	30.89	31.18	0.56	0.53	0.52	0.52	93.03	120.36	101.65	102.86
1700	10%	|	12.00	12.55	12.61	12.65	0.91	0.91	0.90	0.90	20.72	24.59	23.85	24.13
1700	20%	|	15.45	19.05	18.12	18.26	0.82	0.81	0.80	0.80	5.57	11.63	9.24	9.41
1700	30%	|	22.01	25.71	25.46	25.52	0.76	0.76	0.75	0.75	30.58	41.13	39.83	40.07
1700	40%	|	25.79	32.42	29.45	29.69	0.68	0.67	0.67	0.67	60.34	73.28	70.30	71.45
1700	50%	|	32.69	41.29	36.92	37.29	0.56	0.55	0.53	0.54	86.16	128.76	118.39	121.05
1200	10%	|	15.05	16.48	15.98	16.00	0.90	0.90	0.90	0.90	5.76	7.28	5.80	6.23
1200	20%	|	17.22	19.79	19.78	19.82	0.87	0.83	0.84	0.84	20.12	25.85	23.85	24.28
1200	30%	|	27.94	32.57	30.99	31.17	0.76	0.73	0.73	0.73	34.88	49.45	38.99	39.90
1200	40%	|	35.44	41.34	38.75	39.13	0.63	0.61	0.63	0.63	51.67	71.47	66.64	68.67
1200	50%	|	44.44	55.00	51.17	51.94	0.58	0.54	0.56	0.57	52.89	83.42	72.43	74.58

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.2 Panel 2

RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Univariate MAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	6.77	6.24	7.36	7.37	0.84	0.97	0.86	0.86	25.88	9.41	28.80	29.24
3200	20%	8.75	8.25	10.00	10.10	0.80	0.92	0.78	0.79	28.61	11.96	34.91	35.00
3200	30%	14.17	13.73	15.74	15.84	0.78	0.86	0.74	0.73	36.21	21.17	41.15	41.31
3200	40%	20.36	21.94	23.51	23.58	0.71	0.73	0.65	0.65	60.20	54.57	70.64	70.94
3200	50%	23.98	27.83	27.26	27.45	0.55	0.53	0.49	0.50	120.07	129.49	141.57	143.64
2700	10%	8.87	8.48	9.63	9.66	0.89	0.94	0.86	0.86	29.15	19.51	30.85	31.07
2700	20%	12.66	12.24	13.18	13.26	0.79	0.85	0.78	0.78	37.39	28.97	41.20	41.31
2700	30%	17.83	18.09	18.17	18.28	0.72	0.75	0.67	0.67	44.86	42.64	52.47	53.06
2700	40%	23.37	25.44	26.74	26.96	0.63	0.64	0.59	0.59	68.62	75.68	81.87	82.58
2700	50%	25.82	30.93	30.33	30.48	0.59	0.54	0.54	0.54	91.71	102.40	112.23	112.36
2200	10%	10.27	9.89	11.06	11.09	0.88	0.91	0.84	0.85	27.63	20.68	32.44	32.77
2200	20%	16.24	18.20	18.77	18.84	0.81	0.83	0.79	0.79	41.20	38.59	47.66	48.19
2200	30%	25.16	28.04	29.80	29.91	0.68	0.68	0.64	0.65	62.85	64.36	70.43	71.74
2200	40%	29.37	31.26	31.47	31.69	0.64	0.61	0.59	0.58	75.71	87.09	90.69	92.07
2200	50%	32.19	38.61	38.58	38.78	0.58	0.55	0.55	0.55	110.52	145.52	130.32	133.60
1700	10%	13.25	13.53	16.46	16.47	0.88	0.89	0.84	0.84	31.66	30.24	36.47	36.76
1700	20%	22.74	25.96	27.75	27.80	0.76	0.77	0.74	0.74	46.82	45.41	54.03	54.74
1700	30%	25.51	32.19	32.25	32.50	0.73	0.72	0.70	0.69	62.68	65.40	64.77	65.85
1700	40%	28.73	34.98	34.76	35.13	0.63	0.61	0.60	0.61	108.42	131.85	120.76	123.62
1700	50%	36.68	41.97	40.55	41.02	0.60	0.55	0.56	0.54	110.63	136.83	131.17	134.85
1200	10%	17.24	18.38	19.83	19.93	0.86	0.87	0.84	0.84	25.19	23.90	27.09	27.35
1200	20%	25.21	26.37	27.14	27.29	0.79	0.79	0.78	0.78	38.86	46.25	48.83	49.63
1200	30%	29.09	35.49	34.78	35.14	0.71	0.69	0.69	0.70	53.52	66.85	63.44	65.17
1200	40%	39.47	44.92	42.99	43.47	0.62	0.58	0.59	0.59	67.20	85.65	79.67	81.09
1200	50%	48.01	58.43	56.54	57.50	0.54	0.52	0.52	0.53	72.13	85.91	80.05	84.30

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.2 Panel 3

RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Univariate MNAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	8.14	7.59	8.71	8.73	0.91	0.95	0.90	0.90	15.60	2.54	21.83	21.75
3200	20%	10.38	10.45	11.37	11.42	0.82	0.88	0.78	0.78	22.99	10.42	32.80	32.96
3200	30%	12.90	13.85	14.38	14.35	0.77	0.79	0.71	0.72	36.65	28.59	46.04	46.77
3200	40%	18.88	19.57	19.76	19.89	0.68	0.68	0.65	0.64	47.42	46.43	57.98	58.75
3200	50%	26.54	29.33	29.48	29.77	0.54	0.53	0.50	0.51	83.27	95.75	101.86	102.85
2700	10%	8.48	8.33	8.89	8.88	0.87	0.91	0.85	0.86	10.95	4.60	16.55	16.30
2700	20%	12.55	13.00	13.13	13.21	0.83	0.85	0.81	0.81	36.01	32.40	48.19	48.80
2700	30%	16.55	17.89	17.98	18.00	0.77	0.79	0.72	0.73	40.18	41.31	48.35	48.51
2700	40%	20.03	25.66	23.36	23.49	0.68	0.65	0.64	0.63	65.31	70.85	75.88	76.74
2700	50%	25.58	31.02	30.33	30.41	0.54	0.51	0.50	0.52	87.44	112.71	103.88	105.14
2200	10%	10.94	10.86	11.94	11.91	0.90	0.91	0.89	0.89	24.83	22.92	27.66	27.71
2200	20%	19.47	21.03	21.81	21.70	0.77	0.79	0.75	0.76	38.67	36.89	47.92	48.38
2200	30%	21.83	24.54	24.67	24.90	0.72	0.72	0.69	0.69	58.67	68.00	69.88	70.52
2200	40%	27.58	32.42	32.18	32.32	0.63	0.61	0.60	0.61	75.13	90.66	90.99	92.16
2200	50%	30.82	38.30	36.23	36.35	0.53	0.50	0.49	0.50	118.93	157.88	141.83	143.80
1700	10%	13.87	13.70	15.11	15.12	0.85	0.86	0.83	0.83	20.01	20.28	28.41	28.24
1700	20%	18.42	21.88	22.19	22.39	0.77	0.78	0.74	0.74	30.56	38.17	38.28	38.72
1700	30%	28.08	31.99	32.79	32.87	0.66	0.65	0.63	0.63	42.19	60.36	62.15	63.48
1700	40%	30.55	37.08	35.68	36.12	0.60	0.58	0.57	0.57	50.90	62.15	64.42	64.82
1700	50%	37.38	49.16	46.18	46.65	0.56	0.51	0.52	0.52	100.84	126.22	122.96	126.24
1200	10%	20.26	21.36	22.41	22.38	0.83	0.83	0.82	0.82	18.62	22.61	26.77	26.44
1200	20%	23.76	28.17	28.27	28.45	0.79	0.77	0.75	0.75	19.20	26.92	24.35	24.59
1200	30%	33.51	37.65	39.53	39.75	0.64	0.62	0.62	0.62	33.21	48.51	44.39	45.93
1200	40%	36.53	46.39	41.09	41.42	0.60	0.59	0.57	0.58	55.49	85.09	70.37	71.53
1200	50%	41.43	52.64	52.52	52.91	0.56	0.51	0.53	0.54	57.55	84.34	75.20	77.83

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.3 Panel 1

RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Multivariate MCAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	8.48	3.43	4.41	4.36	0.82	0.98	0.94	0.92	5.01	2.88	3.05	3.31
3200	20%	10.21	5.39	6.38	6.31	0.76	0.90	0.82	0.80	40.13	4.59	10.38	10.91
3200	30%	19.52	12.19	11.71	11.45	0.66	0.84	0.82	0.82	57.96	15.44	15.48	17.52
3200	40%	31.77	15.09	11.74	11.66	0.56	0.64	0.70	0.62	142.39	72.12	79.90	86.71
3200	50%	41.75	26.06	19.80	20.50	0.36	0.58	0.60	0.60	212.28	97.85	115.84	119.28
2700	10%	8.31	3.31	5.15	5.12	0.90	0.96	0.96	0.96	33.57	19.09	23.98	24.48
2700	20%	19.94	3.98	5.83	6.00	0.74	0.96	0.94	0.94	39.34	19.22	20.52	21.33
2700	30%	23.90	20.49	18.74	18.87	0.54	0.88	0.82	0.82	43.85	30.11	35.03	35.12
2700	40%	24.39	16.40	16.16	16.85	0.54	0.80	0.72	0.70	149.91	34.18	44.25	44.20
2700	50%	47.66	35.24	34.94	35.12	0.32	0.58	0.60	0.54	118.79	51.25	63.70	66.28
2200	10%	10.44	6.52	7.58	7.52	0.86	0.98	0.92	0.90	5.22	1.40	2.05	2.54
2200	20%	18.62	15.20	15.59	15.52	0.66	0.86	0.80	0.78	35.41	11.19	13.31	14.90
2200	30%	23.90	16.02	17.53	17.75	0.64	0.84	0.82	0.80	63.78	16.39	35.34	36.06
2200	40%	51.60	31.88	27.35	28.04	0.40	0.60	0.50	0.48	121.86	71.85	81.90	82.71
2200	50%	58.64	27.77	23.46	24.00	0.40	0.62	0.64	0.54	346.82	158.35	149.10	155.11
1700	10%	21.08	18.48	19.62	19.75	0.80	0.90	0.84	0.82	11.33	5.58	7.75	7.79
1700	20%	31.44	19.60	23.65	24.21	0.64	0.74	0.74	0.74	43.59	20.35	20.71	22.59
1700	30%	29.12	21.49	23.48	23.48	0.68	0.74	0.70	0.70	66.56	20.73	24.44	26.74
1700	40%	43.72	25.24	24.58	26.07	0.38	0.62	0.62	0.58	171.32	130.13	116.87	123.14
1700	50%	56.10	45.94	38.38	38.61	0.38	0.48	0.42	0.40	130.04	131.74	124.23	126.19
1200	10%	12.22	10.31	13.51	13.88	0.94	0.94	0.92	0.92	4.84	3.30	4.83	4.76
1200	20%	36.52	31.28	29.24	29.33	0.82	0.86	0.86	0.84	73.52	43.46	50.20	51.25
1200	30%	38.56	35.78	30.43	30.71	0.66	0.74	0.72	0.72	48.78	42.40	40.21	42.16
1200	40%	51.81	54.69	49.63	51.10	0.42	0.60	0.52	0.48	53.50	57.13	45.23	52.39
1200	50%	91.25	58.37	54.95	55.86	0.34	0.52	0.50	0.46	126.66	61.55	55.21	62.86

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.3 Panel 2

RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Multivariate MAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	8.74	9.29	9.38	9.41	0.88	0.96	0.82	0.82	12.93	15.52	20.37	20.83
3200	20%	20.11	14.73	17.29	17.31	0.62	0.90	0.60	0.60	36.91	17.20	60.85	61.03
3200	30%	37.04	31.25	35.26	35.20	0.40	0.62	0.30	0.28	226.46	121.25	205.63	206.20
3200	40%	48.63	31.82	37.24	37.21	0.42	0.62	0.36	0.34	387.91	208.64	304.79	304.79
3200	50%	69.19	46.23	58.82	59.00	0.24	0.48	0.16	0.16	412.17	297.55	374.93	375.10
2700	10%	9.47	9.53	13.97	13.90	0.86	0.90	0.86	0.84	21.32	24.87	46.21	46.91
2700	20%	20.55	16.26	22.10	22.18	0.66	0.80	0.66	0.66	62.56	55.57	73.98	74.45
2700	30%	32.36	25.30	29.09	29.05	0.46	0.60	0.46	0.44	222.90	163.91	202.68	204.37
2700	40%	54.40	38.52	43.99	44.11	0.34	0.52	0.28	0.26	185.19	165.76	242.92	244.44
2700	50%	65.03	56.83	61.97	62.03	0.20	0.30	0.12	0.12	404.56	363.55	489.53	493.51
2200	10%	17.41	17.33	19.41	19.20	0.88	0.90	0.86	0.86	33.13	33.83	34.06	34.65
2200	20%	23.84	19.38	24.88	24.39	0.68	0.82	0.70	0.70	83.37	57.46	93.19	94.42
2200	30%	38.05	33.64	41.83	41.80	0.48	0.68	0.56	0.52	237.17	157.63	203.10	205.48
2200	40%	46.81	39.85	51.06	51.24	0.40	0.52	0.34	0.32	188.38	178.65	262.67	266.22
2200	50%	70.15	53.37	68.95	68.83	0.18	0.30	0.24	0.24	328.37	227.60	294.18	295.11
1700	10%	20.06	22.82	22.89	22.94	0.80	0.82	0.76	0.76	11.07	18.74	19.56	20.35
1700	20%	23.61	22.06	22.99	23.03	0.70	0.78	0.76	0.70	75.05	65.27	69.75	71.09
1700	30%	38.23	27.16	32.73	32.97	0.40	0.58	0.48	0.46	134.53	103.79	128.05	128.64
1700	40%	63.39	59.27	67.06	67.15	0.47	0.48	0.36	0.36	234.76	180.89	205.77	209.23
1700	50%	85.34	54.80	63.26	63.34	0.26	0.38	0.30	0.28	455.21	281.69	345.02	350.23
1200	10%	22.49	29.49	33.86	33.31	0.80	0.88	0.84	0.84	35.87	36.65	57.55	57.95
1200	20%	34.93	33.87	39.01	38.93	0.75	0.76	0.74	0.70	75.48	74.84	98.92	103.94
1200	30%	41.18	39.49	60.20	60.69	0.50	0.58	0.50	0.50	130.84	124.35	138.49	138.73
1200	40%	47.48	39.70	39.16	40.18	0.48	0.58	0.56	0.52	133.34	125.57	156.30	158.81
1200	50%	78.73	71.55	77.55	77.99	0.20	0.34	0.26	0.26	224.32	202.19	233.66	235.12

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.3 Panel 3

RMSE and Coverage Rate of Estimated Coefficients for Semi-continuous Dependent Variable (Multivariate MNAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	1.03	1.06	1.23	1.15	1.00	1.00	1.00	1.00	2.22	3.23	3.32	3.25
3200	20%	12.70	6.20	6.72	6.82	0.90	0.96	0.94	0.94	14.92	4.16	4.26	5.18
3200	30%	14.41	6.33	6.85	6.99	0.76	0.92	0.78	0.76	20.52	7.25	3.67	2.45
3200	40%	23.74	14.99	12.15	12.37	0.58	0.76	0.74	0.74	55.13	24.68	24.22	25.37
3200	50%	33.83	18.46	15.87	16.72	0.32	0.62	0.58	0.58	113.21	23.52	12.69	15.36
2700	10%	1.41	1.43	1.63	1.57	1.00	1.00	1.00	1.00	3.91	4.11	4.76	4.47
2700	20%	12.95	3.83	3.89	3.98	0.88	0.92	0.92	0.92	14.04	6.94	7.85	7.77
2700	30%	18.69	13.95	15.41	15.60	0.72	0.86	0.84	0.84	35.92	28.97	20.50	21.83
2700	40%	32.11	18.67	14.45	15.03	0.40	0.76	0.64	0.62	40.18	37.64	22.09	20.04
2700	50%	36.41	30.21	23.48	24.70	0.38	0.61	0.60	0.58	207.51	158.51	120.17	120.25
2200	10%	0.98	1.77	1.90	1.78	1.00	1.00	1.00	1.00	1.07	5.70	3.64	3.92
2200	20%	18.64	4.87	6.14	6.31	0.84	0.96	0.92	0.92	74.40	36.67	33.60	33.99
2200	30%	22.81	14.58	16.64	16.82	0.70	0.86	0.80	0.78	70.70	47.90	33.81	36.68
2200	40%	29.85	14.01	10.77	11.27	0.62	0.82	0.74	0.70	140.61	116.57	103.92	106.54
2200	50%	47.49	35.95	32.35	33.70	0.40	0.56	0.56	0.46	173.14	52.77	43.29	49.54
1700	10%	1.40	1.61	1.43	1.42	0.98	1.00	1.00	1.00	4.40	9.87	8.84	8.23
1700	20%	19.02	13.15	10.71	10.73	0.76	0.90	0.88	0.88	35.08	20.04	9.47	10.55
1700	30%	25.66	26.55	24.62	24.76	0.68	0.80	0.74	0.74	92.87	70.57	55.48	55.79
1700	40%	33.47	23.78	21.79	21.87	0.42	0.68	0.66	0.64	118.91	68.83	53.95	62.05
1700	50%	41.93	41.99	32.75	33.51	0.36	0.56	0.54	0.50	165.64	108.96	74.78	69.83
1200	10%	1.94	4.65	3.01	2.93	1.00	1.00	1.00	1.00	10.49	28.39	11.17	12.24
1200	20%	17.05	11.97	11.88	11.91	0.82	0.92	0.88	0.88	5.61	11.68	4.26	4.67
1200	30%	45.43	40.78	34.57	35.55	0.78	0.80	0.78	0.74	44.54	57.52	42.80	42.59
1200	40%	42.49	28.02	27.23	27.69	0.58	0.72	0.72	0.68	60.25	53.08	49.76	49.76
1200	50%	74.17	53.67	47.31	50.27	0.34	0.54	0.44	0.42	94.61	96.00	92.61	107.48

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.4 Panel 1

RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Univariate MCAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	248.5	273.3	200.7	202.0	1.00	1.00	1.00	1.00	0.26	6.06	0.48	0.51
3200	20%	370.5	402.7	303.1	309.3	1.00	0.98	1.00	1.00	0.55	5.81	0.70	0.99
3200	30%	483.1	504.3	387.8	390.3	0.98	0.96	0.99	0.98	3.97	15.13	4.27	5.15
3200	40%	620.0	621.6	514.6	528.5	0.91	0.88	0.96	0.95	4.33	21.13	5.78	6.03
3200	50%	674.2	747.5	624.4	627.5	0.87	0.86	0.88	0.86	5.11	21.08	5.61	6.19
2700	10%	271.0	287.9	226.5	228.8	1.00	1.00	1.00	1.00	0.23	4.60	1.00	1.17
2700	20%	396.9	416.1	323.8	327.3	1.00	0.99	1.00	1.00	0.32	12.71	2.13	2.14
2700	30%	521.9	540.0	422.6	443.2	0.97	0.95	0.99	0.98	0.99	14.13	3.39	3.71
2700	40%	648.6	638.8	540.0	546.2	0.92	0.92	0.96	0.95	2.07	16.14	3.40	3.90
2700	50%	826.5	812.6	714.0	741.0	0.78	0.82	0.92	0.86	4.01	17.27	4.04	5.02
2200	10%	288.8	308.4	235.8	236.2	1.00	1.00	1.00	1.00	1.04	6.18	1.17	1.18
2200	20%	450.7	471.5	382.1	388.3	0.99	0.98	0.99	0.99	1.77	8.73	2.47	2.57
2200	30%	595.1	568.3	479.9	498.0	0.97	0.97	0.99	0.98	2.06	10.92	3.07	4.07
2200	40%	704.7	667.4	605.8	620.0	0.94	0.94	0.95	0.95	2.28	11.10	4.06	4.98
2200	50%	898.7	745.8	607.3	637.7	0.84	0.92	0.98	0.94	2.77	21.68	4.41	6.44
1700	10%	358.0	357.1	298.3	300.9	1.00	0.99	1.00	1.00	1.15	5.65	1.38	1.39
1700	20%	497.3	497.5	419.7	432.8	0.99	0.98	0.99	0.99	1.37	8.82	3.16	3.77
1700	30%	655.4	653.8	543.3	560.5	0.97	0.97	0.98	0.97	2.44	14.26	4.56	5.02
1700	40%	826.7	783.2	693.8	741.1	0.91	0.93	0.96	0.95	3.02	18.35	6.53	6.59
1700	50%	1092.9	976.8	823.7	929.0	0.82	0.88	0.94	0.90	3.60	23.21	10.27	10.60
1200	10%	391.9	413.4	318.0	325.9	1.00	0.99	1.00	1.00	2.04	3.99	1.22	1.35
1200	20%	618.9	632.8	500.6	540.7	0.99	0.97	0.99	0.99	2.70	6.38	1.37	1.84
1200	30%	740.4	651.1	631.7	642.6	0.98	0.98	0.96	0.96	2.73	10.25	1.44	2.43
1200	40%	1075.9	971.2	824.0	982.2	0.84	0.92	0.96	0.92	2.76	14.14	1.65	2.52
1200	50%	1150.8	1044.9	960.2	1079.0	0.85	0.88	0.90	0.88	3.13	14.56	2.64	3.63

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.4 Panel 2

RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Univariate MAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	296.8	294.4	**230.8**	234.9	1.00	1.00	**1.00**	1.00	6.57	**1.55**	**5.82**	6.13
3200	20%	402.8	390.3	**320.1**	326.2	1.00	1.00	**1.00**	1.00	7.23	**2.02**	**6.59**	7.58
3200	30%	518.0	516.8	**416.7**	419.7	0.97	0.96	**0.98**	0.98	7.82	10.15	**6.80**	7.88
3200	40%	640.2	604.3	**513.8**	520.7	0.90	0.93	**0.96**	0.95	9.41	11.81	**9.38**	9.52
3200	50%	743.8	685.7	**603.6**	636.4	0.83	0.86	**0.92**	0.90	11.45	14.30	**10.21**	10.49
2700	10%	328.0	343.3	**249.1**	254.1	1.00	1.00	**1.00**	1.00	8.79	**2.60**	**5.68**	7.14
2700	20%	414.8	404.0	**348.1**	352.8	1.00	0.99	**1.00**	1.00	8.90	**2.79**	**8.04**	8.77
2700	30%	559.4	536.5	**463.7**	476.7	0.97	0.97	**0.99**	0.98	10.07	**4.41**	**8.96**	9.55
2700	40%	686.6	653.4	**559.9**	580.0	0.92	0.92	**0.96**	0.96	10.45	10.83	**10.44**	11.32
2700	50%	836.5	731.5	**673.4**	716.9	0.81	0.87	**0.89**	0.88	11.11	11.08	**10.73**	11.56
2200	10%	374.4	389.2	**292.4**	298.4	1.00	1.00	**1.00**	1.00	7.93	**2.83**	**6.19**	7.66
2200	20%	474.0	467.0	**380.8**	393.9	1.00	1.00	**1.00**	1.00	9.52	**7.21**	**8.20**	9.11
2200	30%	612.2	575.8	**497.7**	507.9	0.98	0.97	**0.98**	0.98	13.33	**8.49**	**8.72**	9.27
2200	40%	762.9	707.8	**603.1**	632.7	0.92	0.93	**0.96**	0.95	13.36	12.94	**10.59**	11.68
2200	50%	901.6	808.3	**728.7**	799.5	0.85	0.87	**0.92**	0.89	13.66	13.57	**12.99**	13.01
1700	10%	390.2	383.7	**312.4**	323.5	1.00	1.00	**1.00**	1.00	9.68	**3.48**	**6.92**	8.99
1700	20%	541.9	513.1	**429.7**	448.1	1.00	0.99	**1.00**	1.00	10.01	**8.03**	**9.16**	9.67
1700	30%	677.5	630.9	**554.8**	584.0	0.98	0.97	**0.98**	0.98	10.10	**8.67**	**9.65**	10.09
1700	40%	928.8	877.2	**654.1**	678.4	0.96	0.90	**0.96**	0.94	13.19	13.12	**9.99**	10.77
1700	50%	1043.4	824.8	**820.7**	958.3	0.78	0.88	**0.86**	0.86	14.85	14.36	**12.25**	13.83
1200	10%	458.8	484.2	**366.4**	397.3	1.00	0.99	**1.00**	0.99	9.91	**3.49**	**2.19**	2.58
1200	20%	653.1	656.8	**641.8**	676.1	0.98	1.00	**0.98**	0.96	10.31	**8.66**	**5.14**	7.70
1200	30%	855.3	828.9	**641.3**	674.9	0.96	0.96	**1.00**	0.98	11.05	9.88	**7.18**	9.24
1200	40%	946.2	924.9	**853.8**	960.9	0.90	0.90	**0.92**	0.90	12.56	14.60	**8.12**	9.33
1200	50%	1011.6	888.2	**862.7**	951.8	0.92	0.96	**0.94**	0.88	14.30	16.78	**8.36**	9.76

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.4 Panel 3

RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Univariate MNAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	357.3	380.0	291.6	301.7	0.99	0.98	1.00	0.99	1.05	8.31	2.49	2.08
3200	20%	459.3	477.2	375.2	392.5	0.98	0.95	0.99	0.98	1.18	14.50	3.23	2.48
3200	30%	578.7	592.7	474.8	496.4	0.93	0.92	0.96	0.95	1.64	15.30	3.86	3.27
3200	40%	691.7	698.0	571.0	591.8	0.87	0.85	0.92	0.91	1.89	15.42	4.80	3.77
3200	50%	814.8	777.2	686.6	714.4	0.78	0.79	0.85	0.85	2.29	16.56	5.04	3.95
2700	10%	390.5	414.0	304.1	323.2	0.99	0.98	1.00	1.00	1.33	10.67	2.35	1.52
2700	20%	513.3	539.7	430.1	449.9	0.97	0.96	0.98	0.98	1.43	11.14	2.62	1.91
2700	30%	617.3	631.9	507.6	539.9	0.95	0.92	0.97	0.96	1.95	11.46	3.19	2.45
2700	40%	732.6	704.9	608.1	637.9	0.88	0.89	0.92	0.91	4.80	11.48	6.81	5.60
2700	50%	892.5	825.9	742.8	794.0	0.78	0.81	0.87	0.85	6.76	19.93	9.73	8.44
2200	10%	413.5	420.0	332.8	358.2	0.99	0.98	1.00	0.99	0.24	6.03	1.50	0.59
2200	20%	594.6	584.0	495.1	525.8	0.97	0.96	0.98	0.98	0.40	8.87	1.91	1.00
2200	30%	709.0	690.5	600.4	636.2	0.93	0.94	0.96	0.95	0.57	9.29	2.18	1.33
2200	40%	784.2	730.2	665.7	689.3	0.91	0.91	0.93	0.93	0.86	16.26	2.71	2.24
2200	50%	969.7	895.8	826.0	893.7	0.78	0.82	0.86	0.84	1.22	16.34	6.31	4.48
1700	10%	467.4	459.5	362.3	384.9	1.00	0.99	1.00	1.00	0.19	4.06	0.82	0.72
1700	20%	602.5	604.6	495.5	520.2	0.98	0.97	0.99	0.99	0.58	7.42	1.44	1.29
1700	30%	703.2	665.4	583.5	611.6	0.96	0.95	0.97	0.97	1.11	12.45	4.78	3.81
1700	40%	898.2	831.2	721.5	790.8	0.90	0.92	0.95	0.93	1.43	12.91	5.26	4.33
1700	50%	1125.5	931.4	886.3	974.9	0.80	0.89	0.89	0.88	8.01	19.35	12.40	11.28
1200	10%	561.4	563.1	445.5	493.6	0.99	0.98	1.00	0.99	1.45	5.76	1.84	1.51
1200	20%	706.7	659.0	563.9	604.2	0.99	0.99	1.00	0.99	1.48	6.38	2.54	2.34
1200	30%	842.0	776.3	697.2	754.7	0.97	0.97	0.98	0.97	1.51	9.11	5.75	4.49
1200	40%	1233.9	1162.7	1112.3	1250.5	0.80	0.84	0.94	0.82	2.25	12.94	6.17	5.40
1200	50%	1619.5	1233.1	1148.8	1254.8	0.68	0.78	0.82	0.78	5.26	17.42	6.21	5.64

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.5 Panel 1

RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Multivariate MCAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	294.4	225.5	130.6	125.1	1.00	1.00	1.00	1.00	0.10	3.46	0.11	0.10
3200	20%	453.8	328.6	197.0	186.0	0.98	1.00	1.00	1.00	0.17	4.20	0.30	0.62
3200	30%	592.2	392.6	272.6	251.8	0.94	0.99	1.00	1.00	0.28	5.42	0.35	0.78
3200	40%	877.7	508.0	356.2	347.3	0.76	0.96	1.00	1.00	1.30	10.88	1.38	1.85
3200	50%	1132.4	545.5	413.4	422.1	0.62	0.94	0.98	0.98	6.62	19.46	4.23	4.28
2700	10%	311.7	246.1	139.7	132.1	1.00	1.00	1.00	1.00	0.15	4.72	0.72	0.71
2700	20%	534.1	328.3	232.1	216.2	0.98	1.00	1.00	1.00	0.71	12.43	2.01	2.36
2700	30%	741.5	429.1	298.7	292.7	0.87	0.99	1.00	1.00	1.12	13.52	2.31	2.38
2700	40%	923.7	495.6	396.0	384.5	0.73	0.97	1.00	1.00	2.30	16.39	2.82	3.95
2700	50%	1329.0	639.8	490.6	494.7	0.55	0.91	0.98	0.97	7.89	22.08	3.31	4.31
2200	10%	354.3	217.5	182.9	176.1	0.99	1.00	1.00	1.00	0.15	5.62	0.89	0.70
2200	20%	663.6	441.9	318.8	296.1	0.97	1.00	1.00	1.00	0.86	13.52	1.41	2.31
2200	30%	879.1	535.2	415.0	413.7	0.90	1.00	1.00	1.00	1.81	14.92	2.46	3.91
2200	40%	1155.4	565.3	469.2	446.6	0.72	0.99	1.00	1.00	4.99	20.16	5.05	6.38
2200	50%	1491.2	715.9	610.4	611.0	0.58	0.86	0.97	0.97	10.36	23.68	5.11	6.80
1700	10%	378.5	275.1	242.6	239.6	1.00	1.00	1.00	1.00	0.22	6.71	1.98	1.87
1700	20%	782.9	526.8	363.5	346.1	0.97	0.99	1.00	1.00	1.18	14.89	2.70	2.99
1700	30%	1091.9	583.8	514.3	513.5	0.87	0.97	0.99	0.99	2.39	15.12	4.48	5.08
1700	40%	1151.8	691.4	642.6	594.9	0.85	0.98	0.99	0.99	3.63	21.61	4.82	5.22
1700	50%	1277.2	799.9	722.9	737.9	0.61	0.92	0.97	0.96	13.06	32.12	8.87	10.56
1200	10%	449.2	345.6	257.1	242.9	1.00	1.00	1.00	1.00	0.78	8.09	0.99	0.92
1200	20%	788.9	528.4	518.4	488.6	0.98	1.00	1.00	1.00	2.70	19.57	4.01	5.16
1200	30%	990.2	618.3	615.5	560.4	0.96	0.99	1.00	1.00	6.24	23.51	5.26	6.15
1200	40%	1249.1	722.6	676.2	657.5	0.81	0.99	0.99	0.98	14.04	24.08	8.74	8.98
1200	50%	1774.6	857.3	829.5	829.9	0.72	0.96	0.96	0.96	20.23	34.70	9.02	12.01

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.5 Panel 2

RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Multivariate MAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	536.9	223.6	307.4	314.7	1.00	1.00	1.00	1.00	14.73	1.19	9.11	14.40
3200	20%	982.6	397.1	443.0	458.0	0.82	0.98	0.98	0.98	17.96	2.59	12.79	20.13
3200	30%	1207.6	520.1	546.9	570.4	0.63	0.94	0.94	0.94	31.73	2.88	20.07	21.32
3200	40%	1912.4	810.4	813.0	869.1	0.51	0.91	0.88	0.87	41.61	3.05	21.19	22.74
3200	50%	2836.6	803.6	978.2	1117.1	0.28	0.82	0.61	0.60	46.39	5.34	21.12	30.42
2700	10%	571.0	343.0	352.3	353.7	0.96	1.00	1.00	1.00	15.81	1.38	10.93	17.77
2700	20%	984.9	471.2	510.8	527.0	0.73	0.98	0.98	0.98	35.03	3.00	13.49	21.03
2700	30%	1396.1	601.0	608.7	640.3	0.67	0.98	0.96	0.96	35.83	3.03	21.23	22.54
2700	40%	2173.6	809.1	860.8	936.7	0.52	0.83	0.83	0.80	46.74	4.25	21.52	23.53
2700	50%	3230.3	1094.3	1190.5	1381.1	0.41	0.82	0.73	0.68	52.24	6.25	22.60	31.20
2200	10%	685.7	363.9	407.1	407.6	0.93	1.00	1.00	1.00	25.33	1.63	11.49	21.74
2200	20%	1182.0	520.1	546.9	570.4	0.75	0.98	0.98	0.98	28.27	3.04	14.92	16.43
2200	30%	1523.9	683.3	713.7	739.5	0.74	0.93	0.96	0.87	32.88	3.44	22.32	27.70
2200	40%	2234.5	901.7	945.4	1013.9	0.51	0.85	0.83	0.77	48.38	6.88	27.78	32.29
2200	50%	3364.1	1100.7	1295.5	1488.5	0.41	0.82	0.71	0.60	53.57	7.64	32.50	34.97
1700	10%	740.2	373.9	412.7	424.5	0.91	1.00	1.00	1.00	3.91	1.86	12.98	15.36
1700	20%	1207.6	555.0	562.2	600.2	0.73	0.98	0.98	0.98	34.88	7.52	22.66	25.63
1700	30%	1641.0	712.1	734.3	747.7	0.72	0.98	0.96	0.87	39.19	11.75	23.84	28.02
1700	40%	2522.0	1021.4	1049.5	1169.5	0.59	0.85	0.84	0.82	42.11	22.97	29.10	30.07
1700	50%	3808.6	1140.2	1299.0	1516.3	0.32	0.75	0.72	0.58	56.54	26.75	29.57	32.29
1200	10%	794.1	404.8	420.2	444.3	0.86	1.00	1.00	1.00	25.73	9.62	17.30	18.25
1200	20%	1271.1	575.3	695.6	728.2	0.65	0.95	0.89	0.88	37.55	20.62	27.00	29.04
1200	30%	1700.4	809.7	854.3	900.3	0.58	0.87	0.85	0.82	47.61	20.73	28.96	31.18
1200	40%	2546.4	1025.6	1080.9	1192.9	0.40	0.68	0.58	0.54	68.88	34.32	36.50	44.59
1200	50%	4108.6	1224.0	1399.9	1649.1	0.37	0.65	0.63	0.45	52.53	34.90	39.14	48.67

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.5 Panel 3

RMSE and Coverage Rate of Estimated Coefficients for Binary Dependent Variable (Multivariate MNAR Missing Data) – Grouping by Sample sizes

The logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	673.6	405.4	**297.8**	322.9	0.94	**1.00**	**1.00**	1.00	1.93	**0.72**	**1.07**	1.08
3200	20%	1040.1	570.4	**517.1**	539.8	0.70	**0.97**	**0.96**	0.97	3.38	**1.35**	**1.36**	1.44
3200	30%	1490.5	681.2	**592.8**	647.3	0.60	**0.93**	**0.90**	0.92	7.82	**3.93**	**4.38**	5.12
3200	40%	2031.0	**809.1**	**844.7**	853.8	0.46	**0.88**	**0.88**	0.82	11.18	**4.61**	**5.26**	6.89
3200	50%	2395.1	**1050.1**	**1154.8**	1275.6	0.30	**0.73**	**0.67**	0.63	13.85	11.53	**10.62**	10.80
2700	10%	797.7	437.9	**386.1**	400.2	0.92	0.99	**1.00**	1.00	2.40	4.83	2.51	1.12
2700	20%	1057.5	630.3	**535.8**	585.1	0.66	0.97	**0.97**	0.97	6.07	5.56	4.15	1.59
2700	30%	1584.9	728.7	**631.0**	636.1	0.59	0.85	**0.92**	0.92	10.00	10.45	6.18	5.69
2700	40%	2393.3	**815.1**	**847.8**	866.5	0.45	0.82	**0.84**	0.80	13.05	21.54	9.29	8.01
2700	50%	2835.9	**1056.4**	**1392.6**	1471.3	0.30	0.72	**0.73**	0.65	14.66	22.98	14.26	11.97
2200	10%	813.9	507.7	**424.1**	436.4	0.84	0.98	**1.00**	1.00	3.18	6.42	3.68	1.43
2200	20%	1251.9	654.5	**583.1**	590.1	0.66	0.97	**0.99**	0.97	7.95	9.75	5.92	3.85
2200	30%	1825.2	769.4	**787.3**	811.3	0.48	0.82	**0.94**	0.92	16.76	14.58	8.10	8.81
2200	40%	2640.3	**965.1**	**967.2**	1005.5	0.43	0.80	**0.85**	0.78	24.40	23.40	10.81	9.18
2200	50%	3333.4	**1147.5**	**1465.1**	1607.0	0.29	0.68	**0.72**	0.60	26.37	31.10	18.76	14.96
1700	10%	966.0	531.3	**438.7**	443.4	0.76	0.98	**1.00**	1.00	6.01	7.18	4.67	3.62
1700	20%	1312.3	755.9	**612.2**	642.7	0.64	0.97	**0.98**	0.93	8.81	13.47	7.17	4.44
1700	30%	1838.0	902.4	**833.5**	854.4	0.54	0.83	**0.87**	0.92	17.33	20.23	13.10	8.52
1700	40%	2646.3	**973.8**	**1139.6**	1095.5	0.40	0.78	**0.84**	0.74	27.06	25.31	14.74	11.94
1700	50%	3900.2	**1151.8**	**1523.5**	1543.3	0.28	0.65	**0.66**	0.62	47.98	43.40	21.03	16.40
1200	10%	1171.3	725.2	**527.4**	591.9	0.76	0.98	**1.00**	1.00	14.22	20.63	5.08	4.09
1200	20%	1359.3	891.5	**730.2**	793.4	0.53	**0.96**	**0.94**	0.93	15.04	21.50	12.14	7.18
1200	30%	2315.5	955.1	**950.3**	962.1	0.53	**0.82**	**0.79**	0.78	21.25	22.44	13.70	9.82
1200	40%	3063.6	**1000.2**	**1020.3**	1130.3	0.36	**0.74**	**0.60**	0.53	33.46	34.21	20.09	14.62
1200	50%	4871.4	**1228.3**	**1613.2**	1757.4	0.22	**0.58**	**0.56**	0.54	49.61	45.28	24.14	20.15

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.6 Panel 1

RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Univariate MCAR Missing Data) – Grouping by Sample sizes

The ordered logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	65.3	102.7	46.3	51.5	1.00	1.00	1.00	1.00	0.29	6.70	0.63	0.28
3200	20%	100.2	160.5	74.5	76.5	1.00	1.00	1.00	1.00	0.29	12.55	1.31	0.47
3200	30%	125.1	218.2	96.5	107.8	1.00	0.99	1.00	1.00	0.30	18.49	2.26	1.79
3200	40%	159.9	271.4	116.9	119.0	1.00	0.97	1.00	1.00	0.32	24.04	3.19	2.17
3200	50%	189.9	316.1	156.6	163.8	0.99	0.91	0.99	0.99	0.33	28.81	3.50	2.34
2700	10%	76.5	116.8	51.1	56.1	1.00	1.00	1.00	1.00	0.40	7.09	0.69	0.37
2700	20%	114.4	180.5	79.8	89.9	1.00	1.00	1.00	1.00	0.41	12.75	1.34	1.00
2700	30%	139.5	230.8	104.7	104.8	1.00	0.99	1.00	1.00	0.42	18.59	2.43	1.29
2700	40%	172.9	266.5	132.3	135.1	1.00	0.98	1.00	1.00	0.68	23.64	2.94	2.22
2700	50%	222.5	327.3	171.4	167.8	0.99	0.91	0.99	0.99	0.73	30.06	3.07	2.76
2200	10%	92.6	134.9	50.3	68.1	1.00	1.00	1.00	1.00	0.43	7.05	0.39	0.51
2200	20%	141.6	178.7	88.4	95.6	1.00	1.00	1.00	1.00	0.46	10.30	2.12	1.01
2200	30%	184.3	246.4	105.2	133.7	1.00	0.99	1.00	1.00	0.53	18.93	3.81	3.41
2200	40%	240.2	306.7	163.7	162.7	1.00	0.98	1.00	1.00	0.54	23.13	3.90	3.61
2200	50%	295.6	322.1	203.7	200.3	0.99	0.90	0.99	0.99	0.63	26.04	4.41	3.93
1700	10%	108.1	146.4	58.4	74.5	1.00	1.00	1.00	1.00	0.37	6.26	0.30	0.58
1700	20%	171.5	231.5	108.9	111.6	1.00	1.00	1.00	1.00	0.42	13.43	1.61	1.30
1700	30%	221.6	269.6	133.3	126.8	1.00	0.99	1.00	1.00	0.53	14.17	2.04	1.90
1700	40%	280.2	352.6	187.9	180.6	1.00	0.96	1.00	0.99	0.97	17.07	3.43	2.31
1700	50%	289.1	186.8	241.2	210.0	1.00	0.90	0.98	0.97	1.45	26.49	5.65	2.40
1200	10%	98.1	152.6	97.3	93.0	1.00	1.00	1.00	1.00	0.23	3.87	0.22	0.27
1200	20%	115.7	228.7	122.9	105.2	1.00	1.00	1.00	1.00	0.25	7.64	1.28	0.31
1200	30%	175.0	289.9	164.3	144.0	1.00	0.99	1.00	1.00	0.33	10.32	1.64	1.35
1200	40%	421.2	480.7	268.7	260.5	1.00	0.94	1.00	0.99	1.94	31.86	2.15	1.73
1200	50%	609.1	508.3	363.2	356.9	0.99	0.90	0.97	0.97	2.20	33.56	2.23	2.80

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.6 Panel 2

RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Univariate MAR Missing Data) – Grouping by Sample sizes

The ordered logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	104.9	87.8	85.4	73.4	1.00	1.00	1.00	1.00	2.22	5.15	5.44	6.20
3200	20%	144.4	143.3	128.5	124.6	1.00	1.00	1.00	1.00	2.25	7.90	7.07	7.27
3200	30%	170.3	218.2	138.8	133.9	1.00	0.99	1.00	1.00	4.32	16.80	8.21	7.57
3200	40%	200.1	278.1	154.8	162.6	0.99	0.96	1.00	0.99	7.32	25.00	9.52	8.76
3200	50%	260.9	288.8	216.2	223.0	0.92	0.92	0.97	0.95	10.89	23.26	13.52	12.86
2700	10%	83.4	98.3	95.6	89.7	1.00	1.00	1.00	1.00	2.50	6.24	7.02	7.40
2700	20%	154.9	199.3	123.6	121.5	1.00	1.00	1.00	1.00	4.07	13.45	7.52	7.71
2700	30%	188.9	225.8	133.9	151.7	1.00	0.99	1.00	1.00	5.01	16.14	7.68	7.90
2700	40%	244.1	279.5	209.1	212.9	0.99	0.95	0.98	0.95	5.20	20.79	11.57	10.47
2700	50%	301.9	304.9	225.0	234.4	0.91	0.91	0.96	0.94	13.01	24.68	13.07	12.23
2200	10%	120.0	101.5	106.5	109.2	1.00	1.00	1.00	1.00	4.34	6.34	6.64	7.21
2200	20%	195.7	210.5	152.9	152.2	1.00	1.00	1.00	1.00	5.45	10.13	9.43	9.47
2200	30%	250.0	268.9	169.7	185.4	1.00	0.98	1.00	0.99	6.84	16.95	9.65	10.33
2200	40%	294.0	348.5	213.8	216.1	0.99	0.95	0.98	0.99	7.27	26.65	10.08	10.59
2200	50%	330.2	404.2	250.6	264.3	0.95	0.84	0.96	0.96	12.85	34.78	13.26	12.02
1700	10%	137.0	115.6	117.3	116.8	1.00	1.00	1.00	1.00	4.69	6.41	6.60	7.88
1700	20%	166.9	158.3	194.3	172.6	1.00	1.00	1.00	1.00	5.62	6.97	7.63	8.90
1700	30%	214.3	223.6	214.7	231.2	1.00	0.98	1.00	0.99	6.92	15.89	11.69	13.05
1700	40%	270.9	272.3	267.1	284.6	1.00	0.93	0.98	0.95	13.00	25.71	13.31	15.83
1700	50%	381.3	386.6	306.1	322.2	0.93	0.83	0.95	0.95	14.29	35.75	20.77	17.27
1200	10%	152.6	114.6	164.7	156.0	1.00	1.00	1.00	1.00	5.56	6.64	6.82	8.78
1200	20%	190.4	150.9	174.3	152.5	1.00	1.00	1.00	1.00	6.08	7.77	8.85	9.42
1200	30%	244.2	300.9	224.0	266.9	1.00	0.98	0.98	0.98	7.64	9.81	9.28	9.68
1200	40%	456.8	503.3	304.7	369.1	0.99	0.92	0.98	0.95	15.51	33.51	17.84	18.09
1200	50%	607.6	554.1	431.8	462.6	0.90	0.82	0.95	0.94	19.96	36.52	22.32	16.37

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.6 Panel 3

RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Univariate MNAR Missing Data) – Grouping by Sample sizes

The ordered logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	123.5	120.3	114.9	124.4	1.00	1.00	1.00	1.00	2.00	1.15	6.75	7.37
3200	20%	155.6	161.9	128.8	126.2	1.00	1.00	1.00	1.00	3.53	6.80	9.23	7.47
3200	30%	166.5	203.7	141.0	136.4	1.00	0.99	1.00	1.00	5.43	14.45	11.66	9.90
3200	40%	201.8	250.4	150.2	140.5	1.00	0.96	1.00	0.99	6.20	19.68	11.87	11.14
3200	50%	257.2	276.9	197.0	180.0	0.94	0.93	0.99	0.98	6.38	21.97	12.42	11.69
2700	10%	125.8	141.2	116.9	120.2	1.00	1.00	1.00	1.00	1.87	1.60	7.01	7.50
2700	20%	174.0	172.1	128.7	126.7	1.00	1.00	1.00	1.00	2.88	6.85	9.89	9.12
2700	30%	190.4	242.9	184.3	173.6	1.00	0.97	1.00	1.00	4.24	17.29	10.86	10.68
2700	40%	242.1	261.8	208.0	196.2	0.98	0.97	0.99	0.99	6.36	18.44	15.18	13.29
2700	50%	276.2	325.8	209.4	208.8	0.93	0.89	0.99	0.96	6.82	25.54	17.62	16.41
2200	10%	173.9	190.6	157.7	169.1	1.00	1.00	1.00	1.00	3.06	1.97	8.28	9.07
2200	20%	215.1	226.4	176.8	169.2	1.00	1.00	1.00	1.00	4.77	6.86	11.59	9.50
2200	30%	241.4	242.5	179.7	177.4	1.00	0.99	1.00	1.00	6.85	11.32	11.86	10.85
2200	40%	288.6	300.7	202.6	196.6	0.98	0.98	1.00	0.99	7.39	20.31	13.11	13.71
2200	50%	363.9	375.2	222.2	212.0	0.91	0.89	1.00	0.99	7.56	26.59	13.76	13.96
1700	10%	186.2	224.5	145.0	161.6	1.00	1.00	1.00	1.00	3.14	2.98	8.66	9.14
1700	20%	232.8	240.6	169.0	168.0	1.00	1.00	1.00	1.00	5.43	7.78	11.97	10.32
1700	30%	279.7	299.0	182.9	171.4	1.00	0.98	1.00	1.00	7.76	15.96	12.20	11.65
1700	40%	345.5	382.6	206.6	200.0	0.97	0.93	1.00	1.00	8.71	26.41	14.12	12.59
1700	50%	425.0	396.7	353.7	341.2	0.91	0.88	0.95	0.94	8.93	27.87	23.76	20.74
1200	10%	137.0	124.9	111.5	116.5	1.00	1.00	1.00	1.00	4.19	3.89	9.68	10.67
1200	20%	187.4	151.5	174.0	168.8	1.00	1.00	1.00	1.00	6.02	9.00	13.63	12.02
1200	30%	261.7	228.4	259.5	248.7	0.97	0.93	1.00	1.00	8.09	16.67	15.36	15.09
1200	40%	440.0	451.2	283.3	268.1	0.97	0.91	0.98	0.98	9.20	21.92	19.62	18.52
1200	50%	612.3	545.1	377.4	361.3	0.86	0.87	0.98	0.96	9.56	35.52	31.27	29.69

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.7 Panel 1

RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Multivariate MCAR Missing Data) – Grouping by Sample sizes

The ordered logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	108.7	78.3	80.4	97.3	1.00	1.00	1.00	1.00	0.45	8.09	0.42	0.51
3200	20%	191.5	130.0	157.2	189.9	0.99	1.00	0.99	0.97	1.10	11.94	0.80	0.57
3200	30%	266.7	221.9	232.9	261.4	0.94	0.96	0.96	0.89	1.68	22.97	5.68	1.00
3200	40%	302.2	269.9	271.0	300.4	0.85	0.95	0.91	0.87	1.79	26.44	4.80	1.16
3200	50%	489.8	375.6	393.7	417.9	0.63	0.80	0.88	0.88	2.44	35.15	6.85	2.58
2700	10%	130.4	102.7	118.4	148.3	1.00	1.00	1.00	1.00	0.47	6.21	0.54	0.67
2700	20%	193.2	163.7	173.7	195.5	1.00	1.00	0.99	0.98	1.69	13.89	1.56	1.42
2700	30%	293.4	235.8	242.0	273.9	0.93	0.99	0.98	0.90	1.74	19.66	3.03	1.42
2700	40%	399.3	303.9	305.6	355.7	0.84	0.95	0.89	0.82	1.82	27.28	4.51	1.85
2700	50%	500.8	406.1	389.3	429.1	0.62	0.89	0.81	0.73	3.32	36.04	12.94	2.68
2200	10%	161.9	133.7	136.3	170.6	0.99	1.00	1.00	0.98	0.54	8.16	0.45	0.68
2200	20%	267.7	189.5	248.4	281.0	1.00	1.00	0.98	0.94	1.35	11.55	3.25	1.43
2200	30%	354.9	271.3	309.2	360.1	0.93	0.98	0.98	0.90	3.05	20.84	3.83	3.53
2200	40%	465.2	267.9	365.3	433.6	0.83	0.93	0.88	0.81	3.71	21.25	6.23	3.60
2200	50%	653.9	398.3	481.0	548.5	0.62	0.85	0.78	0.67	4.78	31.81	14.02	11.33
1700	10%	188.9	123.3	173.5	178.9	1.00	1.00	1.00	0.97	0.58	5.70	1.14	1.29
1700	20%	248.2	186.5	179.8	208.2	1.00	1.00	0.98	0.94	1.43	13.52	1.60	1.36
1700	30%	308.9	191.9	239.8	238.9	0.93	1.00	0.97	0.90	3.47	15.02	6.31	2.02
1700	40%	310.2	247.6	281.1	255.6	0.82	0.99	0.87	0.80	3.39	22.01	6.31	3.55
1700	50%	855.5	462.8	658.7	611.2	0.55	0.77	0.76	0.64	8.02	34.08	19.20	5.46
1200	10%	170.7	104.0	107.7	98.1	1.00	1.00	1.00	0.97	0.64	3.10	1.64	1.73
1200	20%	206.7	148.3	156.0	153.1	1.00	1.00	0.98	0.94	1.93	7.17	1.87	1.78
1200	30%	322.0	218.2	237.4	223.1	0.93	1.00	0.96	0.90	3.25	14.74	2.80	2.55
1200	40%	415.9	261.5	273.6	262.0	0.81	0.99	0.87	0.79	3.41	15.19	3.62	3.52
1200	50%	616.2	293.0	352.3	352.4	0.54	0.76	0.75	0.63	15.47	16.54	4.40	4.12

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.7 Panel 2

RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Multivariate MAR Missing Data) – Grouping by Sample sizes

The ordered logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	260.1	109.3	188.0	174.0	0.98	1.00	1.00	1.00	10.19	1.23	7.59	15.46
3200	20%	465.2	142.5	286.1	278.9	0.82	1.00	0.87	0.90	13.35	3.83	14.93	25.09
3200	30%	671.8	181.7	390.5	315.2	0.58	0.99	0.75	0.87	16.13	4.66	16.41	26.33
3200	40%	958.8	213.8	550.9	482.4	0.47	0.96	0.56	0.67	24.25	4.67	18.86	29.17
3200	50%	1503.2	314.5	675.1	654.8	0.37	0.83	0.47	0.49	42.20	9.84	27.08	33.51
2700	10%	219.6	200.1	202.9	176.9	0.98	1.00	1.00	1.00	12.88	14.10	5.68	11.53
2700	20%	388.1	223.5	316.1	218.4	0.81	0.99	0.88	0.94	13.60	16.44	5.76	15.27
2700	30%	466.6	281.6	406.2	260.3	0.68	0.98	0.77	0.87	17.69	17.73	8.14	20.98
2700	40%	905.9	303.2	577.6	410.4	0.42	0.91	0.56	0.74	25.98	19.05	10.52	27.83
2700	50%	1336.6	387.7	719.1	489.0	0.37	0.80	0.48	0.71	44.19	22.81	13.33	31.06
2200	10%	293.7	218.5	216.6	176.2	0.98	1.00	1.00	1.00	12.64	11.48	5.88	11.94
2200	20%	467.5	225.4	355.4	311.5	0.80	1.00	0.90	0.94	14.72	11.91	6.80	17.91
2200	30%	760.4	295.3	455.2	342.2	0.56	0.98	0.81	0.87	18.40	15.25	7.65	18.61
2200	40%	1101.1	346.3	634.1	444.1	0.42	0.94	0.64	0.81	27.88	16.30	8.53	22.62
2200	50%	1252.9	392.3	684.9	408.0	0.37	0.83	0.49	0.78	45.57	21.33	9.96	29.49
1700	10%	286.2	145.8	224.6	134.7	0.98	1.00	1.00	1.00	13.40	12.18	4.76	9.95
1700	20%	518.6	254.3	360.6	246.4	0.79	1.00	0.83	0.98	15.80	12.91	4.39	21.27
1700	30%	716.1	388.2	606.9	336.3	0.55	0.97	0.64	0.87	19.28	13.92	4.42	22.07
1700	40%	1101.9	418.8	708.8	401.1	0.41	0.87	0.57	0.79	28.11	20.18	8.04	22.09
1700	50%	1546.4	490.7	726.1	487.7	0.36	0.77	0.57	0.76	47.85	26.59	31.65	22.80
1200	10%	404.6	201.3	255.7	177.5	0.98	1.00	0.95	1.00	14.36	17.37	7.60	10.00
1200	20%	433.8	266.1	440.9	226.0	0.78	1.00	0.98	1.00	16.71	18.03	16.25	10.34
1200	30%	809.2	357.0	447.0	292.3	0.55	0.97	0.84	0.87	24.35	19.12	20.76	14.33
1200	40%	1212.5	439.2	639.4	434.1	0.34	0.86	0.64	0.78	38.44	19.15	19.54	18.21
1200	50%	1498.3	548.5	800.4	448.2	0.35	0.75	0.52	0.75	50.42	28.00	25.76	24.00

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Table 5.7 Panel 3

RMSE and Coverage Rate of Estimated Coefficients for Ordinal Dependent Variable (Multivariate MNAR Missing Data) – Grouping by Sample sizes

The ordered logistic model was used as the estimation method in this table.

		RMSE				CP				BIAS			
size	mis	cca	pmm	mi	ml	cca	pmm	mi	ml	cca	pmm	mi	ml
3200	10%	353.2	138.2	264.7	142.7	0.94	1.00	0.99	1.00	2.67	6.83	2.30	8.35
3200	20%	586.1	192.8	364.1	214.3	0.80	0.99	0.91	0.99	4.36	9.99	3.95	13.89
3200	30%	791.5	249.3	474.3	296.4	0.61	0.98	0.81	0.97	6.33	10.97	5.39	20.46
3200	40%	1287.9	303.7	604.7	435.0	0.45	0.96	0.54	0.96	9.93	13.25	6.00	31.85
3200	50%	1848.8	380.0	892.4	587.2	0.36	0.84	0.45	0.83	28.49	13.95	20.14	33.42
2700	10%	354.4	256.9	265.3	131.4	0.92	0.97	0.98	1.00	2.77	9.32	2.26	1.79
2700	20%	493.1	320.5	404.7	161.9	0.70	0.89	0.81	0.99	4.67	27.91	4.26	1.80
2700	30%	603.4	390.0	488.3	223.3	0.58	0.78	0.58	0.97	6.98	29.79	10.19	5.94
2700	40%	1052.4	430.8	711.3	271.2	0.40	0.75	0.54	0.95	10.44	30.38	10.32	6.60
2700	50%	1695.4	496.4	925.8	433.2	0.32	0.62	0.44	0.82	33.17	36.33	32.29	8.28
2200	10%	375.6	307.2	351.4	140.2	0.89	0.97	0.96	1.00	3.08	10.61	7.63	3.18
2200	20%	613.3	346.2	419.3	200.7	0.64	0.88	0.79	0.98	4.93	25.85	10.46	3.18
2200	30%	922.3	392.7	584.6	260.4	0.57	0.77	0.65	0.97	5.55	26.29	18.10	5.80
2200	40%	1329.7	533.9	785.5	356.7	0.38	0.74	0.53	0.87	10.60	30.84	24.01	6.68
2200	50%	2415.6	529.2	1047.4	442.4	0.30	0.73	0.41	0.81	35.23	38.96	33.53	10.14
1700	10%	272.9	181.0	217.4	163.4	0.83	0.97	0.91	1.00	3.42	11.53	8.51	3.21
1700	20%	694.0	384.0	546.5	215.2	0.60	0.87	0.77	0.98	5.03	24.59	11.74	5.19
1700	30%	911.8	422.0	574.5	284.4	0.48	0.76	0.61	0.96	6.16	27.96	19.34	6.47
1700	40%	974.2	467.1	678.5	376.8	0.36	0.72	0.50	0.79	15.78	30.05	28.92	10.85
1700	50%	2493.3	631.6	1282.5	677.9	0.28	0.71	0.41	0.75	38.82	41.67	48.17	16.54
1200	10%	395.3	263.2	283.9	196.8	0.80	0.97	0.90	0.99	4.29	12.78	9.57	3.94
1200	20%	506.0	289.7	423.5	281.8	0.59	0.85	0.74	0.97	5.79	24.84	12.91	5.23
1200	30%	886.3	387.4	583.7	344.2	0.47	0.73	0.58	0.95	12.68	28.01	21.62	9.68
1200	40%	2383.2	659.0	1263.2	740.4	0.34	0.70	0.47	0.79	21.53	30.18	29.74	14.19
1200	50%	2504.6	737.3	1289.7	748.5	0.22	0.67	0.34	0.68	40.64	42.25	48.86	18.34

Note 1: MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; RMSE = root mean square error; CP = coverage rate of the 95% confidence interval of the coefficient estimated with complete data to each estimated coefficients; BIAS = standard bias in percentage; size = raw sample size before missingness stimulation; mis = stimulated missing data rate; ccp = complete case analysis; pmm = predictive mean matching; mi = multiple imputation; ml = maximum likelihood estimation;

Note 2: RED – CCA performs the best; PURPLE – PMM outperforms MI and ML; BLUE – MI outperforms ML; SHADED – coverage rate under 90%, or standard bias with absolute value < 40.

Chapter 6:

Conclusion

6.1 Introduction

The sustainability of MFIs depends on whether they can efficiently and effectively collect their scattered loans. The issue of loan delinquency has been well recognized and is waiting to be resolved. As one of the most important determinants of loan delinquency, financial literacy has been well studied in the past decades. However, financial awareness, which is an essential component of financial literacy, was rarely mentioned in the literature. Among the several conjectures cited about the determinants of loan default, and how to develop and improve borrowers' financial awareness of interest repayment in the area of microfinance, in the empirical chapters, this thesis test 11 hypotheses using different econometric techniques. In addition, the thesis also studies 4 research questions that are associated with the wider methodological debate on the use of missing data imputation methods with a focus on administrative loan book data in the context of a normal distribution violation. The hypotheses and research questions aim at contributing to the existing academic literature and credit scoring methods related to microfinance in developing countries on the following issues:

1. Is agriculture really associated with a higher probability of loan default as most MFIs expected? Many microfinance activities in developing countries naturally focus on rural areas in which more than 75% of poor people live depending on agriculture.
2. Can previous access to credit increase the borrowers' financial awareness of their interest repayment? To this question, in fact, is trying to study whether access to credit is a virtuous cycle or a vicious cycle on itself.
3. What is the best missing data imputation technique for microfinance loan book data? It is impossible for MFIs to identify all the determinants of missing values and make accurate assumptions of the prior distribution. Answering this question helps us to find out the most robust imputation methods across different mechanisms of missingness.

The thesis focuses on the practical issues that affect the performance of credit scoring systems used by MFIs rather than the determinants of MFIs' financial performance. The following sections of this chapter summarize the results and provide practical implications for each of the three empirical chapters.

6.2 Summary of Results

6.2.1 Delinquency of microfinance

In the first empirical chapter, I have identified the individual socio-demographic and business characteristics that are associated with microfinance loans based on a high-quality administrative loan book data that stems from four MFIs from developing countries. I have replaced the omnifarious binary default indicators used in previous studies with three semi-continuous default indicators: the amount of arrears, the number of days being in delinquent, and PaR30. The explanatory variables are already known from classical banking and the prior literature of microfinance. According to the clustered structure and skewness of the data, a Two-Part model with Box-Cox transformation is applied here.

The results show that agriculture is related to a lower probability of default that measured by the amount of arrears. However, it becomes insignificant when we use the length of delayed repayment as a proxy for the probability of default. In the subsample analyses, we reconfirm that investing in agricultural business associated with a lower probability of default in INSOTEC Ecuador and FINCA Peru. In CACIL Honduras, it is found that agriculture positively relates to the intensity of loan default only. Possible explanations for the inconsistent results include: 1. borrowers involved in agricultural businesses cannot pay back the loan with a high repayment frequency; and 2. while agriculture is claimed to be the safest sector due to high social control and low volatility, it is in line with the prevailing weather conditions and indeterminate natural disasters that happen during the period of interest.

On the other hand, we also found that married borrowers have a lower probability of default and a lower intensity of delinquency, measured by both the amount of arrears and the length of delayed repayment. In the subsample analyses, borrowers have a lower probability of default in

general. This relationship is especially pronounced in MICROCRED Madagascar but is insignificant in CACIL Honduras. What is more, we found that the relation between marital status and the probability of default does not differ with genders. In addition, the relation between marriage and the probability of default is strengthened if a client has completed secondary school.

The results of gender and age are inconsistent in different MFIs, and they are insignificant in most of the subsamples. An interesting finding related to age is that the clients aged between 40 and 49 benefit the most from marriage when it comes to repayment performance.

It is surprising to find that education positively relates to the probability of default in FINCAR Peru, while the association is negative in MICROCRED Madagascar. For CACIL Honduras and INSOTEC Ecuador, no significant relations between education and loan default are detected. Possible explanations for the abnormal results for FINCA Peru include: 1. borrowers participate in business activities that require little education, but lots of experience and skills; 2. borrowers with better education are more likely to be over-indebted as they have much easier access to credit, because education is highly related to parental income and creditworthiness in Peru.

Finally, the results also point out that the estimation performances between Two-Part model and Double Hurdle model are similar, while the algorithm of the Two-Part model is more efficient. By implementing a Two-Part model in credit scoring, the MFIs would obtain better results for the probability and intensity of default with moderate time investment.

6.2.2 What drives the financial awareness in microfinance?

In the second empirical chapter, I have tested test the individual/household effects on the clients' financial awareness of their interest rate by using a large global data set covering 51 MFIs in 27 countries. Financial awareness is studied through the proxies designed by Micro Finanza Rating. They are a pair of dummies which indicate whether a client can accurately (less than 25% different from the actual values) remember his/her interest rate and total interest payment. In order to reduce the biases caused by missing values, multiple imputations and mean imputation techniques have been applied in this study.

Our findings indicate that previous access to moneylenders improved the awareness of their interest. Clients who have had saving accounts before were less knowledgeable about the interest in general. However, previous access to a saving service has a positive effect on the clients with

at least primary education. On the other hand, previous access to microfinance has a positive relation to the financial awareness of the clients who lived in urban areas.

The overall findings on the socio-demographic characteristics suggest that in our sample the association between gender and financial literacy of the interest rate only exists in Latin America and Christian countries. Women may be more financially cautious than men in these areas. The results for education background and living location are all significant. It shows that a more educated client who lives in a rural area has a much higher probability of being financially cautious. However, there are no results for age. Because the missing rate is too high, and the results with multiple imputations and the results with mean imputation are inconsistent.

6.2.3 MI, ML and PMM for semi-continuous missing data in microfinance loan book

In the final empirical study, I investigated how PMM compares to MI and ML for imputing semi-continuous data, binary data, and ordinal data. I also investigated how performance is affected by sample size and missing rate in the data, and look into the effects of the missing data mechanism on imputation methods for imputing different types of data. In addition, I investigate the aforementioned methods in the presence of univariate and multivariate missingness. The main contribution of this paper is to provide a systematic evaluation for the imputation performances of MI, ML and PMM methods with actual administrative loan book data, as there are so few performance comparison studies of different missing data techniques (MDT) available in the current literature.

To make the empirical findings of Monte Carlo studies applicable to real data, we need at least two assumptions for the sample size (or missing rate): 1. the relation between imputation quality and sample size (or missing rate) is strictly linear; 2. the sensitivities between imputation quality and sample size (or missing rate) are the same for the MDT in comparison. However, the findings in this paper suggest that these assumptions are too strong for administrative loan book data. For instance, we found that PMM usually outperforms MI and ML when the sample sizes are large, and the missing rates are low when the missing mechanism is MAR. Compared to MI and ML, PMM is more sensitive to the changing sample sizes and missing rates. It reminds us that we should not overestimate the capabilities of MDT and neglect the size effects.

Generally speaking, all MDT have comparatively the lowest biases and highest coverage rates when the missing data are ordinal categorical. Most of their biases and coverage rates have exceeded the significant criterion when the missing data are semi-continuous. On the other hand, the MDT perform better with univariate missing data than with multivariate missing data. For semi-continuous data, we also found that the sample size will affect the relationship between bias and missing rate. The biases are less sensitive to the changes of missing rates in small samples.

When the missing data are semi-continuous, PMM outperforms MI and ML in most simulations. For binary or ordinal categorical data, MI and ML are generally better than PMM. But we also notice that PMM's performance surpass MI and ML when the sample sizes are very large, the missing rates are low, and the missing mechanism is MAR.

In terms of the comparison between MI and ML, we found that MI performs better than ML when the missing data are semi-continuous, or when the missingness is MAR. Consistent with the findings in the prior literature, ML outperforms MI in small samples in general. However, it should be stressed that the differences between the biases of MI and ML are still marginal.

Finally, we found that MI, ML and PMM underperform the benchmark CCA in many simulations. In univariate missing data, CCA provides more accurate coefficient estimations in most simulations across different data types and missing mechanisms. The only exception is when the missing data are binary with MAR missingness. In multivariate missing data, MI, ML and PMM perform better than CCA in most simulations when the missing data is MAR or MNAR. But CCA is still preferable when the missing data are MCAR, and the missing rates are very low.

6.3 Implications and Recommendations

From a policy perspective, four issues are prescribed for microfinance practitioners and the governments in developing countries. The need for investing in agriculture is increasing due to the rising global population. According to an estimation by the World Bank, demand for food will increase by 70% by 2050, and more than $80 billion needs to be invested annually to resolve this issue. However, the previous market-oriented reforms in the microfinance industry seem to fail

as agriculture receives a very small share of total credit. Most MFIs limit their operations to urban areas with high densities of population. Poor farm households and farm-related business only represent a small share of their portfolios, because the MFIs consider the costs and risks encountered in serving the agricultural sector are high. In fact, our study has shown that agriculture is associated with a lower probability of default in most cases.

We recommend the MFIs to reconsider expanding into rural areas and agriculture to instead of the urban service sector. The risks of agriculture have been overestimated by them. A possible explanation for the overestimated risks is that the MFIs use a one-size-fits-all loan method requiring all clients to adapt to operational needs. It was designed and best suited for households with weekly or monthly cash inflows, but impractical for farmers. Hence, we also recommend the MFIs to design new financial products for the clients in the agricultural sector. On the other hand, another explanation to the overestimation could be that the standard of being delinquency or default is too strict for the clients with seasonal flows. In this case, we recommend the MFIs to establish a new credit scoring method based on the Two-Part model, which can separate the intensity of default from the probability of default and generate better risk evaluation.

There is a common assumption that the extremely poor remain poor due to inadequate financial management, and they need greater 'literacy' that can be improved by financial education programmes. The programmes offered by MFIs usually focus on subjects like financial knowledge, budgeting, saving, investing, financial planning, and how to choose appropriate financial instruments. While the underlying assumption of these programmes sounds right, it neglects the fact that who associated with low probabilities and low intensities of default often exhibit a financial awareness that is rarely captured in the conventional financial education. As a result, these programmes usually fail the extremely poor. Our study shows that financial awareness of borrowers can be improved by their previous access to savings or credits. Practical experience of using microfinance products is an important source of financial awareness, especially for the clients with only primary education.

Non-experimental or observational designs have played a dominant role in the past. However, most microfinance studies in recent years have switched to experimental approaches as the evaluation results heavily depend on data quality, which is a weak spot of the non-experimental studies. Data quality refers to the availability of a rich dataset of appropriate variables related to the participants of microfinance. Researchers have no control over the origination of data in the

case of non-experimental studies. Most of the time, only observable outcomes for participants was implemented. One of the biggest obstacles to acquire rich datasets is the huge amount of missing values in variables of interest. As a result, the researchers have to painfully drop the variables with missing values or delete the data points with missing values before evaluation. Our study suggests that the modern missing data imputation technique Predictive Mean Matching is a much better solution comparing to simple Listwise Deletion, Multiple Imputation, and Maximum Likelihood Estimation, in terms of semi-continuous data with lots of zeros. When the dataset is large enough, Predictive Mean Matching outperforms other methods in terms of binary or categorical data as well. It is a very effective technique to do imputation for missing data in the area of microfinance especially when the missing percentage of data is noticeable.

6.4 Limitations and Future Considerations

The first empirical chapter has the potential of unravelling more interdependence between the intensity/probability of loan default and other covariates, such as loan purposes and individual socio-demographic characteristics. A number of influential factors that mentioned in previous literature are excluded in this study, such as repayment frequency, income, and the number of household dependents. In addition, the observed positive relationship between educational background and the probability of loan default for the borrowers of FINCA Peru requires further scrutiny as it contrasts economic theory of education. This suggests a further study with a greater number and different types of MFIs in each country, such as Peru, so that we can ascertain whether the abnormal relationship between education and default is a country-specific issue or a firm-specific issue. Moreover, the comparison between the efficiencies of 2PM and DH lacks supporting evidence. The statistics of the runtime analysis for the algorithms of models are not included here.

In the case of the second empirical chapter, that examines the determinants of financial awareness of interest repayment, the issue of how to measure financial awareness is controversial. Our study proposes that financial awareness should be indicated and measured by whether a client knows his/her interest rate or interest amount. The biggest limitation of using these indicators individually is that it provides only partial information on the financial awareness of clients. Therefore, a more comprehensive measure of financial awareness is needed to address

questions that have been put forward in the growing financial literacy literature. A recent study by Kalra et al. (2015) has tried to propose a theoretical framework to fulfil the needs. However, they just equally weighted every indicator in their index and underestimated the importance of interest rate. Hence, the relation between the probability of loan default and their financial awareness index might be weak or insignificant. More empirical studies on the effectiveness of these new financial awareness indexes should be done in the future.

In terms of the final empirical chapter that compares the imputation performances between different MDTs, the first limitation is that bias is simply measured as the difference between the true correlation and the recovered correlation, as we only focus on how missing data imputation can be applied to credit scoring in microfinance. Other common imputation evaluation criterions have not been discussed in this chapter. These include: bias of the mean, the bias of the median, preservation of distributional shapes, and plausibility of the imputations (whether the imputed value could have been observed if the data was not missing). The second limitation is that we only consider a specific set of skewness, kurtosis, and point mass of a real administrative loan book data. The actual relationships between the imputation performances and they are beyond the scope of our study. Finally, we display empirical results for simulations with discretised sample size and missing rate. Hence, whether the imputation performances are strictly linearly associated to sample size and the missing rate is still unclear.

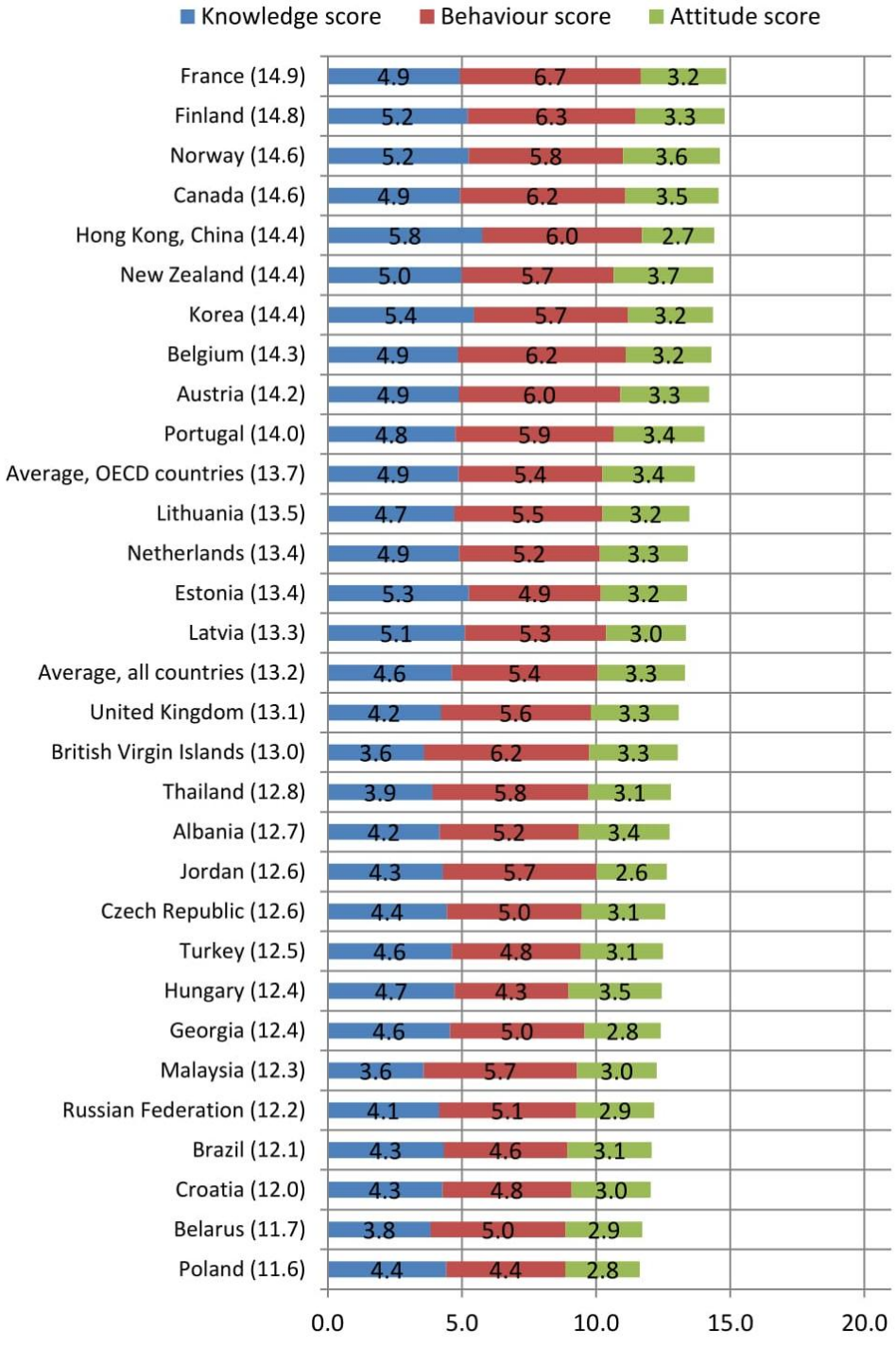
Appendices

Appendix A Top 25 Business Activities in Microcred and Finca Peru by Population

	Microcred			Finca Peru	
Business Activity	Sector	Shares	Business Activity	Sector	Shares
Grocery	Retail	12.08%	Grocery	Retail	9.70%
Other	Retail	10.85%	Clothing	Retail	7.06%
Clothing, Confection	Retail	7.03%	Foods	Services	6.93%
Other	Services	5.88%	Livestock	Retail	5.99%
Passenger transport	Services	4.69%	Cereal	Retail	5.11%
Frippery	Retail	4.38%	Wine	Retail	4.66%
Confection	Production	4.34%	Cosmetic	Retail	4.36%
Lumber, Coal	Retail	4.24%	Vegetable	Retail	3.56%
Bar	Services	3.67%	Confection	Retail	3.17%
Fruits, Vegetables, Eggs	Retail	2.39%	Fruit	Retail	2.87%
Vehicles Rental	Services	2.36%	Crafts	Production	2.55%
Other	Production	2.23%	Other	Services	1.90%
Livestock	Farming	2.19%	Cheese	Retail	1.80%
Joinery, Furniture	Production	2.08%	Bread, Pastry, Chocolate	Retail	1.77%
Haulage	Services	1.94%	Restaurante, Cafe, Bar	Services	1.79%
Butcher, Sausage	Retail	1.94%	Other	Retail	1.59%
Hairdressing, Beauty	Services	1.84%	Butcher, Sausage	Retail	1.36%
Scrap Dealer	Retail	1.63%	Beverage	Retail	1.35%
Other	Wholesale	1.40%	Potato	Farming	0.94%
Fish, Crustacean	Retail	1.38%	Clothing	Production	0.87%
Local products collection	Retail	1.36%	Shoes	Retail	0.76%
Clearance products	Retail	1.27%	Passenger transport	Services	0.71%
Fabrics	Retail	1.13%	Livestock	Farming	0.69%
Hardware, Electrical appliance	Retail	1.09%	Brick	Production	0.65%
Nursery	Services	0.99%	Rootstock	Retail	0.63%
Sample Size		18473	Sample Size		15461

Notes: For Microcred Madagascar, there are 81 types of business activities recorded in the loan book, in which the top 25 business activities take up 84.4% of the sample. For Finca Peru, there are 198 types of business activities recorded in the loan book, in which the top 25 business activities take up 72.8% of the sample.

Appendix B Financial knowledge, attitudes and behaviour (average scores); Stacked points (weighted data): all respondents, sorted by overall score



Notes: Average, all countries and Average, OECD countries report the mean of the country/economy percentages. Each country/economy is therefore given equal weight. Source: OECD/INFE International Survey of Adult Financial Literacy Competencies.

Appendix C Basic Information of the Microfinance survey database

Region	Country	Religion	Year	MFI
Africa	Ghana	Christianity	2010	ID Ghana
Africa	Kenya	Christianity	2009	KADET
Africa	Kenya	Christianity	2011	Jitigemea
Africa	Mali	Muslim	2010	CAMIDE
Africa	Mali	Muslim	2012	Nyesigiso
Africa	Niger	Muslim	2009	Taanadi
Africa	Niger	Muslim	2010	Kokari
Africa	Senegal	Muslim	2010	Microcred
Africa	Uganda	Mixture	2009	BRAC
Africa	Uganda	Mixture	2009	MCDT
Africa	Uganda	Mixture	2010	Hofokam
Africa	Uganda	Mixture	2011	Uganda Finance Trust
Europe	Bosnal Hercegovina	Mixture	2009	Sinergija
Europe	Bosnal Hercegovina	Mixture	2010	Partner
Europe	Bosnal Hercegovina	Mixture	2011	MiBospo
Europe	Romania	Christianity	2010	FAER
Latin America	Bolivia	Christianity	2008	FIE
Latin America	Bolivia	Christianity	2011	Banco FIE
Latin America	Colombia	Christianity	2009	Contactar
Latin America	Ecuador	Christianity	2008	Banco Solidario
Latin America	Ecuador	Christianity	2008	Espoir
Latin America	Ecuador	Christianity	2008	Huellas Grameen
Latin America	Ecuador	Christianity	2010	CACPE Pastaza
Latin America	Ecuador	Christianity	2010	Mushuc Runa
Latin America	Honduras	Christianity	2009	COMIXMUL
Latin America	Honduras	Christianity	2009	FUNED
Latin America	Honduras	Christianity	2010	CACIL
Latin America	Honduras	Christianity	2010	Ceibena
Latin America	Honduras	Christianity	2010	Sagrada Familia
Latin America	Mexico	Christianity	2009	Fampegro
Latin America	Mexico	Christianity	2009	Fincomun
Latin America	Mexico	Christianity	2009	Fojal
Latin America	Mexico	Christianity	2009	Progresemos
Latin America	Nicaragua	Christianity	2009	Prestanic
Latin America	Paraguay	Christianity	2010	Vision Banco
Latin America	Peru	Christianity	2010	EDPYME Proempresa
Latin America	Peru	Christianity	2010	Prisma
Latin America	Republica Dominicana	Christianity	2010	Banco ADEMI
Latin America	Republica Dominicana	Christianity	2011	ECLOF Dominicana
Middle East	Afghanistan	Muslim	2008	FMFB
Middle East	Armenia	Christianity	2011	ECLOF Armeni
Middle East	Azerbaijan	Muslim	2009	Azercredit
Middle East	Jordan	Muslim	2011	UNRWA Jordan
Middle East	Kazakistan	Muslim	2010	FFSA
Middle East	Kazakistan	Muslim	2011	Arnur Credit
Middle East	Kyrgyzstan	Muslim	2010	ABNCU
Middle East	Pakistan	Muslim	2010	ASASAH
Middle East	Palestine	Muslim	2011	ASALA
Middle East	Tajikistan	Muslim	2010	IMON
Middle East	Tajikistan	Muslim	2010	OXUS
Southeast Asia	Philippines	Christianity	2009	ASKI

References

Agier, I., & Szafarz, A. (2013). Microfinance and gender: Is there a glass ceiling on loan size?. World Development, 42, 165-181.

Agnew, J.R., Bateman, H. and Thorp, S. (2012). Financial Literacy and Retirement Planning in Australian. UNSW Australian School of Business Research Paper, (2012ACTL16).

Ahmad, S. A. (1997). Natural hazards and hazard management in the greater Caribbean and Latin America. University of West Indies, Kingston, Jamaica: The University Printers.

Alesina, A.F., Lotti, F., and Mistrulli, P.E. (2008). Do women pay more for credit? Evidence from Italy. National Bureau of Economic Research.

Allgood, S. and Walstad, W. (2013). Financial literacy and credit card behaviours: A cross-sectional analysis by age. Numeracy, 6(2), 3.

Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. British Journal of Mathematical and Statistical Psychology, 55(1), 193-196.

Alvarez, R., & Crespi, G. (2003). Determinants of technical efficiency in small firms. Small business economics, 20(3), 233-244.

Ameriks, J., Caplin, A. and Leahy, J. (2002). Wealth accumulation and the propensity to plan (No. w8920). National Bureau of Economic Research.

Ameyaw-Amankwah, I. (2011). Causes and effects of loan defaults on the profitability of Okomfo Anokye Rural Bank. Retrieved on January, 5, 2015.

Amin, S., Rai, A. S. & Topa, G. (2003). Does microcredit reach the poor and vulnerable? Evidence from northern Bangladesh. Journal of Development Economics, 70(1), 59-82.

Armendariz, A., & Morduch, J. (2010). The economics of microfinance (2nd edition).

Arminger, G., Enache, D., & Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks. Computational Statistics, 12(2), March 26, 1997.

Arrondel, L., Debbich, M. and Savignac, F. (2013). Financial literacy and financial planning in France. Numeracy, 6(2), 8.

Atkinson, A., & Messy, F. A. (2016). OECD/INFE International Survey of Adult Financial Literacy Competencies. Paris: OECD. https://www.oecd.org/finance/oecd-financialliteracy-study-finds-many-adults-strugglewith-money-matters.htm.

Augsburg, B. (2006). Econometric evaluation of the SEWA bank in India: applying matching techniques based on the propensity score. MGSoG Working Paper No. 003.

Avery, R. B., Calem, P. S., & Canner, G. B. (2004). Consumer credit scoring: do situational circumstances matter? Journal of Banking & Finance, 28(4), 835-856.

Baesens, B., W. Verbeke, P. Sercu and J. V. Gool (2011). Credit Scoring for Microfinance: Is It Worth It? International Journal of Finance and Economics, 17(2), 103–23.

Balogun, E. D., & Alimi, A. (1988). Loan delinquency among small farmers in developing countries: A case study of the Small–Farmer Credit Programme in Lagos State of Nigeria. CBN Economic and Financial Review, 26(3).

Bandyopadhyay, A., & Saha, A. (2011). Distinctive demand and risk characteristics of residential housing loan market in India. Journal of Economic Studies, 38(6), 703-724.

Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. American Economic Journal: Applied Economics, 7(1), 22-53.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of school psychology, 48(1), 5-37.

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), 1133-1157.

Bauman, K. J. (1999). Shifting family definitions: The effect of cohabitation and other nonfamily household relationships on measures of poverty. Demography, 36(3), 315-325.

Becker, G. S. (1973). A theory of marriage: Part I. Journal of Political economy, 81(4), 813-846.

Becker, G. S. (1974). A theory of marriage: Part II. Journal of Political Economy, 82(2, Part 2), S11-S26.

Beisland, L. A., & Mersland, R. (2012). The use of microfinance services among economically active disabled people: Evidence from Uganda. Journal of International Development, 24(S1).

Bellucci, A., Borisov, A., & Zazzaro, A. (2010). Does gender matter in bank-firm relationships? Evidence from small business lending. Journal of Banking & Finance, 34(12), 2968-2984.

Berger, A. N., & DeYoung, R. (1997). Problem loans and cost efficiency in commercial banks. Journal of Banking & Finance, 21(6), 849-870.

Bhatt, N. & S. Y. Tang (2002). Determinants of Repayment in Microcredit: Evidence from Programmes in the United States. International Journal of Urban and Regional Research, 26(6), 360–76.

Bloem, A. M., & Goerter, C. N. (2001). The macroeconomic statistical treatment of non-performing loans. In Statistics Department of the International Monetary Fund Discussion Paper.

Blundell, R., & Dias, M. C. (2009). Alternative approaches to evaluation in empirical microeconomics. Journal of Human Resources, 44(3), 565-640.

Bodner, T.E. (2008). What improves with increased missing data imputations? Structural Equation Modeling, 15(4), 651-675.

Brown, M. & Graf, R. (2013). Financial literacy and retirement planning in Switzer-land. Numeracy, 6(2), 6.

Brown, S. L., & Booth, A. (1996). Cohabitation versus marriage: A comparison of relationship quality. Journal of Marriage and the Family (1996), 668-678.

Bucher-Koenen, T. & Lusardi, A. (2011). Financial literacy and retirement planning in Germany. Journal of Pension Economics and Finance, 10(4), 565-584.

Bucher-Koenen, T., Lusardi, A., Alessie, R. & van Rooij, M. (2012). How financially literate are women? Some new perspectives on the gender gap. Netspar Panel Paper, 31.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3), 67.

Caliendo, M. (2006). Microeconometric evaluation of labour market policies (Vol. 568). Springer Science & Business Media.

Calvet, L. E., Campbell, J. Y., & Sodini, P. (2007). Down or out: Assessing the welfare costs of household investment mistakes. Journal of Political Economy, 115(5), 707-747.

Cameron, A. C., & Trivedi, P. K. (2005). Micro-econometrics: methods and applications. Cambridge university press.

Campbell, J. (2006). Household finance. Journal of Finance, 61, 1553-1604.

Carpena, F., Cole, S.A., Shapiro, J. & Zia, B. (2011). Unpacking the causal chain of financial literacy. World Bank Policy Research Working Paper Series, No.5798.

Chemin, M. (2008). The benefits and costs of microfinance: evidence from Bangladesh. The journal of development studies, 44(4), 463-484.

Chen, H. & Volpe, R.P. (2002). Gender differences in personal financial literacy among college students. Financial services review, 11(3), p.289.

Chen, M. A., & Snodgrass, D. R. (1999). An assessment of the impact of SEWA Bank in India: baseline findings. AIMS, for the Office of Microenterprise Development, USAID. http://pdf.usaid.gov/pdf_docs/PNACG038.pdf.

Chen, M. A., & Snodgrass, D. (2001). Managing resources, activities, and risk in urban India: The impact of SEWA Bank. Washington, DC: AIMS. http://pdf.usaid.gov/pdf_docs/Pnacn571.pdf.

Chossudovsky, M. (1998). Global Poverty in the late 20th Century. Journal of International Affairs, 293-311.

Clarkberg, M. (1999). The price of partnering: The role of economic well-being in young adults' first union experiences. Social Forces, 945-968.

Cocco, J.F., Gomes, F.J. & Maenhout, P.J. (2005). Consumption and portfolio choice over the life cycle. Review of Financial Studies, 18(2), 491-533.

Cochrane Collaboration. (2011). Cochrane handbook for systematic reviews of interventions 5.1. 0. Cochrane Collaboration. http://handbook-5-1.cochrane.org/front_page.htm

Coleman, B. E. (1999). The impact of group lending in Northeast Thailand. Journal of Development Economics, 60(1), 105-141.

Coleman, B. E. (2006). Microfinance in Northeast Thailand: who benefits and how much? World Development, 34(9), 1612-1638.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330.

Cook, T. D., Shadish, & W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of policy analysis and management, 27(4), 724-750.

Copestake, J. (2002). Inequality and the polarizing impact of microcredit: evidence from Zambia's Copperbelt. Journal of international development, 14(6), 743-755.

Copestake, J., Bhalotra, S., & Johnson, S. (2001). Assessing the impact of microcredit: A Zambian case study. Journal of Development Studies, 37(4), 81-100.

Copestake, J., Dawson, P., Fanning, J. P., McKay, A., & Wrigh-Revolledo, K. (2005). Monitoring the diversity of the poverty outreach and impact of microfinance: A comparison of methods using data from Peru. Development Policy Review, 23(6), 703-723.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica: Journal of the Econometric Society (1971), 829-844.

Crook, J., & Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? Journal of Banking & Finance, 28(4), 857-874.

Croson, R. & U. Gneezy (2009). Gender Differences in Preferences. Journal of Economic Literature, 47(2), 448–74.

D'agostino, R. B., Belanger, A., & D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. The American Statistician, 44(4), 316-321.

Deaton, A. (2010). Instruments, randomisation, and learning about development. Journal of economic literature, 48(2), 424-455.

Deevy, M., Lucich, S., & Beals, M. (2012). Scams, schemes, and swindles: A review of consumer financial fraud research. Available at: fraudresearchcenter. org.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for non-experimental causal studies. Review of Economics and statistics, 84(1), 151-161.

Dehejia, R., Montgomery, H., & Morduch, J. (2012). Do interest rates matter? Credit demand in the Dhaka slums. Journal of Development Economics, 97(2), 437-449.

Deininger, K., & Liu, Y. (2013). Economic and social impacts of an innovative self-help group model in India. World Development, 43, 149-163.

Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. Journal of Statistical Computation and Simulation, 78(1), 69-84.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.

Dinh, T. H. T., & Kleimeier, S. (2007). A credit scoring model for Vietnam's retail banking market. International Review of Financial Analysis, 16(5), 471-495.

Dunn, L. F., & Kim, T. (1999). An empirical investigation of credit card default. Ohio State University, Department of Economics Working Papers, (99-13).

Duvendack, M. (2010). Smoke and mirrors: evidence from microfinance impact evaluations in India and Bangladesh. Doctoral dissertation, University of East Anglia.

Duvendack, M., & Palmer-Jones, R. (2012). High noon for microfinance impact evaluations: reinvestigating the evidence from Bangladesh. The Journal of Development Studies, 48(12), 1864-1880.

Duvendack, M., Palmer-Jones, R., Copestake, J. G., Hooper, L., Loke, Y., & Rao, N. (2011). What is the evidence of the impact of microfinance on the well-being of poor people? EPPI Centre. https://assets.publishing.service.gov.uk/media/57a08aeeed915d622c0009bb/Micro-finance2011Duvendackreport.pdf.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. Structural Equation Modeling, 8(1), 128-141.

Fernandes, D., Lynch Jr, J.G. and Netemeyer, R.G. (2014). Financial literacy, financial education, and downstream financial behaviours. Management Science, 60(8), 1861-1883.

Fernando, J. L. (1997) Nongovernmental organisations, microcredit, and empowerment of women. The Annals of the American Academy of Political and Social Science, 554(1), 150-177.

FFIEC (2016). Charge-Off and Delinquency Rates on Loans and Leases at Commercial Banks. https://www.federalreserve.gov/releases/chargeoff/delallsa.htm#fn2

Fidrmuc, J., & Hainz, C. (2010). Default rates in the loan market for SMEs: Evidence from Slovakia. Economic Systems, 34(2), 133-147.

Finke, M.S., Howe, J.S. & Huston, S.J. (2016). Old age and the decline in financial literacy. Forthcoming in Management Science.

Fonseca, R., Mullen, K.J., Zamarro, G. & Zissimopoulos, J. (2012). What explains the gender gap in financial literacy? The role of household decision making. Journal of Consumer Affairs, 46(1), 90-106.

Food and Agriculture Organisation of the United Nations. (2005). Gender and farming systems: Lessons from Nicaragua. http://www.fao.org/docrep/008/y4936e/y4936e00.htm#Contents

Gaile, G. L., & Foster, J. (1996). Review of Methodological Approaches to the Study of the Impact of the Microenterprise Credit Programmes. Assessing the Impact of Microenterprise Services (AIMS). http://pdf.usaid.gov/pdf_docs/PNABZ073.PDF.

Gallardo, J. (1997). Leasing to support small businesses and microenterprises. Policy Research Working Paper No. 1857. The World Bank.

García, J., & Labeaga, J. M. (1996). Alternative approaches to modelling zero expenditure: an application to Spanish demand for tobacco. Oxford Bulletin of Economics and statistics, 58(3), 489-506.

Gelman, A., & Hill, J. (2007). Data analysis using regression and multi-level-hierarchical models (Vol. 1). New York, NY, USA: Cambridge University Press.

Gerardi, K., Goette, L. & Meier, S. (2013). Numerical ability predicts mortgage default. Proceedings of the National Academy of Sciences, 110(28), 11267-11271.

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

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

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. Journal of Applied Psychology, 78(1), 119.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. Statistical strategies for small sample research, 50, 1-27.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. The science of prevention: Methodological advances from alcohol and substance abuse research, 1, 325-366.

Graham, J.W., Olchowski, A.E. & Gilreath, T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science, 8(3), 206-213.

Guiso, L., & Jappelli, T. (2005). Awareness and stock market participation. Review of Finance, 9(4), 537-567.

Haddad L (2011) A new harvest of RCTs? Blog post on Development horizons, 15 May. www.developmenthorizons.com/2011/05/new-harvest-of-rcts.html

Hancock, J. (2002). Eighth defined contribution plan survey. John Hancock Financial Services, Boston: John Hancock.

Hand, D., & Henley, W. (1993). Can reject inference ever work? IMA Journal of Management Mathematics, 5(1), 45-55.

Hashemi, S. M., Schuler, S. R., & Riley, A. P. (1996). Rural credit programmes and women's empowerment in Bangladesh. World development, 24(4), 635-653.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programmes, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programmes, and to forecast their effects in new environments. Handbook of Econometrics, 6, 4875-5143.

Hsu, J. W. (2016). Aging and strategic learning: The impact of spousal incentives on financial literacy. Journal of Human Resources, 51(4), 1036-1067.

Hulme, D., Moore, K., & Barrientos, A. (2009). Assessing the insurance role of microsavings. DESA Working Paper, No. 83. http://www.un.org/esa/desa/papers/2009/wp83_2009.pdf.

Hung, A., Parker, A.M. & Yoong, J. (2009). Defining and measuring financial literacy. RAND Working Paper Series WR-708.

IADB (2009). Honduras obtains IDB assistance for disaster risk management. http://www.iadb.org/en/news/news-releases/2009-06-25/honduras-obtains-idb-assistance-for-disaster-risk-management,5488.html.

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

INFE, O. (2011). Measuring Financial Literacy: Questionnaire and Guidance Notes for conducting an Internationally Comparable Survey of Financial Literacy. https://www.oecd.org/finance/financial-education/49319977.pdf.

Javaras, K. N., & Van Dyk, D. A. (2003). Multiple imputation for incomplete data with semi-continuous variables. Journal of the American Statistical Association, 98(463), 703-715.

Johnston, D. & Morduch, J. (2008). The unbanked: evidence from Indonesia. The World Bank Economic Review, 22(3), 517-537.

Jones, A. M. (1989). A double-hurdle model of cigarette consumption. Journal of applied econometrics, 4(1), 23-39.

Kalra, V., Mathur, H. P., & Rajeev, P. V. (2015). Microfinance clients' awareness index: A measure of awareness and skills of microfinance clients. IIMB Management Review, 27(4), 252-266.

Kantis, H. (Ed.). (2004). Desarrollo emprendedor: América Latina y la experiencia internacional. IDB.

Karlan, D. S. & Zinman, J. (2009). Expanding microenterprise credit access: Using randomized supply decisions to estimate the impacts in Manila (No. 976). Center Discussion Paper, Economic Growth Center.

Karlan, D. S. (2007). Social connections and group banking. The Economic Journal, 117(517), F52-F84.

Karlan, D. S., & Appel, J. (2011). More than good intentions. New York: Dutton.

Karlan, D., & Zinman, J. (2010). Expanding credit access: Using randomized supply decisions to estimate the impacts. Review of Financial Studies, 23(1), 433-464.

Karlan, D., Ratan, A.L. & Zinman, J. (2014). Savings by and for the Poor: A Research Review and Agenda. Review of Income and Wealth, 60(1), 36-78.

Kenney, C., & McLanahan, S. (2001). Are cohabiting relationships more violent than marriage? Unpublished manuscript, Princeton University: Office of Population Research.

Kenward, M. G., & Molenberghs, G. (1998). Likelihood-based frequentist inference when data are missing at random. Statistical Science, 236-247.

Khandker, S.R. (2005). Microfinance and poverty: Evidence using panel data from Bangladesh. World Bank Economic Review, 19(2), 263-286.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American political science review, 95(1), 49-69.

Klapper, L. & Panos, G.A. (2011). Financial literacy and retirement planning: the Russian case. Journal of Pension Economics and Finance, 10(4), 599-618.

Klapper, L., Lusardi, A., & Van Oudheusden, P. (2015). Financial Literacy around the World. Standard & Poor's Ratings Services Global Financial Literacy Survey., Access mode: http://media.mhfi.com/documents/2015-Finlit_paper_17_ F3_SINGLES.pdf.

Kocenda, E., & Vojtek, M. (2009). Default predictors and credit scoring models for retail banking.

Kohansal, M. R., & Mansoori, H. (2009, October). Factors affecting on loan repayment performance of farmers in Khorasan-Razavi Province of Iran. In Conference on International Research on Food Security, Natural Resource Management and Rural Development, University of Hamburg, 1-4.

Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. Political Analysis, 22(4), 497-519.

Kunz, R., Vist, G., & Oxman, A. D. (2007). Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev, 2(2).

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American journal of epidemiology, 171(5), 624-632.

Lensink, R., & Pham, T.T.T. (2012). The impact of microcredit on self-employment profits in Vietnam. Economics of Transition.

Little, R. J. (1988). Missing-data adjustments in large surveys. Journal of Business & Economic Statistics, 6(3), 287-296.

Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley & Sons. 87.

Lusardi, A. & Mitchell, O.S. (2007). Baby boomer retirement security: The roles of planning, financial literacy, and housing wealth. Journal of Monetary Economics, 54(1), 205-224.

Lusardi, A. & Mitchell, O.S. (2008). Planning and financial literacy: How do women fare? (No. w13750). National Bureau of Economic Research.

Lusardi, A. & Mitchell, O.S. (2009). How ordinary consumers make complex economic decisions: Financial literacy and retirement readiness (No. w15350). National Bureau of Economic Research.

Lusardi, A. & Mitchell, O.S. (2011a). Financial literacy and planning: Implications for retirement wellbeing (No. w17078). National Bureau of Economic Research.

Lusardi, A. & Mitchell, O.S. (2011b). Financial literacy around the world: an overview. Journal of Pension Economics and Finance, 10(4), 497-508.

Lusardi, A. & Tufano, P. (2009a). Debt literacy, financial experiences, and over-indebtedness (No. w14808). National Bureau of Economic Research.

Lusardi, A. & Tufano, P. (2009b). Teach workers about the perils of debt. Harvard Business Review, 87(11), 22-24.

Lusardi, A., Mitchell, O.S. & Curto, V. (2009). Financial literacy among the young: Evidence and implications for consumer policy (No. w15352). National Bureau of Economic Research.

Maddala, G.S.9 (1991). A perspective on the use of limited-dependent and qualitative variables models in accounting research. The Accounting Review, 66(4), 788-807.

Madden, D. (2008). Sample selection versus two-part models revisited: The case of female smoking and drinking. Journal of health economics, 27(2), 300-307.

Mahdavi, M. & Horton, N.J. (2014). Financial knowledge among educated women: room for improvement. Journal of Consumer Affairs, 48(2), 403-417.

Mandell, L. (2008). Financial education in high school. Overcoming the saving slump: How to increase the effectiveness of financial education and saving programmes, 257-279.

Manning, W. D., & Lichter, D. T. (1996). Parental cohabitation and children's economic well-being. Journal of Marriage and the Family, 998-1010.

Markow, D. (2005). What American Teens and Adults Know About Economics? Prepared for The National Council on Economic Education. Harris Interactive.

Mayoux, L. (1997). Impact assessment and women's empowerment in microfinance programmes: issues for a participatory action and learning approach. Paper presented at the Background Paper for CGAP Meeting.

McArdle, J.J., Smith, J.P. & Willis, R. (2009). Cognition and economic outcomes in the Health and Retirement Survey (No. w15266). National Bureau of Economic Research.

McIntosh, C., Villaran, G., & Wydick, B. (2011). Microfinance and home improvement: using retrospective panel data to measure programme effects on fundamental events. World Development, 39(6), 922-937.

Meier, S. & Sprenger, C.D. (2013). Discounting financial literacy: Time preferences and participation in financial education programmes. Journal of Economic Behaviour & Organisation, 95, 159-174.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105(1), 156.

Moffatt, P. G. (2005). Hurdle models of loan default. Journal of the operational research society, 56(9), 1063-1071.

Moore, D.L. (2003). Survey of financial literacy in Washington State: Knowledge, behaviour, attitudes, and experiences. Washington State Department of Financial Institutions. Social and Economic Sciences Research Center Technical report, 03-39.

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

Morduch, J. (1999). The microfinance promise. Journal of economic literature, 37(4), 1569-1614.

Mottola, G.R. (2013). In our best interest: Women, financial literacy, and credit card behaviour. Numeracy, 6(2), 4.

Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. Journal of health economics, 17(3), 247-281.

Munene, H. N., & Guyo, S. H. (2013). Factors influencing Loan repayment Default in microfinance institutions: The experience of Imenti North District, Kenya. International Journal of Applied science & Technology, 3.

Murray, J. (2011). Default on a loan, United States Business Law and Taxes Guide National Credit Act (2005). Act No. 34 of 2005, Republic of South Africa, 489, No.28619.

Navajas, S., Schreiner, M., Meyer, R.L., Gonzalez-Vega, C., & Rodriguez-Meza, J. (2000). Microcredit and the Poorest of the Poor: Theory and Evidence from Bolivia. World development, 28(2), 333-346.

Odell, K. (2010). Measuring the impact of microfinance. Grameen Foundation, Washington, 1-38.

OECD (2010). Latin American Economic Outlook 2011: How Middle-Class is Latin America? OECD Development Centre, 131-133.

Okorie, A. (1986). Major determinants of agricultural smallholder loan repayment in a developing economy: Empirical evidence from Ondo State, Nigeria. Agricultural Administration, 21(4), 223-234.

Okunmadewa, F. (1998, June). Domestic and international response to poverty alleviation in Nigeria. In Proceedings of the 7th Annual Conference of the Zonal Research Units Organized by Research Department, CBN at Makurdi, 6(1), 296-309.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semi-continuous longitudinal data. Journal of the American Statistical Association, 96(454), 730-745.

Orso, C. E. (2011). Microcredit and poverty. An overview of the principal statistical methods used to measure the programme net impacts. POLIS Working Paper No. 180.

Park, A., & Ren, C. (2001). Microfinance with Chinese characteristics. World Development, 29(1), 39-62.

Pearson Jr, R. V., Fund, R. H. L., & Bureau, T. C. (2006). Causes of default among housing micro loan clients. FinMark Trust Rural Housing Loan Fund, National Housing Finance Corporation and Development Bank of Southern Africa, South Africa.

Pitt, M. & S. R. Khandker (1998). The Impact of Group-Based Credit Programmes on Poor Households in Bangladesh: Does the Gender of Participants Matter? Journal of Political Economy, 106(5), 958-996.

Pitt, M. M. (2011). Response to Roodman and Morduch's 'The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence'. Brown University.

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

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

Pitt, M. M., Khandker, S. R., & Cartwright, J. (2006). Empowering women with microfinance: evidence from Bangladesh. Economic Development and Cultural Change, 54(4), 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), 87-118.

Pudney, S., & Woodbridge, G. (1994). Modelling Individual Choice: The Econometrics of Corners, Kinks and Holes. Economic Record, 70(208), 105-105.

Raghunathan, T. E., Solenberger, P. W., & Van Hoewyk, J. (2002). IVEware: imputation and variance estimation software. Ann Arbor, MI: Survey Methodology Programme, Survey Research Center, Institute for Social Research, University of Michigan.

Reinke, J. (1998). How to lend like mad and make a profit: A micro - credit paradigm versus the start - up fund in South Africa. The Journal of Development Studies, 34(3), 44-61.

RJa, L. and Rubin, D.B. (1987). Statistical analysis with missing data.

Rogers, P. J. (2010). Learning from the evidence about evidence-based policy. Strengthening evidence-based policy in the Australian Federation, 195-208.

Roodman, D., & Morduch, J. (2014). The impact of microcredit on the poor in Bangladesh: Revisiting the evidence. Journal of Development Studies, 50(4), 583-604.

Roslan, A. H., & Karim, M. A. (2009). Determinants of microcredit repayment in Malaysia: The case of Agrobank. Humanity & Social Sciences Journal, 4(1), 45-52.

Royston, P. (2005). Multiple imputation of missing values: update of ice. Stata Journal, 5(4), 527.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics, 4(1), 87-94.

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Salazar, G. L. (2008). An Analysis of Repayment among Clients of the Microfinance Institution Esperanza International, Dominican Republic. Journal of Agricultural Economics. 90(5), 1366–91.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. CRC press. Section 6.4.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological Methods, 7(2), 147.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate behavioural research, 33(4), 545-571.

Schafer, J. L., & Olsen, M. K. (1999). Modeling and imputation of semi-continuous survey variables. In Proceedings of the Federal Committee on Statistical Methodology Research Conference (pp. 565-74). Research, 16(3), 243-258.

Schafer, J.L. & Graham, J.W. (2002). Missing data: our view of the state of the art. Psychological Methods, 7(2), p.147.

Schreiner, M. (2004). Scoring arrears at a microlender in Bolivia. ESR Review, 6(2), 65.

Schuler, S. R., Hashemi, S. M., & Riley, A. P. (1997). The influence of women's changing roles and status in Bangladesh's fertility transition: evidence from a study of credit programmes and contraceptive use. World Development, 25(4), 563-575.

Sebstad, J., & Chen, G. (1996). An overview of the studies on the impact of microenterprise credit. Assessing the Impact of Microenterprise Services (AIMS). http://pdf.usaid.gov/pdf_docs/Pnabz074.pdf.

Setboonsarng, S., & Parpiev, Z. (2008). Microfinance and the Millennium Development Goals in Pakistan: Impact assessment using propensity score matching (No. 104). ADB Institute Discussion Papers.

Sharma, M., & Zeller, M. (1997). Repayment performance in group-based credit programmes in Bangladesh: An empirical analysis. World development, 25(10), 1731-1742.

Sheila, A. L. (2011). Lending Methodologies and loan losses and default in a Microfinance deposit-taking institutions in Uganda. A case study of Finca Uganda Kabala Branch (MDI). Unpublished research paper. Makerere University.

Smith, J. A. & E Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 125(1), 305-353.

Smith, M. D. (2002). On specifying double-hurdle models. Handbook of Applied Econometrics and Statistical Inference, 535-552.

Steele, F., Amin, S., & Naved, R. T. (2001). Savings/credit group formation and change in contraception. Demography, 38(2), 267-282.

Stewart, R., Van Rooyen, C., Korth, M., Chereni, A., Rebelo, N. D. S., & De Wet, T. (2012). Systematic Review: Do Micro-Credit, Micro-Savings and Micro-Leasing Serve as Effective Financial Inclusion Interventions Enabling Poor People, and Especially Women, to Engage in Meaningful Economic Opportunities in Low-and Middle-Income Countries. EPPI Centre.

Storey, D. J. (2004) Racial and gender discrimination in the micro firms credit market? Evidence from Trinidad and Tobago. Small Business Economics, 23(5), 401-422.

Tauras, J., Powell, L., Chaloupka, F., & Ross, H. (2007). The demand for smokeless tobacco among male high school students in the United States: the impact of taxes, prices and policies. Applied Economics, 39(1), 31-41.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica: Journal of the Econometric Society, 24-36.

Ugbomeh, G. M., Achoja, F. O., Ideh, V., & Ofuoku, A. U. (2008). Determinants of loan repayment performance among women self-help groups in Bayelsa State, Nigeria. Agriculture Conspectus Scientificus (ACS), 73(3), 189-195.

Utkus, S.P. and Young, J.A. (2010). Financial Literacy and 401 (k) Loans. Pension Research Council Working Paper No. 2010-28.

Van Bastelaer, T., & Leathers, H. (2006). Trust in lending: Social capital and joint liability seed loans in Southern Zambia. World Development, 34(10), 1788-1807.

Van Gool, J., Verbeke, W., Sercu, P., & Baesens, B. (2012). Credit scoring for microfinance: is it worth it? International Journal of Finance & Economics, 17(2), 103-123.

Van Rooij, M.C., Lusardi, A. and Alessie, R.J. (2012). Financial literacy, retirement planning and household wealth*. The Economic Journal, 122(560), 449-478.

Venkata, N. A., & Yamini, V. (2010). Why do microfinance clients take multiple loans. MicroSave India Focus Note, 33.

Verbeek, M. (2008). A guide to modern econometrics. John Wiley & Sons.

Vigano, L. (1993). A Credit-scoring Model for Development Banks: An African Case Study. Savings and Development, 17(4) 441-482.

Vink, G., Frank, L. E., Pannekoek, J., & Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica, 68(1), 61-90.

Vogelgesang, U. (2003). Microfinance in times of crisis: The effects of competition, rising indebtedness, and economic crisis on repayment behaviour. World Development, 31(12), 2085-2114.

Von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? Sociological Methods & Research, 42(1), 105-138.

Wainer, H. (1976). Robust statistics: A survey and some prescriptions. Journal of Educational Statistics, 1(4), 285-312.

Warue, B.N., (2012). Factors Affecting Loan Delinquency in Microfinance Institutions in Kenya. International Journal of Management Sciences and Business Research, 1(12).

Weber, R., & Musshoff, O. (2012). Is agricultural microcredit really more risky? Evidence from Tanzania. Agricultural Finance Review, 72(3), 416-435.

Weber, R., & Musshoff, O. (2013). Can flexible microfinance loans improve credit access for farmers? Agricultural finance review, 73(2), 255-271.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine, 30(4), 377-399.

Yu, L. M., Burton, A., & Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. Statistical Methods in Medical Research, 16(3), 243-258.

Yuan, K. H. (2009). Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. Journal of Multivariate Analysis, 100(9), 1900-1918.

Yuan, K. H., Lambert, P. L., & Fouladi, R. T. (2004). Mardia's multivariate kurtosis with missing data. Multivariate Behavioural Research, 39(3), 413-437.