Impact of microcredit on poverty in Eastern Indonesia: dealing with selection bias

Joeri Smits and Themba Mthembu

NORWEGIAN UNIVERSITY OF LIFE SCIENCES
Department of Economics and Resource Management (IØR)
Master Thesis 30 credits 2012

Abstract

The popularity of and increasing controversy around microcredit has given rise to the need for rigorous evaluation of its welfare impact. We collected pre- (n=299) and post-treatment (n=209) survey data to determine the impacts of a group loan with individual liability on indicators of household welfare in NTT province, Indonesia. Because of the lack of statistical significance of impact estimates due to the short follow-up period of only one year, we focus disproportionately on methodological issues. The main challenge of an evaluation of a non-randomized program, as in our setting, is endogenous treatment selection. We focus on nonparametric methods when dealing with attrition and selection bias. We propose a novel nonparametric test of instrument validity in a general recursive model. In contrast to existing overidentifying restriction tests for linear IV models such as the well-known Sargan test, our test is consistent even when none of the instruments to be tested is valid and can thus be applied when only one instrumental variable is available. This and other tests failed to refute the validity of our new instrument, a dummy indicating whether the household has at least one treated or previously treated acquaintance, when a wealth index was the outcome. Our main finding is that take-up of TLM's group loan has a negative short-term impact on household wealth, indicating loan-induced distress sales of assets.

Key words: *microcredit, non-randomized evaluation, self-selection, attrition, overidentifying restriction test, instrument validity, Indonesia*.

Acknowledgements

We thank Prof. Arild Angelsen for his very helpful comments on our drafts, his accessibility and his humor. We thank our host institution, the UMB School of Economics and Business, for its financial support and the Norwegian government for making university education free. May others learn from you. Thanks to the Indonesian Ministry of Research and Technology (RISTEK) for granting me a permit to do research in Indonesia.

We are indebted to Pak Heri for his hard work and sacrifices. Thank you for accustoming me to the culture of NTT, for your willingness to take risks for the purpose of good research and for your commitment throughout. Makasi banyak!

Thanks to TLM for their cooperation with this research. We truly hope that it will lead to improvements in welfare. Thanks to the staff of TLM for welcoming and helping me in this effort. I especially would like to express thanks to pak Rozali, Shanty, Vience, Vicky, Jerry and Daniel.

Thanks to Helen Keller for lending the anthropometric measurement instruments. We express our gratitude to the help and suggestions received from the following persons: Prof. Jeffrey Racine, Prof. Stein Holden, Prof. Markus Frolich, Dr. Marrit van den Berg, Dr. Martin Huber, Prof. Daniel Millimet, Prof. Cristopher Woodruff and Selim Gulesci.

Thanks also to friends and family for their helpful suggestions, including but not limited to Abel Djojo, Achamyeleh Tamiru Ewunetu, Daniel Muluwork, Tadesse Mezgebo, Pablo Rouwet, Asif Iqbal, and my love Qiongxian.

My parents, thank you for your continued support in all forms, in the spirit of Michiel. Terima kasih banyak pak Patris, Mama dan keluarga untuk kemungkinan ini. Terima kasih untuk tinggal di Kupang selama masa yang saya melakukan survei di Timor. Terima kasih atas makan yang enak terus dan kermahan kalian. Terima kasih banyak Pak Darius dan keluarga untuk tinggal di Kalabahi dan pakai perahu. Thanks to Themba's parent's-in-law for your hospitality and Lene and Pinky for the delicious food.

Ma Pinky thanks for all your support every step of the way. Lene, my lovely and supportive wife so much is packed into the next two words, thank you. Thandi my inspiration, this is for you. Love you guys.

List of acronyms

ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
BPS	Badan Pusat Statistik (Indonesia's statistical office)
CATPCA	Categorical Principal Component Analysis
GDP	Gross Domestic Product
IV	Instrumental Variable
LATE	Local Average Treatment Effect
LIML	Limited Information Maximum Likelihood
MAPE	Mean Absolute Percentage Error
MAR	Missing At Random
MCAR	Missing Completely At Random
MIV	Monotone Instrumental Variable
MNAR	Missing Not At Random
NGO	Non-Governmental Organization
NTT	Nusa Tenggara Timor (East Nusa Tenggara province)
OECD	Organization for Economic Co-operation and Development
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
QTE	Quantile Treatment Effect
RCT	Randomized Controlled Trial
RT	Rukun Tetangga (smallest administrative unit in Indonesia)
SATE	Sample Average Treatment Effect
SATT	Sample Average Treatment Effect on the Treated
SLATE	Sample Local Average Treatment Effect
SQTE	Sample Quantile Treatment Effect
SUTVA	Stable Unit Treatment Value Assumption
TLM	Tanaoba Lais Manekat
TTS	Timor Tenggah Selatan (South-Central Timor regency)
WHO	World Health Organization

Table of Contents

List of tables

List of figures

1. Introduction

1.1. Policy issue

Microcredit, defined here as the provision of small loans to poor people, has become one of the most popular development interventions as to date. The promise of alleviating, reducing or even eliminating poverty by supporting the poor in building and expanding their microenterprises received support from across the political spectrum. The movement gained momentum in the 1990s and culminated in the Nobel Peace Prize for pioneer Mohammed Yunus and his Bangladeshi Grameen Bank in 2006 *"for their efforts to create economic and social development from below"* (Nobelprize.org 2012). To some observers, these microfinance institutions do not differ much from the old moneylenders with usurious interest rates and an over-indebted clientele (f.i. Bateman 2010). Given (a) that the global number of borrowers reached around 150 million in 2009 (Daley-Harris 2009), (b) the scrutiny aid programmes in many OECD countries receive in a period of fiscal tightening and (c) the highly context-specificity of impacts, the need for rigorous evaluation of its main aims and poverty impacts, across settings arises.

This research investigates the impact of a group loan with individual liability on indicators of household welfare in East Nusa Tenggara Province, Indonesia. Data was collected in two survey rounds in East Nusa Tenggara province (NTT) in Indonesia in 2010 at the time of application for the group loan (pre-treatment) and again in 2011 (post-treatment). 299 households were sampled in West-Timor and Alor Island in 2010 and in 2011 209 of them were successfully re-interviewed. Because of the duration of the master thesis, our research design and identification strategy, we were bounded to a short follow-up (period between baseline and follow-up) of only around one year. We therefore focus disproportionately on methods, in particular on how to deal with selection bias, an important and well-documented problem in non-randomized evaluations of the impact of microcredit programs. The focus is on methods that are nonparametric, i.e. do not rest on parametric functional form assumptions, and include the proposal of a new way of non-parametrically testing instrument validity.

1.2. Aims and hypotheses

We postulate the following null hypotheses for the outcomes wealth index, weekly food consumption, women's BMI, livestock index,

1) The impact of the uptake of TLM's group loan on outcomes at one year follow-up is not significantly different from zero (two-sided test).
2) The impact of the uptake of TLM's group loan on quantiles of the outcomes at one year follow-up is not significantly different from zero (two-sided test).

Apart from testing these hypotheses, we aim to gain insight into the distribution of uses of TLM's group loan product, categorized as productive, consumptive and educational. New nonparametric methods for dealing with attrition and selection bias will be developed.

1.3. Data and methods

Covariates, treatment status and the instrument were all observed in the first round to prevent post-treatment bias. To deal with attrition, we non-parametrically single imputed the missing follow-up outcomes.

When selecting and using statistical methods, the first thing to check is the assumptions underlying a method. Although there is certainly improvement over time, most applied economists still rely on methods that impose restrictive and often untested assumptions, even though methods are available that relax one or more of those assumptions. In order to reduce the impact of assumptions on our findings, we use a range of sophisticated statistical methods. Where possible, we explain these methods in the simplest of terms possible. Not all methods and results are very intuitive though, but what is most important for a practitioner, authors included, is to be aware of the method's assumptions and properties, more so than the route (often mathematically derived) from the assumptions to the properties. We thus focus on the assumptions and properties when discussing the estimators used.

On the methodological front, we focused on identifying interesting treatment parameters under a set of credible assumptions and showed that it is possible to obtain a credible impact estimate from a pre-existing, non-randomized intervention. By opting consistently for nonparametric methods and methods that minimize and test restrictive assumptions in general, we overcome to a high degree the concern that our estimates are driven by the particular assumptions we make, rather than by the data itself. The main methodological innovation is the development of a new nonparametric test of instrument validity in a nonparametric nonseparable triangular model. In contrast to linear IV tests, our test is consistent when none of the instruments to be tested is invalid and can thus be applied in the just identified case where only one instrumental variable is available. In addition, the test allows for conditioning on an arbitrary number of mixed categorical and continuous covariates. Our main instrument, a dummy indicating whether the household has at least one treated or previously treated acquaintance, withstands the instrument validity tests applied, when used with wealth index as outcome.

1.4. Results and interpretation

Upon non-rejection of tests of the exclusion restriction, we confidently report a negative and statistically significant at a 10%-level impact of TLM's group loan up-take on wealth index post-treatment. Although the poorer segment of the sample is more noisy, the effects are more pronounced for that subsample. But the negative point estimates are negative for all quantiles. We conjecture that this finding is due to (a) households' short-term rate of return on loan-induced investments not exceeding the loan's interest rate and fees and (b) high prevalence of non-productive loan use in our sample. This coincides with the increasing criticism microfinance institutions face of over-indebting their clients, a situation that these destitute households to sell off assets. Since 72% of the borrowers in the sample reported allocating their loans mainly to either the education of their children or to other income-generating activities, our results may be a poor reflection of longer-term impacts of loan take-up when gestation periods of investments are long.

1.5. Structure of the thesis

The structure of the thesis is as follows. Section 2 describes the area, the microfinance institution TLM and its group loan product that we aim to evaluate. The theory of microcredit and a conceptual framework for TLM’s group loan impact are put forward in section 3. Section 4 reviews existing quantitative microcredit impact evaluations, including both randomized and observational ones. Section 5 describes the data. Section 6 develops and discusses the methodology. The results are reported in section 7 and discussed in section 8. Section 9 concludes with cautious policy and methodological recommendations.

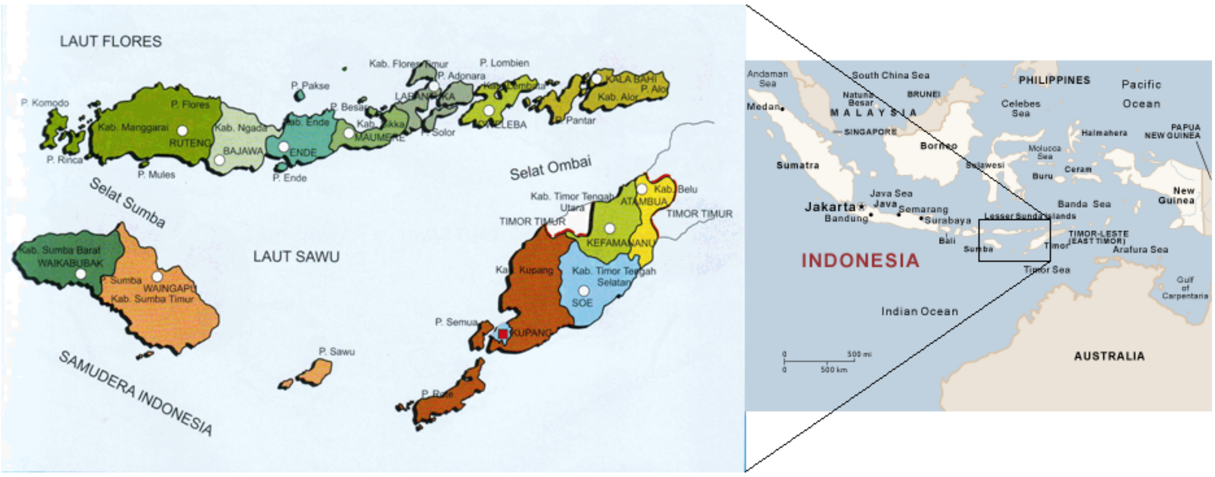
2. Description of the area, TLM and the group lending program

2.1. Description of area

East Nusa Tenggara (NTT) is one of the economically most backward provinces of Indonesia, with its regional GDP per capita in 2008 being the fifth lowest of the 33 provinces, and its per capita GDP reaching only 27% of the national GDP per capita (BPS 2010a). Table 1 gives an overview of some key socio-economic indicators of the province, as compared to Indonesia as a whole and Norway. Poverty is merely a rural phenomenon, with the rural poor making up 89% of the population below the Indonesian national poverty line in the province (BPS 2010c). Most of the poor are subsistence farmers, with income from cash crops that are sold once a year. Steep slopes, erratic rainfall and recurrent droughts are some of the challenges facing livelihoods especially in West-Timor. Livestock is culturally only consumed at wedding and funeral ceremonies. A nutrition survey led by NGO Helen Keller International in el Niño crisis year 2007 found a stunting prevalence among under-five children of 61.1% in West-Timor, well above the WHO threshold for a "very high" public health problem (≥40%). Prevalence of maternal thinness was 24.4%. Furthermore, 58.5% of children aged 3-59 months suffered anaemia, as did 35.8% of non-pregnant mothers (UNICEF 2008). In 2006, the World Food program found a 44.5% prevalence of underweight in West-Timor (excluding the provincial capital Kupang) and a 56.6% prevalence of anaemia among preschool children at the provincial level (GB 2009).

Table 1: Key economic indicators of East Nusa Tenggara Province, as compared with Indonesia and Norway (1(BPS 2010a), 2(BPS 2010b), 3(World Bank 2010a), 4(BPS 2010c), 5(BPS 2010d), 6(World Bank 2010b), 7(World Bank 2010c))

	East Nusa Tenggara Province	Indonesia	Norway
GDP per capita 2010 (Rp. market prices of 21-03-2011)[1,2]	5,916,173	22,238,784	-
GDP per capita (US$ ppp 2010)[1,2,3]	956	3,592	84,538
Poverty headcount (%, 2010)[4]	21.6	13.3	-
Net enrolment ratio, elementary school (2010, Norway 2009)[5,6]	91.0	94.7	99
Net enrolment ratio, junior high school (2010)[5,7]	51.0	67.7	95
Net enrolment ratio, senior high school (2010)[5,7]	34.9	45.6	



Indonesia (right) and East Nusa Tenggara (NTT) Province (left). The survey took place in Kabupaten (regencies) Kupang, Timor Tenggah Selatan and Alor.

2.2. Description of TLM and its group lending program

Tanaoba Lais Manekat (TLM) Foundation is a Christian non-governmental organization based in Kupang, Indonesia and founded in 1995. It currently serves 5 islands of NTT: West-Timor, Alor, Rote, Sabu and Flores. It runs a cattle-fattening program, in which the client receives a cow, which after being raised by the client is being sold with part of the profit going to TLM. Two other programs are a group loan and an individual loan product. TLM has started experimenting with a seasonal credit program, in which households receive food during the lean season and pay it back at harvest time (Basu & Wong 2011). TLM receives donor funding from three international donor organizations: US-based Kiva and the Australia-based Opportunity International and Uniting World.

This research estimates the impact on household welfare indicators of the first-time uptake of the group loan product (repeated borrowings are excluded from the analysis). Groups are most often endogenously formed by their members, but in some cases loan officers link group members. Clients are individually liable and repay weekly typically in one of the group members houses. The Standard repayment period is 104 weeks. First time loan sizes vary officially between Rp. 200,000-5,000,000 (US$ 22-557, on 1/8/12), with the overwhelming majority of loans amounting to Rp. 750,000 (US$ 84). A simple interest rate of 3% per month is charged. In its "Group Lending Manual", TLM outlines the following permitted loan uses:

1. Acquisition of capital.
2. Renovation of business premises
3. Purchases of additional stocks
4. Purchases of Saprodi (Production Facilities of Agriculture and Animal Husbandry)
5. Business Expenditures
6. Public facility development

Interviewing in West-Timor. This its members showing signs of undernourishment. Household is among the poorest in the sample. Their house in the background lacks walls.

Inside a *lopo*, a cone-shaped house in West Timor. Maize is the staple food of the rural poor in West-Timor and Alor and it is stored inside for the lean season.

However, as will appear later in this thesis, many households use their loan for consumptive or schooling purposes. It was our impression that only some loan officers ask for intended loan uses at the time of loan application.

There is no mandatory collateral; however admission to follow-up loans requires the borrower to deposit 20% of the loan size as collateral in a TLM account before loan disbursement. It seems that in practice, if no physical collateral is available at the moment of evaluation of the application of the first loan cycle, the maximum first-time loan size is Rp. 750,000. When the household possesses assets that can suit as physical collateral, higher loan sizes are approved.

TLM states the following eligibility criteria related to loan size:

1. Maximum loan amount is 150% of household capital
2. Total client liabilities including loan repayment and expenses cannot be more than 80% of income
3. Maximum debt expenses is at maximum 70% of household expenses
4. Frequency of client income is daily or weekly.

The fourth criterion is not observed according to our experiences. Farmers selling their produce few times per year, or teachers with monthly salary are among our sample of borrowers.

3. Theory

The prevalence of microcredit is a puzzling one: why do the poor not save their way out of poverty? And even if they are not able to save, why does the private credit market not take care of the credit demand of the poor? The first question is answered in section 3.1. The answer to the second question is that there are credit market imperfections, as discussed in section 3.2. Section 3.3 discusses how microcredit aims to overcome these imperfections. Section 3.4 presents the conceptual framework of the poverty impact of TLM's group loan product.

3.1. Why do the poor not save their way out of poverty

If the rates of return to capital of microenterprises are so high and microcredit is so popular, one may ask why the poor do not save more to invest their way out of poverty. There are at least three explanations for the lack of savings amongst the poor. First, poor people behave myopically (non-forward looking); they may be biased towards the present in that they are even more likely to opt for a lower amount if presented a choice between receiving a lower amount right now or a larger amount in the future. (Lawrance 1991) empirically found differing consumption and savings patterns across socioeconomic classes with the poor having a higher discount rate. The higher discount rate of the poor may lead them to save less and increase present consumption by borrowing.

Second, poor people may lack the self-discipline needed to regularly save voluntarily, given unmet primary consumption needs and the need to keep up with peers (Banerjee & Duflo 2007). In behavioural economics, hyperbolic discounting is a model of time-inconsistent discounting; normally a higher discount rate is observed when the intertemporal consumption trade-off is closer to the present. To illustrate, consider the revelation of time preferences by the choices: (A) "Would you prefer Rp. 50,000 tomorrow or Rp. 60,000 three weeks from now?" and (B) "Would you prefer Rp. 50,000 one year from now or Rp. 60,000 one year and three weeks from now?" When dealing with real monetary payoffs, many subjects will take the lesser amount tomorrow in choice (A), but are willing to wait a little longer to receive the higher payoff in choice (B). Neurological and behavioural economic experiments have shown evidence of hyperbolic discounting (Bauer et al. 2012; Pine et al. 2009).

A third explanation is that social pressure prevents household from accumulating capital. This is the well-known argument that sharing and social commitments acts as an income tax that gives weak incentives to save. When the rumour spreads that someone has accumulated a substantial sum of savings, he then becomes more susceptible to requests for grants and loans from family members and other acquaintances. Within households, women (men) may not be able to keep their savings from their husbands (wives). When writing from a rural Kenyan context, (Dupas & Robinson 2009) found that when forced to make an emergency purchase, a majority of respondents actually preferred to take up a microloan rather than draw from their stock of savings. The rationale behind this was that to neighbours the sight of a loan officer collecting weekly repayments at the house would serve as a signal of lack of liquidity to lend out.

The inability of the poor to accumulate savings gives rise to credit demand. Why can the private sector not meet the credit demand of the poor? The following section gives some insights.

3.2. Credit market imperfections

Assuming, as is standard in microeconomics, a concave production function with diminishing marginal returns to capital, it follows that, ceteris paribus, those poor in capital have a higher rate of return to capital. In a competitive market, the poor should therefore be willing to pay a higher rate of interest on loans (Armendariz & Morduch 2010). Hence, from the perspective of an investor or lender who is maximizing his profits, the poorer an entrepreneur is, the more attractive he should be as a potential borrower. Therefore, money should flow from rich depositors to poor entrepreneurs. However, the contrary is often true in capital-poor countries: larger firms and businesses enjoy better access to capital. The first four explanations for this puzzle are rooted in risk, specifically, the risk of default from the perspective of the lender. These are related to principal-agent problems; they concern the asymmetry of information between the lender (the principal) and the borrower (the agent), so that the lender cannot ensure that the borrower is acting in the lenders' best interest. The fifth explanation is a credit market imperfection brought about by high transaction costs.

The first explanation is that in poor societies, banks lack good mechanisms to collect funds profitably. To compensate for the risk of default on loans, banks may require collateral that has stable value, is easy to seize and to liquidate (land being the prime example). The poor often lack such assets. And even when they have assets that can suit as collateral, the problem often is that property rights are not clearly defined. Furthermore, lenders face enforcement problems as in many countries judicial systems are weak, in which case it is difficult to get a loan contract enforced in court.

Second, the lender cannot perfectly observe the riskiness in terms of default probability of the borrower when the latter is applying for credit. As risk premium, lenders could raise interest rates to offset this default risk. However (Stiglitz & Weiss 1981) showed that, when they would do so, safe borrowers, with a low probability of default and thus a low expected rate of return, would end up paying back a relatively large amount to the lender rendering their ventures unprofitable. Thus, the interest rate has a sorting effect: raising interest rates would drive safe borrowers out of the credit market, a phenomenon termed adverse selection. The resulting risky clientele will contribute to lower profits trough higher default rates. On the other hand, very low rates of interest will also generate low profits to the bank. The bank thus faces an inverse U-shaped credit supply curve as a function of its interest rate, depicted in figure 1(a). The maximum of the supply curve is the bank optimal rate that optimizes the trade-off between interest income per non-defaulting client and loss through default. Figure 1(b) shows the credit market equilibrium under credit rationing. With demand curve 1, there is excess supply, whereas with demand curve 2, there is excess demand, with some borrowers willing to pay higher interest rates not able to obtain a loan, which is the common definition of credit rationing.

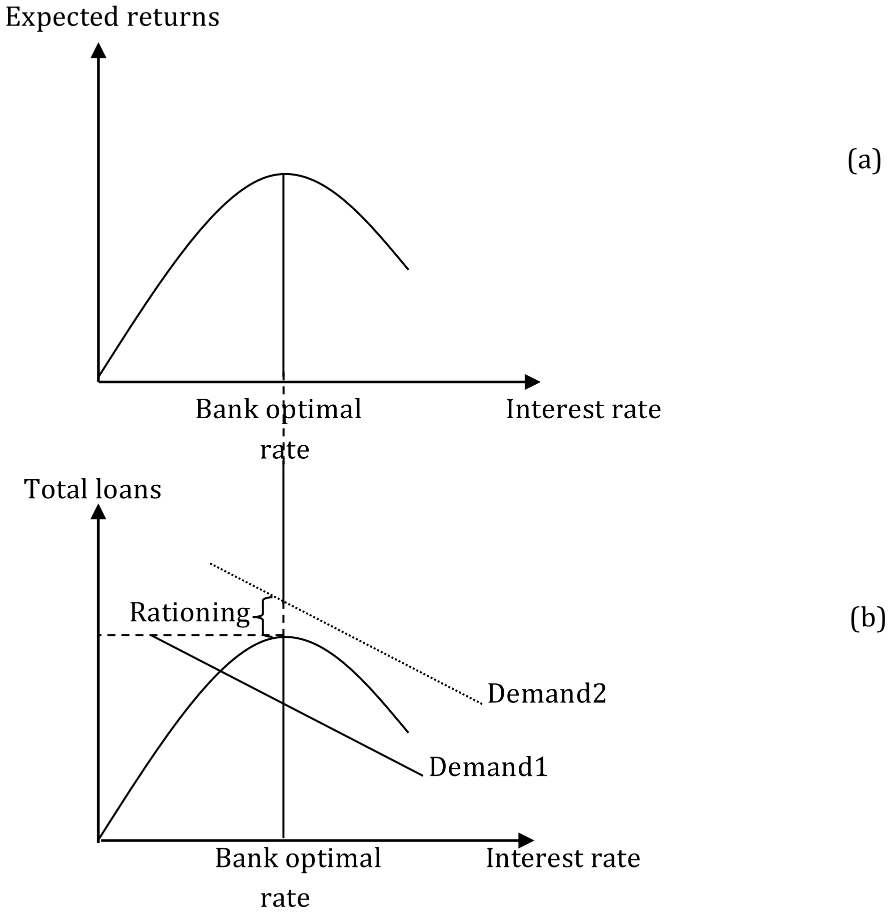


Figure 1: Equilibrium credit rationing. Source: adapted from (Stiglitz & Weiss 1981).

The third explanation is linked to the first one. Imagine, as in the model of (Stiglitz 1990), that the borrower faces a choice between investing his loan in a project with returns that are can be large or zero and a project which is perfectly divisible. For illustration, the indivisible project may be buying and trying to sell a television and the divisible project buying and selling salt. Even when the project of selling salt in small amounts is not as successful as expected, at least some returns would likely be generated, from which loan repayments can be made. In contrast, a failure of the television project will generate zero returns to the borrower from which loan repayments were supposed to be made. From the perspective of the lender, the television project may be preferred, as it maximizes his expected net returns. In a situation with limited liability, after a borrower has received his loan, he realizes that if he defaults, it is the bank that will turn up for the consequences. Therefore, the borrowers' expected net return is higher when he chooses the indivisible television project.

The borrower not only makes a project choice after loan disbursement, he also has to choose his effort level. Effort has a cost to the borrower. A higher interest rate will lead the borrower to take more risk and lower his effort. Since the borrower does not have to bear the consequences of default, his effort level may be lower than the level maximizing the lender's profits. As a result, in the absence of collateral to insure the lender, the lender will ration credit. The higher than optimal risk (from the perspective of the lender) taken by the borrower in his choices of project and effort level is termed ex-ante moral hazard, since these choices are made after loan disbursement but before project returns are realized.

As fourth explanation, consider what happens when project returns are realized. Upon project realization the borrower may willingly decide to default. This happens because the lender cannot fully observe the project outcomes and/or the borrower can choose to falsely reports losses. Wilful default may occur even when the lender has full knowledge of the project outcomes. This scenario is influenced by weak judicial systems that cannot enforce credit contracts that are weak to begin with, for example when project outcomes cannot be verified.

A final explanation for the lack of lending in poor areas is the high transaction costs associated with small loan sizes and even lower repayments. The cost of paperwork and loan officer salaries is relatively high compared to the returns to the bank from the loan. Poor rural borrowers living in remote places may simply not afford to head to a credit branch frequently to repay. Living remotely makes it also more costly to recruit new clients and screen loan applicants. For the aforementioned reasons, private lenders may either not be available to, nor find lending to these "unbankables" profitable (Morduch 1999).

3.3. Microcredit as a welfare-improving intervention in a second-best world

Market imperfections such as the ones described above lead to economic inefficiencies if capital-poor entrepreneurs cannot invest and expand their businesses and poor consumers cannot sufficiently smooth their consumption. When credit is rationed, the introduction of microcredit institutions can expand credit access for the population and move the economy closer towards pareto efficiency. In an environment with limited competition among microlenders, the offer of repeat-borrowings conditional on full repayment of the previous loan cycle creates a dynamic incentive to repay. TLM indeed offers follow-up loans if the client had a perfect repayment record in the first loan cycle. The incentive scheme reduces the risk of both ex-ante and ex-post moral hazard. Also, even though liability is individual rather than joint, peer pressure may have a positive influence on repayment records. Microcredit organizations also try to find alternatives to physical collateral for example compulsory savings accounts. For follow-up loans, TLM for instance requires the client to open an account at its bank and deposit 20% of the loan size upfront in it as collateral substitute.

Being a non-profit organization, a microcredit organization like TLM has a different objective than traditional banks. While the former may scale back credit supply to the point where net returns are maximized, the latter may try to maximize outreach to the poor. This is reflected in the vision statements of BRI, the largest commercial bank in Indonesia (also present in NTT) and TLM. The former's vision is "to become to most prominent commercial bank that puts its clients first"[i], while TLM's website states "The Vision of TLM is "To show the love of God to the world" which is expressed through the creation of small businesses throughout the NTT region, the poorest region in Indonesia." The TLM group lending manual for its staff states, "Besides to cover the operation cost, the product must also focus on serving the poor". Being a charity backed by donor funding, it can bear the losses of higher rates of loan defaults resulting from, selection and offer loans demanding little or no collateral. Moreover, it enables them to cover the transaction costs associated with weekly client meetings far from the credit branch.

3.4. Conceptual framework

Figure 2 presents an overview of the different causal channels from the offering of microcredit by TLM to short-term poverty impact on the population. The figure is to be read from top to bottom. First, the applicant can be poor or not so poor; he can be accepted or rejected. When accepted, the loan can be used for consumption or it can be invested in either education or (other) income-generating activities. Business investments can fail or they can succeed, but even if they succeed, they will only have a positive short-term impact on poverty if the rate of return on capital exceeds the portfolio yield (interest rate + fees) on the loan. There have been a few randomized experiments with cash or working capital transfers to owners of microenterprises. (De Mel et al. 2008) for instance found rates of returns of 4.6-5.3% per month in Sri Lanka, (McKenzie & Woodruff 2008) found monthly rates of 20-30% in Mexico and (Pearlman 2012) found monthly rates of 3.5-21% in Ecuador, all substantially higher than the 3% monthly interest rate charged on TLM's group loan.

When we take a holistic perspective and include school attendance in our definition of poverty, then parent's investment of the loan in the education can lower the probability of their dropout of school and thus decrease poverty in the long-term. When the loan is used to smooth consumption, this can prevent the loss of human capital through nutritional or caloric deficiency and hence prevent the decline of work productivity (not shown). It may also lower the probability of the sale of (productive) assets, decreasing short-term poverty. Note that positive signs alongside immediate arrows leading to the node "short-term poverty impact" indicate that poverty is reduced, while negative signs correspond to increases in the extent, depth and/or severity of poverty. Of course, multiple loan uses are possible, but that does not impede the usefulness of the graph in conceptualizing the impact channels of TLM's group loan on short-term poverty.

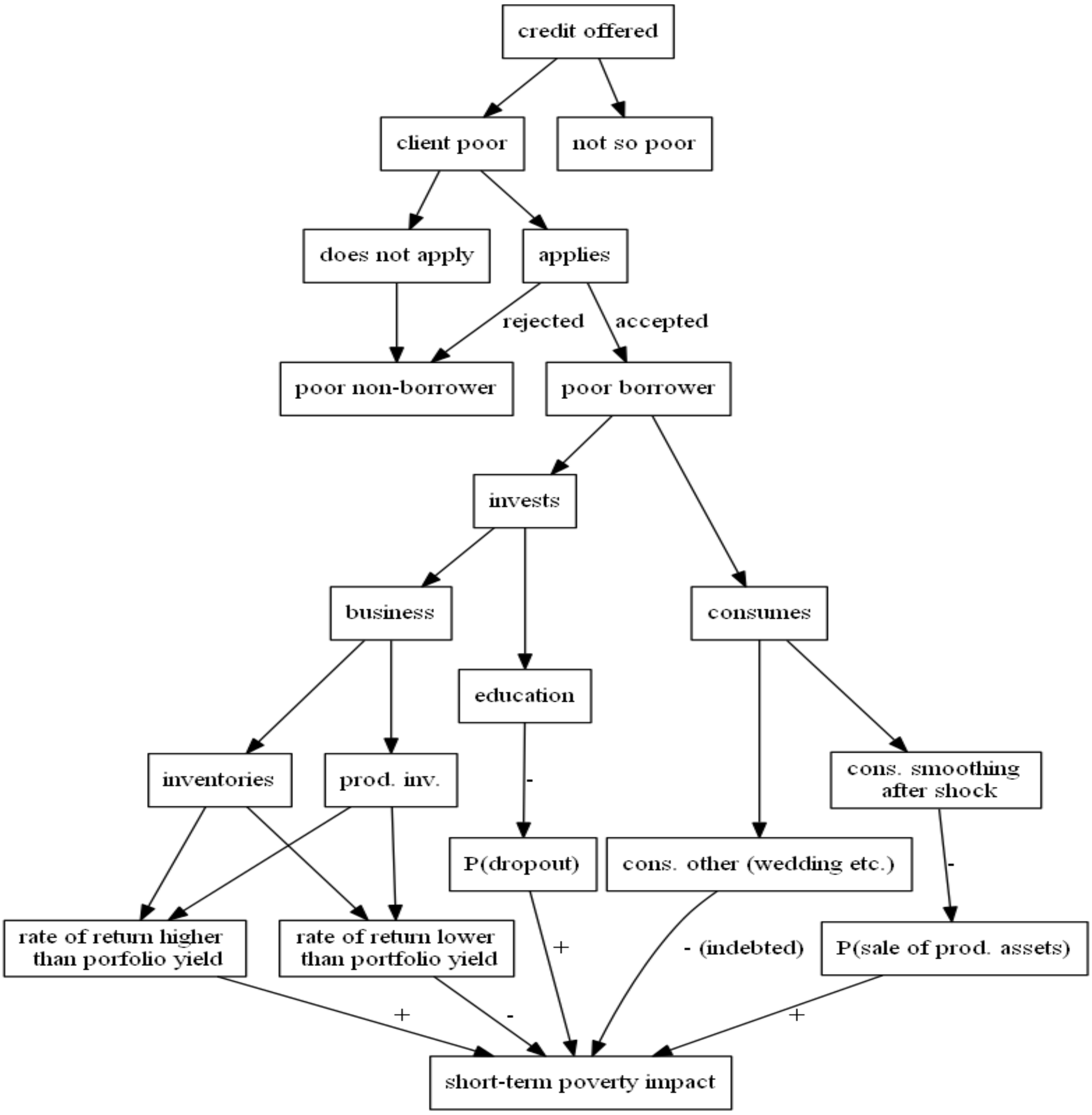


Figure 2: conceptual framework of the short-term impact of TLM's microcredit on poverty.

4. Review of microcredit impact studies

The many impact evaluations of microfinance programs worldwide can be classified in at least two categories: the randomized, controlled experiments and the evaluations based on observational, that is, non-experimental data. In impact evaluation, one is interested in a what-if question: what would have happened would the non-treated household been treated? What would happen to a non-borrowing household had its members borrowed? The fundamental problem of impact evaluation is that of missing data: one can at one point in time only observe the household in a treated state or in a non-treated state. Therefore, in order to answer the what-if question the counterfactual has to be estimated, by means of a control group. Randomized controlled trials are considered the “gold standard” for partial equilibrium impact evaluations, since, if carried out properly, the difference in outcomes between treated and control units can be attributed exclusively to the intervention under study. For non-randomized studies, additional assumptions are typically necessary to be invoked in order to gain useful information on the effects of a treatment. Many different methods exist to do that, and many different techniques have been applied to the microcredit impact question. The randomized evaluations are briefly reviewed in section 4.1, the observational ones in section 4.2.

4.1. Randomized studies

Randomized controlled trials have made an upsurge in development economics during the last decade. With newly established microfinance-focused research institutes such as Poverty Action Lab and Innovations for Poverty Action conducting exclusively randomized impact evaluations and the majority of World Bank impact evaluations now being randomized ones. To the best of our knowledge, there have been conducted 7 randomized controlled trials of microcredit impact in a wide range of settings. They are reported in table 2. In general, these studies have two things in common: they contain mostly female borrowers and their follow-up periods are short, one to one-and-half year.

The studies of (Karlan & Zinman 2010) in South Africa and (Karlan & Zinman 2011) in urban Philippines applied randomization at the level of the household. They provided half of a group of rejected loan applicants identified as marginally rejected with credit. In the Philippines, where the borrowers were not so poor, household welfare impacts were insignificant, although profits went up, consistent with productive investments. In South Africa, consumptive microcredit with high interest rates was found to have a positive welfare impact, presumably through consumption smoothing. This is a salient finding, as many development practitioners are sceptical of such loans; indeed, the behavioural economics literature predicts that people with limited self-discipline may over-borrow.

The general findings in the other RCT’s at best give mixed results on the impacts of microcredit on poverty related measures, with no sign that microcredit affects poverty. The closest contact with poverty impact came in the Mongolia trial where the group microcredit induced increased food consumption (Attanasio et al. 2011). In contrast those that too up individual microloans in (Augsburg et al. 2012)’s study spent less on food. It should be noted that these trials were all evaluated over a 12-36 month period, leaving open the possibility of positive impacts over the long-term.

Table 2: Randomized controlled trials of microcredit impact

Citation	Where	When	Level of randomization	Liability (group or individual	Follow-up (months)	Impacts
(Karlan & Zinman 2011)	South Africa	2004-2005	Individual	G	6-12	welfare: +
(Banerjee et al. 2009)	Hyderabad, India	2006-2008	District	G	12-18	Profits + Wellbeing 0
(Crepon et al. 2011)	Morocco	2006-2009	Village	G (mostly)	24	0
(Attanasio et al. 2011)	Mongolia	2008-2010	Village	G,I	8-17	Group: food spending +
(Karlan & Zinman 2010)	Manila, Philippines	2006-2008	Individual	I	11-22	0
(Augsburg et al. 2012)	Bosnia and Herzegovina	2008-2010	Individual	I	Approx 14	Food spending: -
(Desai et al. 2011)	Ethiopia	2003-2006	Village	G	36	Wellbeing: mixed Education: +

4.2. Non-randomized studies

Non-randomized evaluations suffer from bias due to non-random placement of microcredit programs and self-selection by microcredit clients. Statistical methods need to be used to account for the endogeneity of treatment selection. There is long gradient of approaches in the impact evaluation literature, with some designs being stronger than others. We will discuss the different approaches and some findings here very briefly.

Some rely on difference-in-difference designs, assuming parallel time trends for treatment and control groups. When that assumption does not hold, difference-in-difference and fixed effects estimation are biased. Matching approaches produce biased estimates when there is selection on unobservables, that is, when there are unobserved factors that affect both the propensity of selecting into treatment as well as the outcome. These are rather strong assumptions, so we do not report the findings of those studies here.

In a well known and often-cited research, (Pitt & Khandker 1998) and (Khandker 2005) use an eligibility criterion of owning less than half-an-acre land in Bangladesh as an instrument for the demand for microcredit. The exogeneity of their instrument is questionable however, and Roodman and Morduch (2009) showed that there was actually no discontinuity in the uptake of microcredit with respect to landholdings. In another attempt, (Schroeder 2010) uses the same dataset and exploits heteroskedasticity to identify the credit variable. The identification result critically depends on the conditional covariance being constant. Following her reasoning on the error structure, this condition fails for example if households that have higher financial literacy, or have less (or greater) access to valuable social networks are more likely to respond to a negative economic shocks by seeking microcredit. She finds that microcredit has a positive and significant effect in Bangladesh. Finally, (Berhane & Gardebroek 2011) investigate dynamic longer-term effects of repeat borrowings using a long panel dataset and a random trend model, accounting not only for time-invariant but also for linear trends in household-level unobserved heterogeneity. They found that borrowing increases consumption and housing, and that there are long-term cumulative effects. Short-term impact estimates may thus underestimate impacts of microcredit programs, a fact to bear in mind when interpreting our findings.

5. Data

5.1. Sampling strategy

Random sampling is the gold standard sampling strategy since it ensures representativeness of the sample to the population from which it is drawn, so that sample parameters equal their population counterparts in the limit. Our population under study consists of households from the non-institutionalized civilian population of the areas in NTT where TLM operates. Random sampling from this population was not practically feasible due to logistic constraints and the non-availability of a population register of these Kabupaten (regency). Moreover, it would include too few borrowers and be incompatible with our identification strategies. As we will discuss further in the Methods section, to prevent post-treatment bias, relevant time-varying covariates need to be measured before the treatment (the uptake of the group loan in our case) takes place. A more realistic stratified sampling scheme was therefore employed, with purposive oversampling of treated households. The first survey round, pre-treatment, took place from April to October 2010, while the second survey round (post-treatment) took place from May to October 2011.

First, the three Kabupaten Kupang, TTS and Alor were selected by convenience: the ability to have guest families that could host the enumerators. The guest families were located in Kupang for Kabupaten Kupang, in So'e for Kabupaten TTS and in Kalabahi for Kabupaten Kupang. Within these Kubupaten, borrowers were selected from branches of TLM that were reachable by motorcycle from the respective locations of the guest families. Borrower households were sampled from the following branches: Kupang C1, Kupang C2, Alor, So'e, Kapan, Niki-Niki. Borrowers were first surveyed after they applied for the loan but before they received it. We obtained the names and addresses of the new loan applicants from the administrative records of the respective branches of TLM, to which the management of TLM foundation agreed. In some cases, borrowers were interviewed just after their loan was disbursed, at most two weeks after. This could have led borrowers to temporarily increase (food) consumption expenditures using the obtained loan funds, invaliding the measurement of certain covariates and outcome measures in the first survey round as being pre-treatment. Such effect is less likely to occur when considering durable assets. The choice-based nature of our sampling of treated households potentially induces bias arising from the same underlying mechanism as the selection bias within our sample that we discuss in length in the Methods section.

Households that did not borrow from TLM were included in the sample as counterfactuals at a ratio of 1:1 to treated households. Initially, an attempt was made to select control households randomly within the same RT (Rukun Tettanga, smallest administrative unit in Indonesia) from the list of households obtained from the head of the RT. When the head of the RT was not present, we proceeded to either the RT with one number lower or higher, determined by a flip of a coin and tried to sample the control household from that RT. However, we found that the gain of random selection within the RT or from a neighbouring RT relative to convenience sampling was relatively small, as many households first selected from the lists of the head of the RT were not at home. Ideally, repeated attempts at interview should be made to ensure random selection of control households, but we did not have the time and resources for that.

We opted for convenience sampling: we selected nearby located control households whose household head or his (all of the household heads in households with a married couple were male) wife were at home, without purposefully favouring one particular direction from the treated household to search from. We conjecture that there are two potential ways in which these control households differ in systematic ways from arbitrary households that satisfy the inclusion criteria. Firstly, households living remotely or far from the road may have been under-sampled. We deem this bias limited as most of our treated households lived next to a road, making them comparable on this aspect. Secondly, households that were at home were oversampled and the probability of the household head (or his wife) not being at home is likely to be positively dependent on hours worked away from home. This may be particularly true for the farmers in our sample, since the holders of kiosks are typically at home during the day. To the extent that the factors causing differing number of hours spent outside of the home between arbitrary households and our control households lead to differing household incomes, we try to control for this source of bias by including a pre-treatment household asset index as control variable. Being away from home for differing amounts of time could also be due to more or less taking part in non-income generating activities. A main concern for potential bias here is that the control households sampled differ in their access to valuable social networks which may impact on their nutritional status. Households with heads (and their partners) that are more likely to be at home might be less likely to eat at their neighbours' houses, thus influencing nutritional status, considering that meals served for guests may have higher nutritional or caloric value than meals cooked and consumed exclusively for the household itself. This is an issue to keep in mind when interpreting our estimates that use anthropometric measures as outcomes. Note that this concern would have been taken care of only when returning to the control household for sufficiently many times to get to interview them, a time-consuming task since many farmers leave by sunrise and return home at sunset, after which interviewing would be deemed impolite and could create distrust among the population.

The deviations from random sampling have negative bearing on the external validity of our estimates, whereas they may still uncover useful policy information. Although limiting the generalizability of our findings to the population from which our sample was drawn and to other populations, this approach allowed us to get the most statistical efficiency per Rupiah spent on data collection. A choice-based sampling of treated units is not uncommon in applied settings. In medical studies, treated individuals ("cases") are similarly oversampled in hospital settings relative to the general population in order to include enough treated units. We do not take issues of identification lightly and thus only claim to identify sample parameters of interest (such as sample average treatment effects), rather than the respective population parameters.

5.2. Sampling inclusion/exclusion criteria

To increase comparability of treated and controls, inclusion criteria for control households were based on TLM's eligibility criteria for its group loan product. TLM's eligibility criteria are (TLM 2010):

1) The applicant must possess an official identity card showing the area in which they currently live, which the area TLM serves (i.e.: to get that means the client must have been there for a reasonable amount of time, minimum 6 months)

2) The applicant must demonstrate ownership of a business/having regular income every month. The business has been managed for at least 6 months when the application is submitted.
3) The Applicant must be at least 18 years old and less than 63 years old (to allow insurance and for legal reason).
4) The Applicant must be prepared to open a bank account with TLM's BPR (TLM's commercial bank).
5) The applicant should have to be honest in providing information if they have any outstanding debts with other financial institution or people.
6) The client's current business can cover the repayments & interest of the upcoming loan.

We included/excluded control households by evaluating criteria (2), (3) and (6). Criterion (1) could potentially be fulfilled by households lacking ID seeking a loan, while criteria (4) and (5) are unobserved to the enumerator. Further, we excluded households with heads over 80 years old.

5.3. Measurement error

As econometrics textbooks explain, in the OLS model under Gauss-Markov assumptions, when the left-hand side error is classical, that is,

$$y_i = y_i^* + u_i$$
$$E[y_i^* u_i] = 0$$

where y_i, y_i^* are the observed and true value of the outcome, then the coefficients of the model will be consistent, the cost being only inflated standard errors. However, when the measurement error depends on the true value of the outcome, $E[y_i^* u_i] \neq 0$ generally, and the measurement error is mean-reverting, i.e. biasing the coefficients away from zero.

(Hoderlein & Winter 2010) show rigorously that in nonseparable models, which we will use in our estimations, local average structural derivatives , such as local average treatment effects or quantile treatment effects, are identified only if the individual uses more information in his or her back-cast than the information set the econometrician has available from the regressors of the model. This condition is difficult to interpret or evaluate in applied settings. Importantly, they show that quantiles may be more sensitive to ill-behaved measurement error than averages, something to bear in mind when interpreting our quantile treatment effect estimates. (Hoderlein & Winter 2010) also conclude that applied researchers should try to let respondents recall means rather than medians (typical values) of variables of interest, since only then (and under the foregoing assumption on active use of information), local structural average derivatives are identified. In our research we did not systematically favour eliciting either the mean or median value for consumption back-casts, and the interview approach may differ among enumerators, an additional source of (interviewer) bias. It could be doubted upon whether respondents are able to give more reliable estimates of mean rather than median consumption values from a cognitive neuroscience point of view, a topic for further research.

(Millimet 2011) ran a series of Monte Carlo simulations to evaluate the robustness of 10 different treatment effect estimators to different forms of measurement error in different components of the model. One important finding that stood out was that even slight nonclassical measurement errors in outcomes induce large biases in effect estimates across all estimators studied. Measurement error in covariates biases coefficients towards zero, a phenomenon that is also termed attenuation bias.

5.4. Outcome measures

The interest of this study lies in the impact of the uptake of TLM's loan product on poverty and malnutrition. Since the gap between the baseline and follow-up survey is around one year, outcome measures sensitive to short-term changes in household welfare status are needed. Household income and consumption are two obvious candidates. During the survey work, we found that many households had considerable difficulty in accounting for their household income and profits. In addition, income streams in this area are highly seasonal (Basu & Wong 2011) and often agricultural produce is sold once or a couple of times a year. Hence we opted for household consumption, including consumption of goods that do not show up in households' expenditures. These non-monetary sources of consumption include the consumption of food that is not bought by the household and the use of firewood that is collected by the household. Environmental dependence is important to the livelihoods of rural communities in the region of study as it is in most poor rural communities around the world and neglecting this "hidden harvest" can lead to severe misrepresentation of rural realities and misleading inferences (Angelsen et al. 2011). Respondents were asked to value the consumption of food that was not bought at local market prices. Including the consumption of food not bought by the household is important in our setting, as loan funds could be invested in increased agricultural productivity for self-consumption, thereby improving the nutritional status of the household members, while not showing up in monetary household income, consumption or assets. Consumption of firewood was obtained by multiplying the times a family member collects firewood from the forest by the local market value of one such collection of firewood, making the distinction between adult and child head loads, as they may differ in quantity and price.

The problem with weekly (for food and firewood) and monthly (for other items) recall consumption data is that it is likely to be strife with nonclassical measurement error. The dependence between consumption and recall error could plausibly be due to factors such as age, education, and mental ability that affect both the ability to recall and the share of durables/non-durables consumption (which will likely affect recall error) on the one hand and consumption on the other hand. (Hoderlein & Winter 2010) investigated the Health and Retirement Study (HRS) and a supplemental survey, the Consumption and Activities Mail Survey (CAMS) and found huge measurement errors in food consumption data in the which are typically assumed to be relatively accurately measured.

Our sample differs from the US population in that food consumption is much smaller and contains more 'unadorned' food and more monotonous diet. Furthermore, we asked respondents for their weekly food consumption amount and value in a much disaggregated manner: we used no less than 229 food categories. The food consumption table was taken from the Indonesian national socio-economic survey (SUSENAS[ii]) of 1999, when food consumption was still recorded in a highly disaggregated manner. We found that respondents have considerable difficulty recalling their clothing expenditure of the last year. Hence, given the above concerns of non-classical measurement error, we do not use total consumption as outcome.

To further overcome concerns of left-hand side measurement error, we also asked the respondents for each of 21 assets whether their household possessed at least one of them, see table 5 in the Methods section. Using categorical principal component analysis, we constructed a continuously distributed wealth index from these binary asset indicators, livestock counts and housing infrastructure (flooring and roofing types). The resulting wealth index is expected to be accurately measured. A drawback of this measure is that, given the short follow-up period and the durable nature of these assets, household welfare may have changed as a result of the loan, not yet showing up in asset holdings. The component of the wealth index we expect to be most sensitive to changes in household welfare is the livestock holding. We describe in the Methods section how we constructed the livestock index.

Our nutritional outcome is the BMI of a woman in the household. If the head of the household is not a woman, the wife of the head of the household's height and weight are measured. If the head of the household is not a woman and he does not have a living wife that is available, then another female household member of at least 18 years old is selected first by availability and second randomly for anthropometric measurement. Height was measured on bare feet with non-stretchable measuring tape, while fixating the head's position using a headboard. In case of unsurfaced flooring, ample effort was put in to locate a straight surface area to minimize measurement error. A digital weighting scale, A&D UC321, (A&D, Abingdon, UK) with precision up to 0.05 kg was used. Unlike the household asset index, anthropometric features are able capture changes in hygiene and diet from food consumption that is not purchased by the household. Furthermore, the error in these measurements is likely to be limited in size and of a classical nature. We also measured serum haemoglobin on a consenting subsample of women of productive age (18-40 years old), but exclude this outcome measure from our analysis since the sample is too small to uncover useful policy information.

5.5. Treatment and covariates

The main predictor of interest is a binary treatment indicator taking on the value of 1 if the household is treated and zero otherwise. We believe there to be no misclassification of treatment status. Treated households were selected based on administrative data from TLM. Potential control households were first asked of their borrowing status before being selected as control households and being explaining the survey and its aims.

Table 3 shows the names of the covariates measured and their descriptions. Excluded from the table are the asset indicators and the livestock variables. They are described in the discussion on the construction of the wealth and socio-economic indices in the Methods section. All covariates are measured in round 1, pre-treatment. Controlling for variables that are themselves affected by a treatment in a regression estimation or matching model leads to post-treatment bias, since one causal pathway from the treatment to the outcome is controlled away (Angrist & Pischke 2009; Lee 2005; Wooldridge 2005).

Table 3: Description of key variables.

Variable name	**Description**
(Outcomes)	
wealthindex2	Index constructed from asset and livestock holding using CATPCA (round 2)
wealthindex2_ imputed	*wealthindex2* with imputed missing values
BMI_woman2	Body Mass Index of woman in hh as described under data (round 2)
foodconsweek	Monetary value of weekly household consumption in Rp.
livestockindex2_ imputed	Index constructed from livestock holdings using CATPCA (round 2) with imputed missing values
(Treatment)	
treatment	1=household borrows from TLM, 0=otherwise
(Instruments)	
iv1	The number of people the hh head or other adult respondents of the hh know who borrow or ever borrowed from TLM, not including the members of the current borrowing group for treated hhs
iv2	1=iv1 bigger than 0, 0=otherwise
(control vars.)	All measured in round 1
hhsize	Number of hh members
age	Age of the hh head
gender	Gender of the hh head
primaryschool	1=hh head completed primaryschool, 0=otherwise
smp	1=hh head completed smp (junior high school), 0=otherwise
sma	1=hh head completed sma (general senior high school), 0=otherwise
smk	1=hh head completed smk (technical senior high school), 0=otherwise
dip1orhigher	1=hh head completed a post-secondary school degree, 0=otherwise
religion	1=protestant, 2=catholic, 3=muslim
spm	Raven's Standard Progressive Matrices, a proxy for reasoning ability
finlit	The number of correct answers to a test consisting of 5 questions, proxying financial literacy
friends	Total number of friends, as answer to the question "About how many *close friends* do you have these days? These are people you feel at ease with, can talk to about private matters, or call on for help."[iii]
wealthindex1	The wealth index excluding livestock in round 1.
oownland	1=landarea>0, 0=otherwise
savings	1=has savings account, 0=otherwise
muslim	1=muslim, 0=otherwise
catholic	1=catholic, 0=otherwise
hasbusiness	1=hh runs a business, 0=otherwise
incomediv	1=more than 1 income generating activity, 0=otherwise
educyears	Years of education of the households head
trustgeneral	"In general, can people be trusted?" 1=yes, 0=no
mainloanuse2	Main (according to share of the loansize) loan use as reported in round 2. 1=production; 2=consumption, 3=education.

In our main estimates, we try to select those variables that are less prone to measurement error, since (Millimet 2011) found that measurement error in covariates can have negative repercussions in terms of bias and mean absolute percentage error in his simulations, $MAPE = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\tau}-\tau}{\tau}$, with $r = 1,2,\dots,R$ being the data replications and τ and $\hat{\tau}$ being the true respectively estimated Average Treatment Effect. For this reason we categorize schooling into dummies of the household head's highest educational attainment rather than years of schooling, which may be more prone to recall errors. An exception to this rule is *friends*, which is probably (nonclassically) mismeasured, but we use it in some of our analyses in order for the exclusion restriction to hold.

5.6. Instrumental variable

Our instrumental variable (IV) is the number of people known by the household head who borrow or ever borrowed from TLM, recorded at baseline. We "binarize" this variable to make it less prone to outliers and nonclassical measurement error, with it taking on the value 1 if the respondent knows at least one person who borrow(ed) from TLM at baseline and 0 otherwise. We thus have four types τ of households, defined by their potential treatment status $D(.)$ by instrument value, see table 4. Note that the potential treatment status of an observation is not observable, as we can only observe either $D(1)$ or $D(0)$; the typology will be useful later when we introduce the IV estimators to be used.

Table 4: Types according to potential treatment status by instrument value.

Type τ	$D(1)$	$D(0)$	Notion
a	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
n	0	0	Never takers

To the best of our knowledge, this instrument has not yet been used in the literature, be it in microfinance impact evaluation or in the evaluation of some other program. An instrumental variable, most generally, must be (a) relevant, i.e. strongly related to the potentially endogenous variable it aims to instrument and (b) it must be valid, that is, it must not contain any information about potential outcomes conditional on the treatment (and possibly a set of control variables) (Wooldridge 2010). We will define these conditions more precisely for the particular models we use in the Methods section. In the Results section, we assess the relevance and validity of our instrument empirically.

With regard to condition (a) relevance, when adult household members get to know people who borrow from TLM, then this is likely to *ceterus paribus* affect the likelihood of applying for TLM's group credit, at least for the part of the population. The sharing of information about a product by a neighbour or friend, when perceived as positive, is likely to increase the propensity to borrow. Based on interview questions to non-borrowers and informal discussions around the interviews, we believe that the subsample of households whose treatment status was affected by the instrument consists mostly of compliers and few or no defiers.

The exclusion restriction, i.e. the validity of the instrument is crucial to the credibility of the IV estimation and inference. We assume that our instrument is valid conditional on the nine variables *wealth_index1*, *friends*, *finlit*, *spm*, *primaryschool*, *smp*, *sma*, *smk* and *dip1orhigher*. In the models we use, the exclusion restriction requires that (1) the instrument Z does not affect the outcome Y exclusive of (i.e. other than through) the treatment D and our seven additional covariates X and (2) that there is no common cause of Z and Y other than D, X. The latter requirements are not implied by the instrument validity condition for linear IV models $cov(D, \varepsilon)$=0 stated in econometrics textbooks, but they are necessary to give a causal meaning to IV estimates (Reiss 2005). These possible violations (1) and (2) of the exclusion restriction are depicted by what (Chalak 2012) named direct causality graphs in figure 3 and 4. A direct causality graph depicts causal relations by directed edges (arrows) between nodes representing variables.

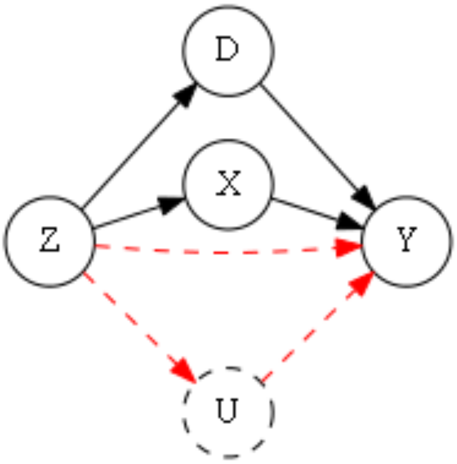


Figure 3: possibility 1: Z U affects Y exclusive of D, X, directly or via U. Red edges indicate violations of the exclusion restriction. U represents an unobservable.

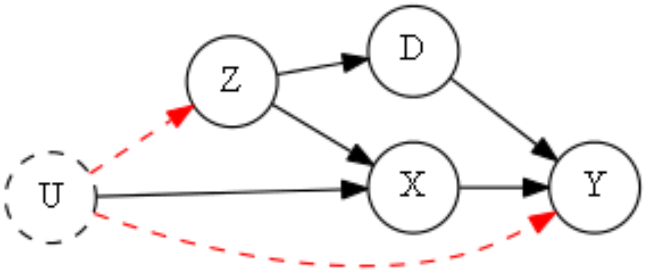


Figure 4: Possiblity 2: Z and Y share unobserved common cause U.

Possibility 1 is not likely to occur with our instrument; it is difficult to see how getting a borrower as friend in itself affects your income, other than trough unlikely and probably small treatment externalities. Possibility 2 is of more concern. The analysis is the same for all four outcome measures, nutritional status and wealth index, socioeconomic status index and child schooling, since these variables are directly determined by household income. These A household or its members may possess certain characteristics that affect the number and type of people they get to know. The more people are known to a household, the more likely a (former) borrower from TLM is known to the household. Since the "number of people known to the household" is difficult to measure, we try to capture this variable by controlling for the quantity of good friends by including *friends* as covariate. The number of people one knows is likely to be strongly associated with the number of good friends one has. The type of friends one makes is likely to depend on a household's characteristics. People in general tend to make friends that are more like themselves, a phenomenon called "assortative matching" in sociology. We therefore included covariates capturing financial literacy (*finlit*), intelligence (*spm*) (see table 3 and the appendix for their descriptions) and education level of the household head (the education dummies) in X. We assume that any other shared determinants U of Z and Y is blocked by D, X, i.e any arrow from U to Z and/or from U to Y passes through D, X, thereby preventing a violation of the exclusion restriction.

To illustrate, when U contains a measure of how socially active or outgoing the household head is, then this could affect the number of acquaintances, but not their type. The pathway from U to Y would then be blocked by $friends$, which is in X, thereby preventing a violation of the exclusion restriction. Imagine now that U contains a measure of the average entrepreneurial skills of the households' acquaintances. Then U will affect Z if entrepreneurial skills are related to the expected rate of return on investment and thereby with the propensity to apply for credit of the acquaintances. On the other hand, having friends who are more entrepreneurially skilled could lead the household and its members to increase their skill levels too, through skill transfer or emulation of business practices. If a household's skill level affects its outcomes, then this would constitute a violation of the exclusion restriction and render the instrument invalid. Such concerns are legitimate. Note that this is not a problem specific to our application, as potential violations of the exclusion restriction can be conjectured for practically all non-randomized instruments, even those generated by natural variation. For this reason we will test the instrument's validity in different ways.

5.7. Qualitative data on loan use

The impact of a microcredit program will among others depend on the use of the loans by the clients. In order to obtain valid and true answers regarding loan use, ample time is reserved to emphasize to the respondent in the introductory explanation prior to commencement of the interview that all information given is dealt with strictly confidential and will not affect their chances of receiving future loans from TLM or any other institution. We believe the answers on loan use reflect reality of most of the respondents' situation. Respondents are asked to give some details about the use of their loan, generating qualitative data in an otherwise primarily quantitative study.

6. Methods

6.1. Construction of wealth index

To address these concerns, we asked households about their ownership of a range of assets, including durable and productive assets (chairs, wardrobe, electronic devices, motorcycle, sewing machine etc.), housing infrastructure (flooring (dirt/non-dirt) and roofing (zinc/non-zinc) and animals (number of pigs, cows, goats (merged with sheep) and poultry). From these ordinal variables we constructed a wealth index. Numbers of animals are count variables and the other asset indicators are of a binary nature. Principal component analysis (PCA) is a feature extraction method often used for construction of socio-economic indices assuming that the first principal component corresponds to the latent concept of "wealth". Since PCA assumes continuously, multivariate normally distributed data, its assumptions are clearly violated in the construction of a wealth index from discrete data. Polychoric PCA, which is an alternative that has become more popular for the construction of wealth indices, assumes the discrete variables to be discretiziations of underlying continuous variables, but still assumes these to be normally distributed (Kolenikov & Angeles 2009). Instead, we opt for categorical principal component analysis. CATPCA consists of two steps. First, it transforms the variables nonlinearly in such a manner that there is maximal data reduction while preserving information. With the option SPORD in the CATPCA command in SPSS 20, this transformation is a smooth monotonic piecewise polynomial of a predetermined degree. The second step is classical PCA to estimate the factors and their loadings. The method assumes neither underlying continuous variables that are multivariate normal nor a linear relationship between them. The wealth index WI_i for the i^{th} household is then obtained by a linear prediction of the variables a_{ik} and the factor loadings w_{ik} from the first factor

$$WI_i = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} a_{ik} \tag{6.1}$$

which can then be interpreted as the linear combination of original variables which captures the maximum amount of information by optimizing the explained proportion of variance (Larrea 2002). (Meulman 1998) offers a detailed explanation of the method.

Supply and demand for assets such as electronic devices relative to supply and demand for other goods can change rapidly in the same setting within a short time span, rendering inter-temporal comparisons invalid (Ferguson et al. 2003). We therefore pooled the data on asset and livestock holdings of both survey rounds to obtain the factor loadings. Table 5 reports the obtained factor loadings, for CATPCA with 2nd degree polynomials with 2 knots. Changing the degree of the polynomials and the number of knots had a negligible impact on the factor loadings. Reassuringly, all factor loadings except poultry are positive, implying that wealth increases with the holding of more of these assets. Poultry means chicken for all but one household, which holds 30 ducks. We merge chicken and ducks into one "poultry" variable, since they have the same FAO livestock unit in Indonesia (FAO 2005). As we want our wealth index to increase with increasing holdings of assets/livestock/social infrastructure, we drop poultry from the construction of our wealth index. In addition, this is the only variable for which we suspect substantial measurement error; bigger animals like pigs, cows and goats are fewer in number and thus more easily correctly counted by the respondent.

Table 5: Factor loadings of the first factor obtained by CATPCA.

Variable	Loadings factor 1 (2nd degree polyn., 2 knots, n=499)		
	With poultry	Without poultry	Round 1, exc. Livestock
Mobilephone	.500	.500	.508
Housephone	.212	.212	.220
Chairs	.101	.101	.103
Wardrobe	.387	.387	.392
Desk	.410	.410	.404
Clock	.490	.490	.491
Ricecooker	.582	.582	.587
Iron	.647	.648	.656
Fridge	.569	.569	.574
Oven	.689	.689	.693
Fan	.499	.499	.505
Sewingmachine	.488	.488	.484
Radio	.366	.366	.361
Television	.712	.712	.721
Dvdplayer	.693	.693	.706
Computer	.514	.514	.504
Bicycle	.507	.507	.501
Motorcycle	.532	.532	.536
Car	.464	.465	.443
Camera	.448	.448	.433
Gaslamp	.296	.296	.282
Jewelryfromgold	.585	.586	.590
Pig	.318	.318	-
Cow	.247	.248	-
Goat	.147	.147	-
Poultry	-0.64	-	-
Floor	0.506	.505	.512
Roof	0.407	.406	.407
Cronbach's alpha	.872	.873	.873
% of variance accounted for	22.458	23.277	25.554

When seen as a proxy for wealth, the index contains little reporting error, but still may contain measurement error, since it is an imperfect measure of the underlying latent variable wealth. We conjecture that for a measure of household welfare, it is a less imperfect measure than consumption or income, since the discrete nature of the asset ownership indicators makes them unlikely to be subject to systematic reporting errors (cf. the discussion in the Data section 5.3 and 5.4).

To see the impact of TLM's group credit on livestock, we also constructed a livestock index using the FAO livestock unit conversion factors of Indonesia (FAO 2005) as factor weights. Here we did not opt for using empirical weights by CATPCA since the number of variables is only 4, potentially giving rise to the problems of 'clumping' and 'truncation' (Vyas & Kumaranayake 2006). Clumping is the grouping together of households into a small number of clusters. The resulting index will then not be able to differentiate sufficiently between different households. Truncation refers to a situation where the asset level is too equal to differentiate between differentiate between different socio-economic groups, for instance the moderately poor and the very poor. The FAO livestock unit conversion factors are: poultry cow 0.65, pig 0.25, goat and sheep 0.1 and poultry 0.01 (FAO 2005).

The other type of response of interest is anthropometric. To check the accuracy of the height measurement, we check the difference in height of women between the first and second round. For adult women, this difference should be negligible. However, only 105 out of the 154 women for which we had a height measurement in both rounds had a height measurement difference of 1 cm or less and 109 a difference of 1.5 cm or less. The fixation and proper use of the height tape is easier with two enumerators, as was done in the first round. Since in addition one of the authors took the height measurements of the women of the first 207 households in the sample, we limit our analysis with women's BMI to that subsample. For the construction of the BMI variable, we combine the height value of the first round with the weight value of the second round for this subsample, since the weighing scale is easy to use and thus likely measured the subjects' weight accurately.

6.2. Spillover effects

Through the selection of control households from the same or neighbouring RT (smallest administrative unit) as the treatment households, bias within the sample due to non-random loan branch placement is eliminated. The drawback of drawing treatment and controls from the same small administrative unit is that of spillover effects, violating one of the basic assumptions of causal inference, namely the Stable Unit Treatment Value Assumption (SUTVA). The SUTVA is a necessary condition for identification of various treatment effects. It states that the potential outcomes of observation i do not depend on the treatment assignment of other observations in the population, thereby ruling out treatment externalities and general equilibrium effects. Given potential outcomes $Y_i(d,z)$, treatment D_i, binary instrument Z_i and additional covariates X_i, for our IV models SUTVA can mathematically be stated as (Frölich 2007),

$$Y_i(d,z) \perp (D_j, Z_j) \text{ and } D_i(z) \perp Z_j, \quad \forall j, d \epsilon \{0,1\}, \text{ and } z \text{ in the support of } Z \qquad (6.2)$$

There are a number of possible spillover effects. First, in a marginal business environment with limited demand, the expansion of a borrower's business may crowd-out non-borrowers' neighbouring businesses. When these non-borrowers are poorer (richer) than the borrowers, this may increase (decrease) the depth of poverty in the area.

Since around 17% of the borrowers in our sample utilized their loans for educational expenditures, there may be long-term dynamic effects not captured by our study. When these children of the borrowers graduate and obtain better-paying jobs, they may generate employment income. Educational spillovers have been documented empirically by (Rudd 2000). Educational spillovers operate over the long term and are thus not likely to bias our short-term impact estimates. It highlights however both the need for longer-term follow-ups and the limitations of partial equilibrium analyses like ours.

For the households investing their loans into productive activities, the increased (decreased) household income from profitable investments may lead to increased (decreased) transfers from borrowers to poor non-borrowers and increased (decreased) demand for goods and services provided by non-borrowers. To assess these possibility, we asked the control household in the same or neighbouring RT of the treatment household whether they in the past (name of variable between brackets)

(1) have received cash from the treated household(s) (*spillover2*);
(2) have received business advice from the treated household(s) (*spillover3*);
(3) sold goods to the treated household(s) (*spillover4*).

The response is recorded on a 4-point likert scale reflecting the frequency of these exchanges: never, seldom, sometimes or often. For future studies, it would be useful to also ask non-borrowers whether they consumed meals in the borrowers' house, as that could have an effect on their health and work productivity.

What we have discussed are intra-village spillovers, however, there may also be regional or national level general equilibrium effects. When changes in income from credit-induced business activities affect aggregate demand for food, this may lead to increased income for the agricultural sector, potentially even in other parts of Indonesia or even outside the country. But, given the small-scale of the program, with around 21,000 borrowers in 2010, these effects are not likely to be very strong, if at all present.

6.3. Attrition and missing data: theory

Our sample is characterized by high attrition: out of the 299 first-wave households, only 209 were re-interviewed in wave 2 – a 30% attrition rate. Three types of missing data can be distinguished: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), see table 6.

Table 6: Types of missing data.

Abbreviation	Notion	Mathematically	Description
MCAR	Missing completely at random	$P(r\|o,u) = P(r)$	The missingness mechanism does not depend on observables nor unobservables
MAR	Missing at random	$P(r\|o,u) = P(r\|o)$	Given the observed data, the missingness mechanism does not depend on unobservables
MNAR	Missing at random	$P(r\|o,u) \neq P(r\|o)$	Given the observed data, the missingness mechanism still depends on unobservables

Attrition results in a missing outcome. Some households dropped out of the sample due to hospital visits or being away from home for some time. In other cases, household members migrated, which is probably non-random. Considering the push factors (drivers) of rural-urban migration, (Hare 1999) suggested that higher productive capital accumulation by households increases their work productivity at home and increases their urban reservation wage. She also pointed out that the presence of imperfectly functioning capital markets, wealthier household may find it easier to bear the cost of migration. Education levels and skill sets may affect the expected wage elsewhere and thereby affect the migration propensity. Finally, increased access to social networks is likely to *ceterus paribus* lower a households' propensity to migrate.

Given missing data, there are at least three ways to proceed. First, a complete case analysis using only observations that have no missing data rests on a MCAR assumption. A second method is to impute the missing values trough a model prediction using the observed data as input. Given that the predictive model is correctly specified and that the data are MAR, this yields unbiased estimates. However, the standard errors of estimations using a dataset obtained by single imputation are generally too optimistic (i.e. too narrow), since this method does not account for the uncertainty generated in the prediction step. (Rubin 1987) termed such imputation methods that do not account for the extra uncertainty resulting from the prediction improper. Indeed, (Hens 2005) showed via Monte Carlo simulations that nonparametric single imputation is unbiased but overstates statistical precision, reflected in too narrow confidence intervals. In contrast, for multiple imputation the researcher randomly samples a number of copies from the predictive posterior distribution and imputes these copies to generate multiple datasets. Multiple imputation is thus a proper method in Rubin's sense. The standard errors obtained by averaging the estimates from these datasets using an averaging formula are consistent.

6.4. Dealing with attrition and missing data

All our covariates are measured in round 1; the outcomes are measured in round 2. We assume that any missing covariate value is a human error by the enumerators and thus MCAR. We thus exclude observations with missing data from the set of covariates we use for our treatment effect models. This creates a small bias if some of those observations also have missing outcomes, when the outcomes are in fact MAR, as we assume they do. Given that we do not want to introduce bias in our estimates by assuming a parametric form for the predictive model, and given that to date there does not exist a software package that does multiple imputation without assuming a parametric functional form while allowing for both continuous and categorical predictors, we restrict ourselves to nonparametric single imputation and complete case analyses. In the Results section, we show that there are indeed statistically significant differences in means of the attrited and non-attrited subsamples, implying non-random attrition.

There exists no software package for nonparametric single imputation, so for the wealth index outcome we proceeded as follows. (1) We fitted a nonparametric predictive model for the outcome using multivariate kernel regression (we explain below what kernel regression is) on the complete case sample (i.e. the subsample of observations with observed outcomes) using the R package –np- due to (Hayfield & Racine 2008). (2) On a sample containing only the observations for which the outcome is missing, we predicted the outcome using the model fitted in step (1). And (3) We use those predicted outcomes as imputations for the missing outcomes[iv]

Kernel regression is one often-used nonparametric regression method used to smooth a regression curve over sample data. In our imputation, we use a local linear smoother. The word "local" in nonparametric regression just means that the function (in this case a linear one) is fitted in a small neighbourhood of the data rather than globally (i.e. for the whole sample of data points). As we have women BMI data in round 2 for only 163 women, any imputation method will likely perform poorly. We therefore employed only a complete-case analysis for women's BMI as outcome.

6.5. Sampling bias

As we described in section 5.1, the departures from random selection in our sampling scheme may limit the external validity of our research. It should, however, not have a bearing on the internal validity of the study. It is useful here to distinguish between two sources of bias: non-random sample selection by the researcher (affecting external validity) and self-selection into treatment status within our sample (affecting internal validity). We discuss the former here and the latter in section 6.7.

The treated households were most likely non-randomly selected from the reference population (the population of the three Kabupaten in the province NTT, see introduction) because of (de Aghion et al. 2007)

1) Non-random program placement: the decision of TLM of where to place the branch may be related to factors also influencing household economic status, such as distance to a city or closeness to an existing branch.
2) Non-random client recruitment: loan officers have targets of how many new clients they need to recruit per month. The decision in which village and in which RT to market the loan product may depend on the potential creditworthiness of households, which is likely to be directly or indirectly related to their economic status. Personal preferences of loan officers may lead them to market the organization and its products in areas that deviate systematically from the population averages on relevant characteristics.
3) Self-selection by clients into treatment based on household social, health and economic factors, the stocks of physical, human and social capital, personality traits, preference sets, etc.

The control households were also not strictly randomly selected, as explained in the Data section. The sources of bias stemming from the non-random selection of community controls are described in the Data section. Given the non-random sample selection, we claim to identify only sample versions of the various treatment effects, and express caution when extrapolating the findings to different geographical areas of the these Kabupaten, since the above three sources of non-random sample selection may make our sample different from the reference population on aspects that affect the return to TLM's group loan product.

6.6. Modelling assumptions

We distinguish between the following types of structural assumptions:

- Functional form assumptions: linearity of effects of observables, additivity of the effects of observables, additivity of the effects of observables and unobservables, monotonicity of the error term;
- Distributional assumptions: normality of the error term;
- Causal assumptions: exogeneity, unconfoundedness, instrument independence;
- Other assumptions: completeness, rank conditions.

Non-structural assumptions include common support and regularity conditions such as continuity and smoothness of regression functions and surfaces. Given how controversial the topic of microcredit's poverty impact is, we do not want our results to be driven by overly restrictive assumption rather than by the data itself. To fix ideas, we focus here on the first type of assumptions, those that restrict the functional form of the assumed data generating process.

To see which functional form assumptions are typically invoked, consider first the restrictive model

$$Y = \sum_{k=1}^{K} \beta_k X_k + \varepsilon \tag{6.3}$$

where Y is the outcome, β_k is a vector of k coefficients, X is a matrix of covariate data with k columns (the number of covariates) and ε the error term. This model assumes that the effect of each covariate is linear and that the effect of the covariates is additive. When one also assumes exogeneity ($E[u_i|x_1, \ldots, x_n] = 0$), estimation by least squares is possible. OLS assumes that the effect of the covariates is the same across observations, i.e. effect homogeneity. A random coefficient model relaxes the assumption of effect homogeneity and assumes that each observation has its own, random effect.

The additive model

$$Y = \sum_{k=1}^{K} f_k(X_k) + \varepsilon \tag{6.4}$$

drops the linearity assumption of the covariates, and replaces it with a weaker smoothness assumption on f_k. It still assumes that the effects of the observables are additively separable.

The nonparametric model

$$Y = f(X) + \varepsilon \tag{6.5}$$

allows for arbitrary interactions among predictors. The relationship between the predictors and the outcome is still deterministic, however, as interactions with unobservables are assumed away.

The nonseparable model

$$Y = f(X, \varepsilon) \tag{6.6}$$

is still stochastic after conditioning on X, as effects of the observables are allowed to depend arbitrarily on the effects of the unobservables contained in ε determining Y. Assuming unconfoundedness of a binary component of X, usually called the treatment D, i.e. $Y^D \perp D|X$ (Y^D being potential outcomes), various treatment effects of D are identified by a range of nonparametric matching methods, including propensity score matching. Unconfoundedness holds when selection into treatment is only on observables, i.e. when all common causes of D and Y are included in X. An added advantage of a nonseparable model is that if one is interested in the effect of treatment D, its effect estimate will not be biased if other covariates X are endogenous, as long as D itself is exogenous. In parametric regressions, endogeneity of control variables potentially biases coefficient estimates on all covariates, including the exogenous ones (Frölich 2008). Often it is necessary to control for endogenous control variables in order for the unconfoundedness assumption to hold; this underscores the utility of nonparametric matching methods when unconfoundedness is likely to hold.

6.7. Testing for selection bias under H$_0$: no treatment effect

Within our sample, there is likely to be nonrandom self-selection by households into treatment status. When this source of endogeneity is not accounted for, it may bias estimates of sample average treatment effects. The sources of this bias are the same as described above for nonrandom sample selection, except for non-random program placement. Within the sample, there is no bias from non-random program placement, considering that we sampled controls from the same or neighbouring RT's at a ratio of 1:1.

One way to test for selection bias is to estimate treatment effects on outcomes that cannot have been affected by the treatment, such as pre-treatment outcomes. Under the null of no selection bias, the treatment effect should then not be statistically significantly different from zero (Rosenbaum 1984). This approach can refute the null of no selection bias, but cannot confirm it. One scenario in which selection bias is present but not detected by the above test is when by coincidence the pre-treatment outcomes are similar across treatment regimes but there is dynamic selection bias, i.e. unobservables that affect the pre-post treatment time trend in outcomes differently across treatment regimes.

As explained above, nonparametric matching methods are preferred over regression when unconfoundedness holds. When the covariates in propensity score matching or Mahalobonis matching have ellipsoid distributions (f.i. normal or t-distributed), then each covariate reduces bias equally, a property named Equal Percent Bias Reduction (EPBR) by (Diamond & Sekhon 2006). Such property may not be desirable, consider for instance the data generating process $Y = X_1^4 + X_2$. In such case, bias reduction in X_1 is more important than in X_2, and X_2 should have a higher weight. Genetic matching uses an evolutionary search algorithm developed by (Sekhon & Mebane Jr 1998) to determine the weight given to each covariate in order to maximize covariate balance across treatment and control groups. Simulations by (Diamond & Sekhon 2006) showed that genetic matching performs better than matching based on the propensity score or Mahalobonis distance, both when EPBR holds and when it does not hold.

6.8. Worst case bounds on SATE

The worst-case bounds on the average treatment effect, which are mostly too wide to be informative, are given by (Manski 1990)

$$\begin{gathered} E[Y(1)|X, D = 1]P(D = 1|X) \\ \leq E[Y(1)|X] \leq \\ E[Y(1)|X, D = 1]P(D = 1|X) + P(D = 0|X) \end{gathered}$$

and

$$\begin{gathered} E[Y(0)|X, D = 0]P(D = 0|X) \\ \leq E[Y(0)|X] \leq \\ E[Y(0)|X, D = 0]P(D = 0|X) + P(D = 1|\square) \end{gathered}$$

The resulting bounds on the average treatment effect are

$$\begin{gathered} E[Y(1)|X, D = 1]P(D = 1|X) - E[Y(0)|X, D = 0]P(D = 0|X) + P(D = 1|X) \\ \leq E[Y(1)|X] - E[Y(0)|X] \leq \\ E[Y(1)|X, D = 1]P(D = 1|X) + P(D = 0|X) - E[Y(0)|X, D = 0)P(D = 0|X) \end{gathered}$$

6.9. Nonparametric identification and estimation of SLATE

One popular solution to the endogeneity problem in econometrics is the use of an instrumental variable Z. Consider the nonseparable model

$$\begin{gathered} Y = f(D, X, U) \\ D = g(Z, X, V) \end{gathered} \qquad (6.7)$$

where D is the treatment, X is a set of additional covariates, Z is the instrument. U and V are unobservables capturing all other common causes of (D, X) and Y. Endogeneity arises due to the dependence of U and V and the dependence of D and U. The nonseparability of this model is an attractive feature, as it allows X to depend on U and V (Frölich 2008). Controlling for endogenous control variables may be necessary for the exclusion restriction to hold. Note that the model is triangular or recursive, as right-hand side variables can affect left-hand side variables, but not the other way around. (Frölich 2007) showed that the local average treatment effect is identified under a mean instrument independence condition

$$E(Y(d,1)|\tau = t) = E(Y(d,0)|\tau = t) = E(Y(d)|\tau = t) \\ \text{for } d\epsilon\{0,1\} \text{ and } t\epsilon\{at, c, d, nt\} \tag{6.8}$$

where τ is the households' type, as categorized in section 5.6. The Local Average Treatment Effect is the effect on the average observation for which the instrument induces a change in treatment status. We will test this assumption in section 6.10. Another assumption necessary is monotonicity of the first stage,

$$\Pr\big(D(1) \geq D(0)\big) = 1 \tag{6.9}$$

This assumption states that the potential treatment state of any observation does not decrease in the instrument, ruling out the existence of defiers (type d). Therefore, the LATE reduces to a Complier Average Causal Effect (CACE). Another, more innocuous assumption is that the relative size of the subpopulations of always-takers, compliers and never-takers are independent of the instrument,

$$P(\tau_i = t|X_i = x, Z_i = 0) = P(\tau_i = t|X_i = x, Z_i = 1) \;\text{ for } t\epsilon\{a, n, c\} \tag{6.10}$$

Nonparametric regression in general suffers from the curse of dimensionality: the number of data points needed to obtain a reasonably small variance increase rapidly with the number of covariates included, since the regression curve is fitted locally (in a small neighbourhood of the data) and thus uses few data points to fit the curve. Recall that most estimators are asymptotically (that is if n tends to infinity) normally distributed. The rate at which this happens is called the rate of convergence of an estimator. OLS is a root-n-consistent estimator of regression coefficient $\hat{\beta}$ because

$$\sqrt{n}\big(\hat{\beta} - \beta\big) \to N(0, \sigma^2\Sigma_{xx}^{-1}) \tag{6.11}$$

where → means 'is asymptotically distributed as' and $\sigma^2\Sigma_{xx}^{-1}$ .is the variance-covariance matrix. The variance-covariance matrix is just a matrix with the variance of the variables on the diagonal and the covariance between the variables as the remaining elements of the matrix. In contrast, the rate of convergence of nonparametric regression is typically slower and decreases as the number of regressors increases, because the regression curve is fitted locally rather than globally. Multivariate kernel regression for instance converges at a rate of $n^{-2/(4+d)}$ when n goes to infinity, d being the number of covariates. Nonetheless, (Frölich 2007)'s nonparametric IV LATE estimator based on the propensity score achieves root-n-convergence. The estimator is thus equally efficient as OLS, making it very attractive for applied researchers (including ourselves) seeking to infer useful information from finite samples. The math in (Frölich 2007) is rather heavy, so for details and derivations we refer to his original paper.

When it comes to covariate selection, those variables that need to be conditioned on in order for the exclusion restriction to hold must be included. Thereafter, (Frölich 2007) explains that the inclusion of extra covariates that are not needed for instrument validity assist to lower the variance of his estimator. When the covariates contain too much noise however, they may increase rather than decrease variance. This logic of selecting covariates to maximize statistical precision given that instrument validity holds guides our selection of covariates that are not needed for the exclusion restriction to hold.

A last note on a choice that has to be made by the practitioner when implementing Frölich's estimator; in his estimator, $E(D|Z,X)$, the conditional treatment given the instrument and covariates, and $E(Z|X)$, the conditional instrument given the covariates have to be estimated. We choose to use local logit for these estimations (i.e. a logit estimation in a certain neighbourhood of the data), since (Frölich 2006) found via Monte Carlo simulations that this nonparametric binary regression method performs strictly better in terms of precision (i.e. variance) compared to other nonparametric binary regression methods. When performing Monte Carlo simulations, multiple synthetic datasets are stochastically generated (i.e. via the programming code specified data-generating process containing random noise), and averages of the estimates of these replications reveal the estimators' finite-sample properties, such as bias and variance.

6.10. Testing the mean exclusion restriction

Exclusion restriction is a key assumption enabling identification of interesting causal effects in the presence of endogeneity. It is imperative that we test this identifying restriction. For the nonparametric identification of LATE with a binary instrument, (Huber & Mellace 2011) derived four testable moment inequalities implied by the mean exclusion restriction. He proposes to use the minimum p-value type test by (Bennett 2009) for joint inequality moment constraints. Higher testing power can be achieved by imposing the assumption that mean potential outcomes of always takers and compliers under treatment and of never takers and compliers under non-treatment are equal. We will discuss the plausibility of the latter assumption in the Results section.

6.11. Nonparametric identification and estimation of unconditional SQTE

Besides the effect of a treatment on the mean counterfactual outcome distribution, interest may also lie in the effects of the treatment on quantiles of the dependent variable, i.e. the quantile treatment effect (QTE). Assume that the quantiles are unique and well-defined, that there exist compliers but no defiers, and that the common support condition is satisfied

$$0 < p(X) < 1 \quad a.s. \quad \text{where } p(X) = P(Z = 1|X = x) \tag{6.12}$$

We believe the assumption of no defiers to be plausible in our sample, given the lack of negative stories told by borrowers during interviews. Assume in addition full instrument independence

$$(Y^D, D) \perp Z|X \tag{6.13}$$

Then the QTE is identified in the nonseparable model (6.7) (Frölich & Melly 2008). Note that since $A \perp B|C$ and $A \perp D|B,C$ implies $A \perp (D,B)|C$ (Dawid 1979), assumption (6.13) implies (a) a full exclusion restriction

$$Y^D \perp Z|X \tag{6.14}$$

And (b) an unconfounded instrument

$$\tau \perp Z|Y^D, X \tag{6.15}$$

Neither of these conditions is directly testable, since they contain potential outcomes, which are unobserved (a household can only be treated or non-treated, it cannot in both states at the same moment in time). However, in the next section, we show that it is possible to have a test that can reject (6.14). Under the aforementioned assumptions, (Frölich & Melly 2008) proposed a root-n consistent estimator based for the unconditional QTE on the propensity score. That is the same rate of convergence as OLS; it is therefore a pretty efficient estimator.

6.12. Limitations of existing instrument validity tests

The problem of existing instrument exogeneity tests for linear IV models such as the widely used (Sargan 1958), (Hansen 1982), (Anderson & Rubin 1949) and (Kleibergen 2002) overidentifying restriction tests is that they are built on the assumption that there are a number of undisputed valid instruments, at least as great as the number of endogenous variables to be instrumented (Verbeek 2008). Many econometrics textbooks, including the widely used introductory (Wooldridge 2009) and (Gujarati 2003) even the advanced level (Wooldridge 2010) do not state this important fact; they only state that a the instrument count needs to exceed the number of endogenous variables. However, if a researcher already has available a number of undisputed valid instruments to begin with, then the identification problem is solved, and the test will only add value if the researcher wants to increase the instrument count to increase statistical precision. Note then that in general, when essential heterogeneity is present, different instruments will generate different estimates (LATE) and LATE estimates with multiple instruments will typically be less straightforward to interpret. At a deeper level, the assumption that the above overidentifying restriction tests aim to test is not sufficient to give the IV estimate a causal interpretation (Reiss 2005). For that the two possibilities depicted in figures 3 and 4 have to be ruled out.

Instead, the problem when one resorts to instrumental variable estimation is the identification problem itself, i.e. finding at least one valid instrument for each regressor. Unlike some econometrics textbooks do appear and unlike many applied papers in top economic journals seem to suggest, the above instrument validity tests will not in general be consistent (Doko & Dufour 2008). Those instrument validity tests are thus not identification robust

6.13. A new test of the full exclusion restriction

Here we propose a new way of testing the full instrument independence assumption. To the best of our knowledge, no test of the instrument validity condition (6.14) has been proposed in the literature yet. If Z causes Y exclusive of (D, X) (possibility 1) and/or Z and Y share an unobserved cause U (possibility 2) then the (6.14) does not hold and the instrument is invalid.

(Chalak & White 2012) proved in a settable systems framework an interrelationship between causality and probabilistic conditional independence that is known as the Conditional Reichenbach Principle of Common Cause. Their paper is very formal, but their main result can be stated in simplified form as follows:

Proposition 1. If (6.14) does not hold, i.e. if Z and Y are conditionally dependent given (D, X), then either

(i) Z causes Y exclusive of (D, X)
(ii) Z and Y share an unobserved common cause U
(iii) Y causes Z exclusive of (D, X).

The last possibility is ruled out by assuming triangularity of the (6.7); it is trivially ruled out in our application, since the outcome Y is measured later in time than the instrument Z, so the former can never affect the latter. Therefore, a rejection of $H_0: Y \perp Z|(D, X)$ implies a violation of the exclusion restriction (6.14). $H_0: Y^D \perp Z|X$ is thus refutable, but it is not confirmable since there exist possible violations of the exclusion restriction that are not detectable by testing $H_0: Y \perp Z|(D, X)$. Figure 5 depicts an example where Z causes Y exclusive of (D, X), yet $Y \perp Z|(D, X)$ does not hold. In this case different causal pathways from Z to Y cancel out each other, rendering Z and Y independent given (D, X).

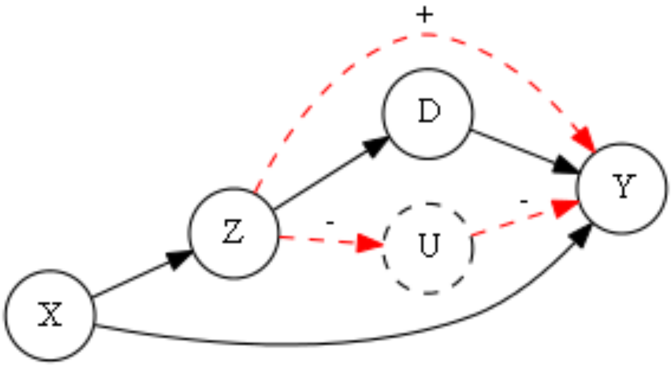


Figure 5: two causal pathways from Z to Y cancelling out, rendering Z and Y independent conditional on (D, X).

An example of Z and Y sharing an unobserved common cause U even though Y and Z are independent conditional on (D, X) can also easily be constructed. Take the situation depicted in figure 4, with Z and Y being normally distributed with mean zero and variance 1 (our instrument is binary, but the present discussion holds irrespective of the nature of the data). Then Z and Y are independent conditional on (D, X) if and only if the correlation between Z and Y is zero.

In sum, a rejection of $H_0: Y \perp Z|(D, X)$ implies that $Y^D \perp Z|X$ does not hold, meaning that the exclusion restriction does not hold. We can thus refute the null hypothesis of instrument independence EQ. While the consequences of rejection of $H_0: Y \perp Z|(D, X)$ are clear, one must be careful when failing to reject. $H_0: Y \perp Z|(D, X)$ can hold even though the exclusion restriction does not. It is therefore imperative to test the exclusion restriction in multiple ways. Another way to test it is to estimate treatment effects on pre-treatment outcomes, in the same way as selection bias can be tested section 6.7

The question remains how to test $H_0: Y \perp Z|(D,X)$. We exploit the fact that $Y \perp Z|(D,X)$ if and only if $f(Y|D,X,Z) = f(Y|D,X)$, where f is a probability density. The intuition is that Z should not contain any information about Y once (D,X) have been accounted for. For Z being multinomial taking on c values, $H_0: f(Y|D,X,Z) = f(Y|D,X)$ is equivalent to

$$H_0: f(Y|D,X,Z=0) = f(Y|D,X,Z=1) = \cdots = f(Y|D,X,Z=c) \qquad (6.16)$$

which can be written as

$$H_0: g_0(Y|D,X) = g_1(Y|D,X) = \cdots = g_c(Y|D,X,Z=c) \qquad (6.17)$$

Hence, by splitting the sample based on the values taken on by the instrument, we obtain a testable condition that contains the same variables across densities whose equality is to be tested. The test can thus be cast in the framework of the nonparametric test of equality of conditional densities of (Li et al. 2009). The theoretical underpinnings of the test statistic of (Li et al. 2009) directly apply.

Their paper describes how to estimate the conditional densities with mixed data by kernel smoothing both the continuous and categorical variables and selecting the bandwidth by means of cross-validation. The bandwidth is just a smoothing parameter. Intuitively, if set very high, the curve will oversmooth the data and incur a high bias, while a bandwidth chosen to small results in high variability, since only few data points are used to fit the curve. This is known as the bias-variance trade-off of bandwidth selection in nonparametric kernel density estimation and regression. Rather than using a rule-of-thumb bandwidth, we select the bandwith in a data-driven way by minimizing the mean squared error, which is the squared sum of bias and variance.

Our new way of testing instrument validity in nonseparable triangular models based on the nonparametric test of equality of conditional densities of (Li et al. 2009) was proposed to Jeffrey Racine of McMaster University, one of the authors of the latter paper. He showed interest in these ideas and conducted Monte Carlo simulations to investigate the finite-sample properties of the test for a continuous outcome and a multinomial instrument. The results of these simulations, reported in appendix for the case of n=250 and a binary instrument, are encouraging: they show that the test is correctly sized and power that increases monotonically with a mean shift in the conditional outcome and with the sample size. The size of a test is the frequency it does not reject H_0 when it is true, while the power of a test refers to the frequency it rejects H_0 when it is not true. Given our small sample size, the power of the test is not very high and we have to be careful when interpreting a moderate p-value as a lack of evidence against the null. A small p-value can unequivocally be interpreted as evidence against the null of instrument validity.

6.14. Identification without an exclusion restriction: Heckman's BVN and Millimet's MB & MB-BC estimators

In order to increase the robustness of our findings to the identifying restrictions, we also deploy alternative estimators that do not rely on an exclusion restriction for identification. These estimators include Heckman's Bivariate Normal model (BVN) (also referred to as 'Heckman model') and (Millimet & Tchernis 2012)'s Minimum-Biased and Minimum-Biased-Bias-Corrected (MB-BC) estimators. The identified object is not the same as when we relied on an exclusion restriction for identification: here we identify and estimate the Sample Average Treatment Effect (SATE).

We discuss here the two-step version of the BVN model due to (Heckman et al. 1999). They showed that in a linear selection model, selection bias can be viewed as an omitted variable problem. To see this, let y, x and d be defined as usual and consider the model

$$\begin{aligned} &y = x'\beta + \gamma d + \delta r + \varepsilon \\ &d^* = x'\alpha + \eta \\ &d = I(w > 0) \end{aligned} \tag{6.18}$$

where $r = \varepsilon(1) - \varepsilon(0)$ are unobserved relative advantages of the treatment group and it is assumed that $\varepsilon(0) = 0$ and that conditional on x, the variables $\varepsilon(1)$ $\varepsilon(0)$ are independent of d. Let $p(x) = \Pr\,(d = 1|x)$ denote the propensity score. When assuming joint normality of η and ε, and denoting $\sigma_\eta^2 = 1, \sigma_\varepsilon^2, \rho$, their respective variance and correlation, (Heckman et al. 1999) showed that the reduced form is equal to

$$\begin{aligned} E(y|x,d) &= x'\beta + \gamma d + \rho\sigma_\varepsilon^{-1}\big[\lambda d - \tilde{\lambda}(1-d)\big] + E(r|x^*, d=1)\times p(x^*) \\ &= x'\beta + \gamma d + \rho\sigma_\varepsilon^{-1}\big[\lambda d - \tilde{\lambda}(1-d)\big] + E(r|p(x^*))\times p(x^*) \end{aligned} \tag{6.19}$$

where $\lambda = \phi(x'\alpha)/(1 - \Phi(x'\alpha))$ and $\tilde{\lambda} = \phi(x'\alpha)/\Phi(x'\alpha)$, with $\phi(.)$ and $\Phi(.)$ denoting respectively the normal pdf and cdf. The specific form of treatment effect heterogeneity that arises when $E(r|p(x^*)) \neq 0$, was termed essential heterogeneity by (Heckman et al. 2006). It occurs when observations have partial or full knowledge of their idiosyncratic treatment response, and this knowledge affects their self-selection propensity. In the presence of essential heterogeneity, the Heckman model and linear IV estimation break down. When $E(r|p(x^*) = 0$, we have the BVN selection model and γ in the outcome equation of (6.18) can be estimated without bias using OLS by including the omitted variable $\rho\sigma_\varepsilon^{-1}\big[\lambda d - \tilde{\lambda}(1-d)\big]$ (called the Inverse Mill's ratio) to the outcome equation. Hence, theoretically, the BVN model is identified from functional form alone. Many researchers, including (Bushway et al. 2007; Puhani 2000). have noted however, that without an exclusion restriction, the BVN model is poorly identified, since it heavily relies on the normality of the tails of the error term for identification. Moreover, multicollinearity between the covariates and the Inverse Mill's ratio may be a problem.

In the presence of essential heterogeneity, the BVN model is inconsistent. The mathematical derivations of their results are not particularly revealing. The intuition behind the minimum-baised (MB) estimator is that there is a propensity score, estimated from the data that minimizes the bias when unconfoundedness fails. The MB estimator uses only observations in a small radius around that propensity score. Again there is a bias-variance trade-off: the smaller the radius, the smaller the bias and the larger the variance. When the researcher is willing to assume the functional form of the selection model (6.18), then an expression for the bias of the MB estimator can be derived. The bias correction also accounts for essential heterogeneity, unlike the BVN model. Both the BVN and Millimet's estimators were implemented in STATA SE12 using Millimet's –bmte- ado-file[v].

7. Results

7.1. Descriptive statistics

Table 7 reports summary statistics of key continuous and ordered multinomial variables, including the differences across treatment regimes and their statistical significance. For this, we employ the nonparametric Kruskal-Wallis rank test, which only assumes that both samples (treated and controls) are independently drawn from the same population. Statistically significant differences in averages between subsamples defined by treatment regime are found on pre-treatment outcome (*wealthindex1*), *landarea* and social network related indicators, including the main instrument *iv1*.

Table 7: Descriptive statistics of key continuous variables.

Var. Name	# of obs.	Min.	Mean	Max.	S.D	Mean Treated	Mean Control	P-val. Diff.[1]
Variable name								
(Outcomes)								
wealthindex2_ imputed	292	0	3.79	42.1	3.2	4.0	3.6	0.64
wealthindex2	209	0	3.91	42.1	3.7	4.1	3.7	0.74
BMI_woman2	122	14.8	22.0	45.0	4.5	21.9	22.1	0.66
foodconsweek2[2]	209	51.5	212.4	1529	157.6	212.7	212.2	0.58
consmonth2[2]	209	468	1548.6	18452	1632.9	1521.5	1576.5	0.99
(Instrument)								
iv1	292	0	3.38	30	4.10	4.8	2.15	0.0001
(Control vars.)								
Hhsize	299	1	5.40	20	2.62	5.5	5.3	0.58
Age	298	22	43.7	78	11.7	43.1	44.3	0.41
Educyears	299	0	43.7	19	11.7	8.0	7.8	0.46
Rspm	299	0	7.5	16	3.2	7.5	7.5	0.75
Finlit	283	0	1.5	5	1.2	1.3	1.7	0.07
Friends	292	0	6.4	90	7.2	5.5	7.1	0.51
Friendsinfo	293	0	2.9	20	2.7	3.0	2.8	0.0072
Friendsgoodinfo	288	0	2.0	15	1.9	2.0	2.0	0.098
Landarea	297	0	0.7	10	1.3	0.5	0.8	0.0022
wealthindex1	296	0	3.21	10.484	2.24	3.47	2.98	0.08

[1] Kruskal-Wallis test of difference of group means with df=1.

[2] In thousands of Rupiah.

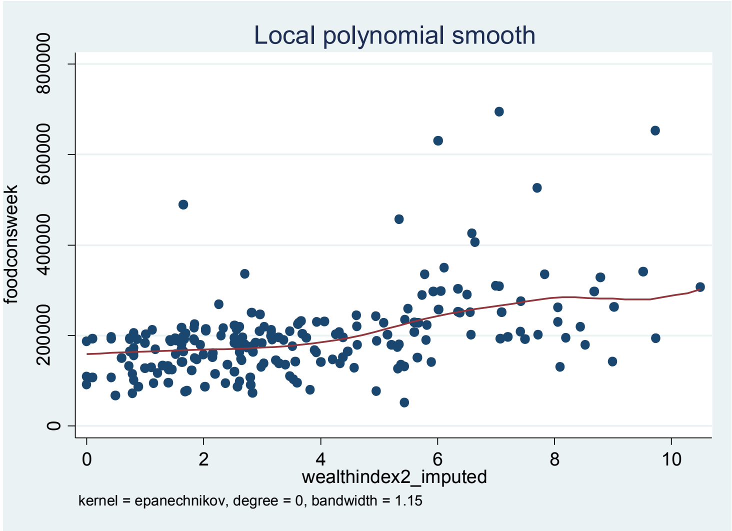


Figure 6: A positive relationship between *wealth2_imputed* and *foodconsweek*, as can be expected.

Treated households were asked for their main loan use retrospectively in survey round 2. As can inferred from figure 7, a majority of treated households reported productive investments as their main loan use. Examples are farmers buying agricultural inputs including seeds, fertilizer and pesticide. Borrowers who are motorcycle taxi-drivers (ojeks) used loan funds to obtain a driving license, pay credit of their motorcycle, or buy spare parts to repair their motorcycle. Owners of kiosks bought inventories. We also found one borrower who used his loan to obtain the letters to become civil servant (counted as productive use), buy land (one household).

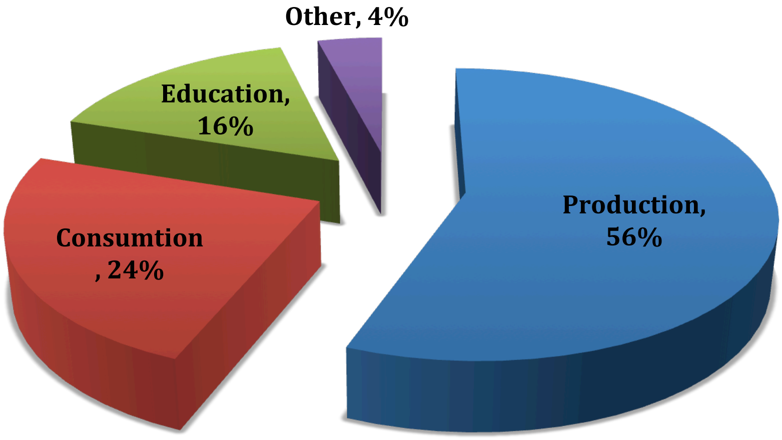


Figure 7: Distribution of main loan use of TLM's group loan (*mainloanuse2*) among treated households as reported by them post-treatment.

About a quarter of borrowing households utilize the loan for consumptive purposes, most often a wedding or funeral ceremony of a family member. Others seem to smooth consumption by seeking credit and buy household necessities from the loan. Of the treated households, 16% uses their loan mainly for education of their children, which typically means paying school fees.

7.2. Checking for outliers

Many regression methods, including linear and nonparametric kernel regression, are highly sensitive to anomalous observations since they are based on a (local) least squares approach (Boente et al. 2009). This results from the fact that when minimizing least squares of residuals, a large residual value squared will be very large and have much influence on the estimate and its standard error. To check for outliers, we plotted the kernel densities of the outcome variables, illustrated in figure 8 for *wealthindex2_imputed*.

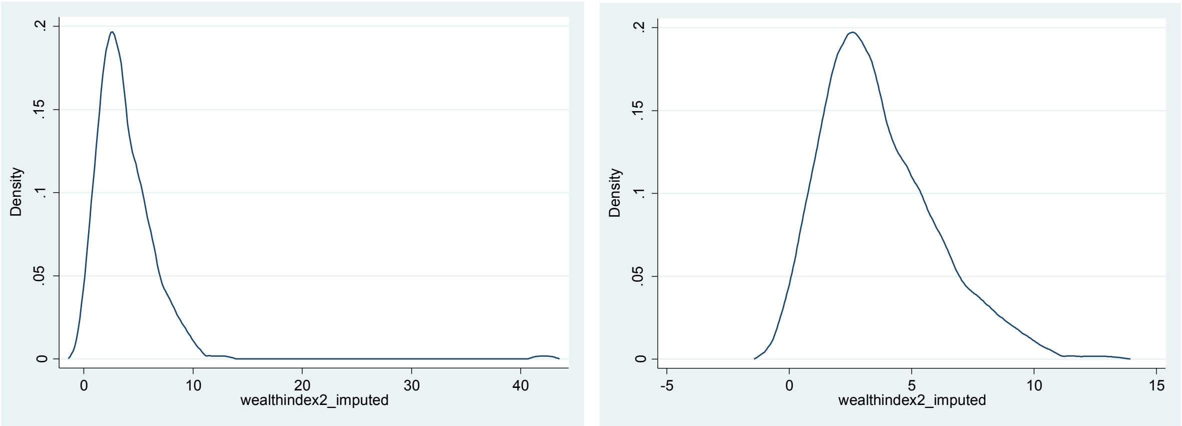


Figure 8: kernel density estimates of wealthindex2_imputed before (left) and after (right) dropping outlying observation for which *wealthindex2_imputed*=42.1.

7.3. Spillover effects

By sampling control households from within the same or neighbouring RT as the treated households, there is a concern about possible treatment externalities. We asked control households whether or not they knew a nearby residing (i.e. in the same or neighbouring RT) treated household. Out of the 125 control respondents who answered this question, 111 knew at least one nearby residing treated household. Three subsequent questions were then posed to these 111 households. For the readers convenience, these are the variable descriptions:

(1) have received cash from the treated household(s) (*spillover2*);
(2) have received business advice from the treated household(s) (*spillover3*);
(3) sold goods to the treated household(s) (*spillover4*).

Table 8: Responses of control households to questions regarding spillovers (n=116).

Variable\response	Never	seldom	sometimes	often
spillover1	101	12	3	0
spillover2	87	17	9	3
spillover3	87	13	7	9

The results in table 8 indicate that there is little concern of spillovers, as the large majority of control households do never or seldom receive cash, goods or business advice from their treated counterparts.

7.4. Attrition

Since our attrition rate is quite high (30%), it is important to test for selection bias. We assess whether differences in observable characteristics between those households that were re-interviewed in the second wave, and those that dropped out, are statistically significant. Table 9 reports the test statistics and p-values for a set of covariates deemed important and relatively accurately measured.

Table 9: Kurskal Wallis rank test results for non-random attrition

Pre-treatment Variable	$N_{\text{non-attrited}}$	N_{attrited}	Mean for non-attrited	Mean for attrited	χ^2 (1 df)	p-value
wealthindex1	206	90	3.40	2.78	4.400	0.036
spm[1]	209	90	7.61	7.23	1.084	0.298
friends	204	88	6.92	5.05	3.636	0.057
finlit	202	81	1.49	1.46	0.007	0.935
educyears	209	90	8.02	7.56	1.302	0.254

[1] spm stands for Social Progressive Matrices, our intelligence test.

The picture that emerges is that the attrited households are less wealthy pre-treatment and that their heads have lower access to valuable social networks than the households not dropping out of the sample. Hence, a complete-case analysis is likely to generate biased estimates.

7.5. Testing for selection bias under H_0: no treatment effect

When applying genetic matching with wealthindex1 as outcome on the single imputed sample, we find a SATE of 0.335 (s.e.=0.273, p-value=0.219), suggestive evidence of selection bias.

As explained in the Methods section, we start the analysis by estimating the Manski (1990) worst-case bounds while not making any structural assumptions. This delivers – (mostly) uninformative – bounds on the (in our case sample) average treatment effect.

7.6. Worst-case bounds

The method requires the outcome, wealth index of round 2, to be bounded between 0 and 1. We followed and transformed the wealth index of round 2 with 0 (the lowest value of the wealth index) as lower bound and the highest value recorded as upper bound (we dropped one outlier), and did the same thing for the round 1 wealth index. We evaluated the treatment effect at a grid of 88 covariate values, the pre-treatment wealth index ranging from 0 to 1 with jumps of 0.1 and the household size ranging from 1 member to 20 with jumps of 4 (the grid cannot contain more values than the length of the dataset in STATA). The resulting bounds are, as expected, uninformative, [-.864 ; 1]. When using then number of good friends the respondent has pre-treatment as covariate in the grid the bounds become [-1,1], the largest possible range for a treatment effect given that we have an outcome bounded between 0 and 1. Clearly, additional structural assumptions are needed to obtain more informative inferences.

7.7. Testing mean instrument independence

The statistical power of the mean exclusion restriction test put forth by (Huber & Mellace 2011) is inversely related to the complier share. In our sample, the compliance share was quite high, 0.469, which explains the fact that the p-value of the test was equal to 1. Testing power is increased when one assumes that mean potential outcomes of always takers and compliers under treatment and of never takers and compliers under non-treatment are equal. In that case the validity of the instrument can be tested by a simple equality in means test for $E(Y|D=1,Z=1)=E(Y|D=1,Z=0)$ and $E(Y|D=0,Z=0)=E(Y|D=0,Z=1)$. We ran a joint test for both hypotheses by regressing Y on a constant, D, Z and $D*Z$ (the interaction of D and Z) and used a Wald test to check whether the coefficients on Z and $Z*D$ are jointly zero. The test statistic and p-value returned, $\chi^2_{df=2}=1.0$, p-value=0.59 show that the null of instrument mean independence cannot be rejected at any conventional significance level. For *foodconsweek* as outcome, , $\chi^2_{df=2}=2.4$, p-value=0.30. To the best of our knowledge, this is the first implementation of (Huber & Mellace 2011)'s test in an applied paper. Note that the latter assumption of equality of means of potential outcomes may be violated when the credit demand of compliers depends on whether or not they know at least one person who borrows or has borrowed from TLM in the past, whereas credit demand of always takers depends more on financial criteria such as interest rates and fees.

We side with (Murray 2006) in discouraging applied researchers to test one-by-one a set of variables to select an instrument, since (a) test p-values unadjusted for multiple testing can be misleading, (b) such data mining can lead to pre-test bias if the same sample is used for estimation, (c) with small to moderate sample size, the test may have low power, especially in the presence of noise and (d) the tests discussed can only refute the null of instrument independence, not confirm it. One must have a strong prior on the validity of a certain instrument, and then let it undergo the above test to see whether the instrument withstands it.

We can also test for instrument validity the same way we tested for selection bias in section 7.5, under the null of no treatment effect. For the wealth index, we also have pre-treatment measures. When estimating the LATE of program participation on the pre-treatment wealth index, we should find an estimate of LATE that is not statistically significantly different from zero. The result from this exercise is very reassuring in terms of the credibility of the exclusion restriction when *wealthindex2_imputed* is used as outcome: LATE is exactly zero and the p-value of exactly 1 when the standard errors are analytically estimated; the bootstrap standard error estimation failed probably due to the closeness of the different LATE estimates.

7.8. Nonparametric IV estimation of SLATE

(Frölich 2007) showed that LATE is identified in a nonseparable model by among others a mean exclusion restriction. For all our estimations in this section, we use iv2 as instrument. The estimates for this estimator with the partially imputed wealthindex2 as outcome are reported in table 10 below. For comparison, the table also lists LATE estimates from 2-stage least squares (2SLS) estimation with heteroskedasticity-robust standard errors, using the same set of covariates (listed in the appendix). The results in the table are for the sample with imputed outcomes for missing outcome values; the results for the complete-case analysis are reported in the appendix. The SLATE (the sample version of LATE) is negative and statistically significantly different from zero at the 10% level in both model estimates. The p-value is an informal measure of evidence for an alternative hypothesis: it is the probability of obtaining data as extreme, or more extreme, than those observed if the null hypothesis were true.

Closeness of the estimates from the two estimators adds confidence to the findings, since they rest on different identifying assumptions. The 2SLS estimator invokes restrictive assumptions on the functional form of the structural equations and assumes exogeneity of control variables. Frölich's estimator in contrast imposes no functional form assumptions but assumes that there are no defiers and that the type is unconfounded. The 2SLS estimate is slightly biased upwards. The share of compliers (i.e. type *c*), 0.44, is high, revealing that our instrument is a strong predictor of treatments status (cf. section 5.6 for the discussion of types). The instrument is also highly significant in the first stage of 2SLS.

Table 10: Parametric and nonparametric IV estimation of SLATE on (partially imputed) wealth index (n=271). The set of covariates used is listed in APPENDIX.

	2SLS	(Frölich 2007) Bootstrapped st. errors	(Frölich 2007) Analytical st. errors
SLATE (st. error)	-0.78 (0.41)[1]	-0.893 (0.539)	-0.893 (0.440)
p-value	0.051	0.098	0.042
90% conf. interval	[-1.5 ; -0.11]	[-1.780 ; -.006]	[-1.616 ; -0.170]
Proportion of compliers			0.44
p-value instrument (*iv2*) in first stage	0.00		

[1] White heteroskedasticity-robust standard errors.

It is a well-known fact that analytical standard errors tend to overstate statistical precision in small samples. Note, that the bootstrapped p-value and confidence interval are based on the assumption of normal distribution of LATE estimates of the replications, which may not hold in a finite sample setting. For this reason, we use not less than 10,000 bootstrap replications.

To check whether the negative impact on household wealth is due to the sale of livestock, we also constructed a livestock index, as explained in the Methods section. Unlike the wealth index, this livestock index also includes the poultry count. Urban households do not typically own livestock, generating many zeros. For those households, changes in household welfare are not reflected in their livestock holdings. To limit this source of bias, we limit this estimation to the subsample of households owning an agricultural plot of land. From the experience of one of the authors, there is an almost one-to-one correspondence between the households living rurally and those owning an agricultural plot of land. Results are reported in table 11.

Table 11: Parametric and nonparametric IV estimation of SLATE on (partially imputed) livestock index *livestockindex2_imputed*. The set of covariates used is listed in the appendix.

	2SLS	(Frölich 2007) Bootstrapped st. errors	(Frölich 2007) Analytical st. errors
SLATE (st. error)	0.21 (0.25)	0.31 (0.28)	0.31 (0.23)
p-value	0.39	0.28	0.18
90% conf. interval	[-0.19 ; 0.61]	[-0.16 ; 0.77]	[-0.07 ; 0.68]
n	230	230	230
Proportion of compliers			0.46
p-value instrument (*iv2*) in first stage	0.00		

[1] White heteroskedasticity-robust standard errors.

Table 12 reports the estimates for BMI of the women as outcome; BMI of the women in round 1 is added as control variable. The estimates of LATE are negative but statistically insignificant at p=0.1. The huge bootstrapped standard errors indicate that there is a lot of noise in the data. The statistical insignificance when the standard errors are estimated analytically is to be expected given the small sample size of only 119 observations.

Table 12: Parametric and nonparametric IV estimation of LATE on BMI of women as outcome – complete-case analysis. The set of covariates includes is listed in the appendix.

	2SLS	(Frölich 2007) Bootstrapped st. errors	(Frölich 2007) Analytical st. errors
LATE (st. error)	-2.1 (2.1)[1]	0.15 (68.7)	0.15 (1.5)
p-value	0.34	1.00	0.92
90% conf. interval	[-6.2 ; 2.1]	[-112.8 ; 113.1]	[-2.2 ; 2.5]
N	115	114	114
p-value instrument (*iv2*) in first stage	0.00		

[1] White heteroskedasticity-robust standard errors.

The LATE on weekly food consumption (*foodconsweek*) as reported in table 13 are small and positive, but statistically insignificant.

Table 13: Parametric and nonparametric IV estimation of LATE on *foodconsweek* (in thousands) as outcome (n=189). The set of covariates included is listed in the appendix.

	2SLS	(Frölich 2007) Bootstrapped st. errors	(Frölich 2007) Analytical st. errors
LATE (st. error)	17.7 (21.1)	19.7 (54.0)	19.7 (21.8)
p-value	0.40	0.72	0.37
90% conf. interval	[-17.1 ; 52.6]	[-69.2 ; 108.6]	[-16.1 ; 55.5]
proportion of compliers			0.45
p-value instrument (*iv2*) in first stage	0.00		

7.9. Testing full instrument independence

The empirical p-values from our new test of instrument validity in nonseparable triangular models based on the nonparametric test of equality of conditional densities of (Li et al. 2009) with 1,000 replications are: p=0.115 for *foodconsweek2*, p=0.575 for *wealthindex2_imputed, 0.275* for *livestockindex2_imputed* and 0.275 for *bmi_woman2*. The covariates conditioned on are listed in the appendix. We conclude that the instrument iv2 is invalid for the model with *foodconsweek2* as outcome, while we did not find evidence against the null of full exclusion restriction for models with *wealthindex2_imputed* as outcome. Given the small sample size, the p-value of 0.275 should be interpreted as suggestive evidence against the validity of *iv2* when used in a model with *bmi_woman2* as outcome. We therefore only use wealthindex2_imputed as outcome for the nonparametric IV quantile estimation, since that method relies on the full exclusion restriction just tested. The results of these estimations are reported in the following section.

7.10. Nonparametric IV estimation of unconditional SQTE

Given that *iv2* is only a valid instrument when using *wealthindex2_imputed* as outcome, we report only the SQTE estimates of for that outcome. Table 14 reports the SQTE on the logged *wealthindex2_imputed*, so that the coefficients can be interpreted as the percentage change in the outcome.

Table 14 : SQTE on wealth index.

Quantile	SQTE	st. error (250 bootstr. repl.)	Empirical p-value	Normal-based 90% conf. interval
1	-0.85	0.83	0.31	[-2.2 ; 0.52]
2	-0.76	0.46	0.10	[-1.5 ; 0.01]
3	-0.65	0.37	0.09	[-1.3 ; -0.03]
4	-0.68	0.33	0.04	[-1.2 ; -0.13]
5	-0.42	0.29	0.15	[-0.90 ; 0.06]
6	-0.37	0.25	0.14	[-0.78 ; 0.04]
7	-0.26	0.22	0.23	[-0.62 ; 0.10]
8	-0.24	0.21	0.24	[-0.60 ; 0.10]
9	-0.16	0.21	0.45	[-0.51 ; 0.19]
10	-0.15	0.22	0.48	[-0.51 ; 0.20]
11	-0.19	0.22	0.38	[-0.55 ; 2.17]
12	-0.23	0.21	0.28	[-0.58 ; 0.12]
13	-0.31	0.21	0.13	[-0.65 ; 0.02]
14	-0.38	0.20	0.05	[-0.71 ; -0.06]
15	-0.29	0.19	0.12	[-0.60 ; 0.02]
16	-0.27	0.18	0.12	[-0.56 ; 0.02]
17	-0.33	0.17	0.05	[-0.61 ; -0.05]
18	-0.29	0.17	0.08	[-0.56 ; -0.01]
19	-0.36	0.16	0.03	[-0.63 ; -0.10]
20	-0.32	0.16	0.04	[-0.58 ; -0.06]

21	-0.32	0.16	0.04	[-0.58 ; -0.06]
22	-0.33	0.16	0.04	[-0.60 ; -0.07]
23	-0.26	0.18	0.14	[-0.55 ; 0.03]
24	-0.26	0.19	0.16	[-0.57 ; 0.04]

In order to appreciate the distributional effects of treatment on wealth, we plot the SQTE against the quantiles of the outcome (*wealthindex2_imputed*) in figure 9. What appears from the figure is that the negative effect of TLM's group loan on wealth is more pronounced amongst the poorer segment of the sample, although the uncertainty as expressed in the variance is also higher for the lower quantiles.

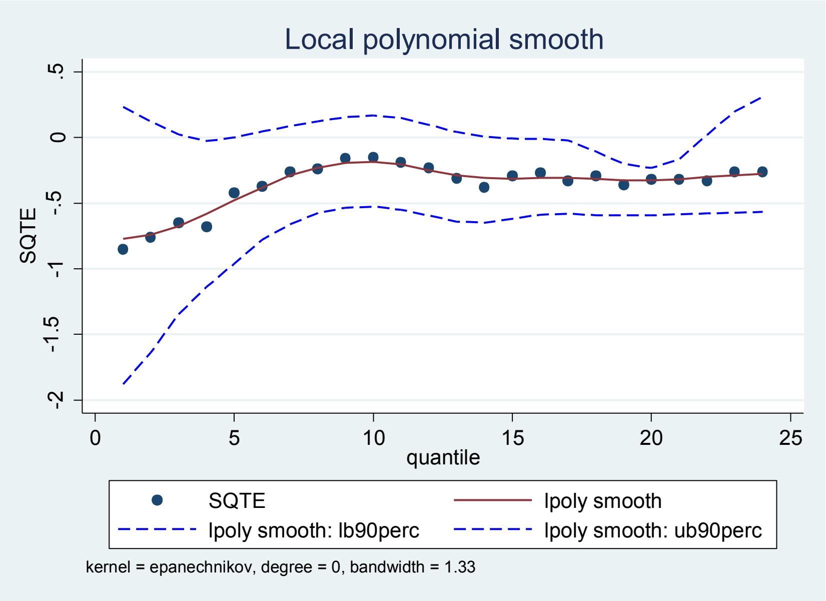


Figure 9: the sample quantile treatment effect (SQTE) as a function of the outcome, *wealthindex2_imputed*. Dashed lines are the lower and upper bound of the 90% bootstrapped confidence interval (38,600 successful bootstrap replications).

7.11. Identification without an exclusion restriction: Heckman's BVN and Millimet's MB & MB-BC estimators.

As reported in the previous sections, statistical tests gave evidence suggesting *iv2* is not a valid instrument when used in a model with *foodconsweek* as outcome. We therefore deploy alternative identifying strategies for this outcome variable that do not rely on an exclusion restriction and at the same time do not require the unconfoundedness assumption to hold. Furthermore, even for *wealthindex2_imputed*, for which we do not suspect iv2 to be an invalid instrument, estimations based on different identifying assumptions can increase the robustness of our findings and therewith their credibility. Table 15 reports the estimates and their uncertainty statistics. MB and MB-BC give estimates of the same sign as the LATE, when *wealthindex2_imputed* is the outcome. The estimates of *foodconsweek* and *bmi_woman2* are unrealistically large and unstable. We discuss possible reasons for this in the Discussion section.

Table 15: ATE estimates of BVN, MB and MB-BC on *wealthindex2_imputed, foodconsweek, bmi_woman2*.

Estimator	*wealthindex2_imputed*	*foodconsweek (in 1000's)*	*bmi_woman2*
Radius	ATE [90% conf. interval]	ATE [90% conf. interval]	ATE [90% conf. interval]
BVN	0.72 [-1.2 ; -0.09]	209 [-190 ; 350]	32.7 [-21.0 ; 38.5]
MB			
0.01	-2.5 [-2.8 ; 1.9]	-98.7 [-110 ; 130]	4.9 [-9.4 ; 8.1]
0.02	-0.94 [-2.5 ; 1.5]	-98.7 [-84 ; 110]	4.9 [-6.7 ; 5.8]
0.03	-1.1 [-2.2 ; 1.1]	-98.7 [-77 ; 87]	4.9 [-6.7 ; 5.8]
0.04	-1.2 [-2.0 ; 0.95]	-98.7 [-73 ; 85]	-0.64 [-6.2 ; 5.5]
0.05	-1.2 [-1.7 ; 0.85]	-39.1 [-66 ; 68]	-0.64 [-6.1 ; 5.1]
0.1	-1.2 [-2.0 ; 0.94]	-41.0 [-51 ; 52]	-0.8 [-4.4 ; 3.6]
0.25	-0.88 [-1.5 ; -0.08]	-12.6 [-35 ; 37]	-1.3 [-3.1 ; 2.5]
MB-BC			
0.01	-2.8 [-3.8 ; 2.5]	105 [-210 ; 390]	37.7 [-18.6 ; 41.4]
0.02	-1.2 [-3.1 ;2.2]	105 [-220 ; 360]	37.7 [-18.6 ; 40.4]
0.03	-1.4 [-2.6 ; 1.6]	105 [-210 ; 370]	37.7 [-18.6 ; 40.4]
0.04	-1.5 [-2.4 ; 1.3]	105 [-210 ; 360]	32.2 [-17.7 ; 39.8]
0.05	-1.5 [-2.5 ; 1.4]	165 [-210 ; 360]	32.2 [-17.7 ; 39.4]
0.1	-1.5 [-2.1 ; 2.7]	163 [-220 ; 350]	32.0 [-20.6 ; 38.5]
0.25	-1.1 [-1.7 ; 0.3]	191 [-210 ; 330]	31.5 [-20.5 ; 37.6]

8. Discussion

8.1. The impact of microcredit on poverty in Eastern Indonesia

The main finding that stands out is the negative short-term effect of group loan uptake on household wealth. We have high confidence in the key identifying assumption (mean instrument independence) behind this result as it withstood three different statistical tests comfortably. We conjecture that households' short-term rate of return on loan-induced investments do not exceed the loan's interest rate and fees. This coincides with the increasing criticism microfinance institutions face of making households over-indebted, a situation that deprives them and forces them to sell assets (Schicks 2010). One cautious note is that the standard errors are probably too optimistic as we single imputed rather than multiple imputed the outcomes that were missing due to attrition, as explained in the Methods section.

The negative effect on wealth is more pronounced among the poorest segments of our sample. This could be due to the fact that poorer households rely on fewer income sources, as is the case in our sample.[vi] Income diversification is an ex-ante risk coping strategy that may mitigate the need for distress sales of assets. A word of caution is needed here, as the uncertainty in terms of variance is also higher amongst the poorer quantiles. The negative effect on wealth dampens out towards the wealthier quintiles, yet our estimates become more precise among these quantiles reflected in tighter confidence intervals. Our confidence in the main finding is further strengthened by the negative effect estimates found in the MB and MB-BC estimators that rely on distinct identifying assumptions: whereas the nonparametric and parametric IV estimators critically depend on causal identifying assumptions, the source of identification of the MB-BC is restrictions on the functional form. The opposite sign found by the BVN model ATE estimate could be due to failure to meet its assumptions, in particular the absence of essential heterogeneity.

We then tested whether these distress sales consist of sale of livestock, as we believed these components of the wealth index to be most sensitive to household wealth. That belief was driven by the fact that households in poor economies often use livestock as a coping mechanism in the face of shocks to disposable income (Fischer & Buchenrieder 2010). The impact on livestock was positive and statistically insignificant however. This finding may be due to (a) the fact that for the livestock (*livestockindex2_imputed*) estimations were performed on the rural subsample, while the (*wealthindex2_imputed*) estimations were done on the whole sample, (b) the different weights on the different animal counts, (c) households sold other assets than livestock and/or (d) (nonclassical) measurement in the livestock count, more so than in the *wealthindex_imputed* which does not include the poultry count.

We found evidence against the null of validity of our instrumental variable for use in models with weekly food consumption (*foodconsweek*) and women's body mass index (*bmi_woman2*) as outcome. The estimators not relying on an exclusion restriction, BVN, MB and MB_BC, gave unrealistically large and unstable estimates of the ATE of credit uptake. This could be due to the failure of their assumptions or (nonclassical) measurement error. Measurement error in household expenditures is typically found to be mean-reverting, i.e. biases effect estimates away from zero (Millimet 2011). Poorer households who have had less education may have more difficulty in accurately recalling their expenditures or consumption from self-production. For the objectively measured weight and height that determine the body mass index, such nonclassical measurement error is less likely. We do suspect substantial (probably classical) measurement error in our length measurements. In all, we de-emphasize our estimates on food consumption and women's body mass index.

The strength of the study lies in optimizing internal validity, while the weakness lies in its external validity. The forces of client recruitment strategies by loan officers and self-selection by households into treatment status will likely be the same for new clients from the branches from which we have included treated households in our sample. Therefore, if TLM were to add staff to those branches and recruit extra clients, the findings of this thesis would apply. However, for other branches of TLM, the findings may not hold, since these areas, and the people living in them, may differ from our sample in ways that affect the effects of the TLM group loan product. The same holds for TLM expanding its operations into new territories. Islands as geographical barriers drive cultural evolution; this may include savings and investment behaviour, which can affect returns to microcredit.

Certain investments have a gestation period before they manifest in outcomes. To take the utilization of loan funds to purchase livestock for example, chicken eggs may start increasing household income shortly after loan purchase, while a goat or a pig has to be raised for often more than a year (the times span between baseline and follow-up) before having grown enough to be sold and contribute positively to household income[vii]. Investments in education have a typically long gestation period too, of possibly many years. Since 72% of the borrowers in the sample reported allocating their loans mainly to either education or other income-generating activities, our results may be a poor reflection of longer-term impacts of loan take-up. Furthermore, we only included first-time borrowers in our sample. Repeat-borrowings may have either positive or negative dynamic effects on the welfare of borrowers and their households, the more so given that these repeat-borrowings require compulsory savings of 20% of the loan size. Future studies should therefore aim to infer the longer-term impacts of credit programs in the area.

8.2. Methodological issues

On the methodological front, we focused on identifying interesting treatment parameters under a set of credible assumptions and showed that it is possible to obtain a credible impact estimate from a pre-existing, nonrandomized intervention. By opting consistently for nonparametric methods and methods that minimize and test restrictive assumptions in general, we overcome to a high degree the concern that our estimates are driven by the particular assumptions we made, rather than by the data itself. Furthermore, by relying on distinct identification strategies, we increase the robustness of our findings to these identifying restrictions.

Given our high attrition rate of 30%, we needed to impose strong assumptions about the mechanism generating the missingness of outcomes. We invoke perhaps the strongest assumption of our research that the missingness of outcomes is random (MAR). We single imputed the missing outcomes without specifying a parametric model for the prediction. Existing nonparametric imputation methods for continuous missing outcomes assume that the covariate data are continuously distributed. To the best of our knowledge, we are the first to nonparametrically single impute missing outcomes by prediction based on multivariate kernel regression with a mix of categorical and continuous predictors.

The main methodological innovation though is the development of a new test of instrument validity. It tests a full exclusion restriction in a nonseparable recursive model, but it could also be applied when the researcher uses linear IV models, since causal assumptions underlying them (depicted in figure 3 and 4) are equal to the ones in nonseparable model. Existing overidentifying restriction tests need at least one undisputedly valid instrument to work. The problem is that in most interesting settings where endogeneity is a problem, no such undisputed instrument is available. Even if it is available, adding instruments may increase statistical precision, but the resulting LATE parameter interpretation becomes less straightforward and often a less interesting policy parameter. Recall that LATE is the average effect of the treatment for those observations that are induced by the treatment to change treatment status. When using more than one instrument, LATE becomes more difficult to interpret. More useful would be a test that is robust to a lack of identification, i.e. a test that still works even if the researcher does not have access to an undisputedly valid instrument. We offer such a test. It can be applied for recursive models with mixed categorical and continuous data, as long as the instrument is multinomial. The test does not impose any parametric assumptions. Monte Carlo simulations conducted by Jeffrey Racine indicate that the test has desirable finite sample properties in terms of power and size.

The novel instrument used in this study, the number of people household members know who are being treated or were treated in the past appears to be a strong predictor of treatment status. Our new instrument validity test found some evidence against the validity of the instrument with certain outcome measures (most notably weekly food consumption). The software code we used allowed for only 2 continuous covariates and 1 binary, so the instrument may be valid for that outcome conditional on more covariates (common causes of the instrument and the outcome), something we intend to test in the future. When valid conditional on a set of common causes of the instrument and the outcome, the instrument can be seen as approximating a randomized encouragement design, with the encouragement being getting to know a treated or formerly treated person.

The rigor we apply in our methodological choices is demanding in terms of software requirements. Software add-ons for most of our methods used exist but for different software packages; we used SPSS, STATA and R. For the nonparametric single imputation method we had to program manually, as we lacked the programming skills to automatize the procedure, a time consuming task.

9. Conclusions and recommendations

The main conclusion of this thesis is that take-up of TLM's group loan has a negative short-term impact on household wealth. We conjecture that this finding is due to (a) households' short-term rate of return on loan-induced investments not exceeding the loan's interest rate and fees and (b) high prevalence of non-productive loan use in our sample. This coincides with the increasing criticism microfinance institutions face of over-indebting their clients, a situation that these destitute households to sell off assets. The negative effect on wealth is more pronounced among the poorest segments of our sample. This could be due to the fact that poorer households rely on fewer income sources or the lack of entrepreneurial skills to achieve high rates of return on investments. Since 72% of the borrowers in the sample reported allocating their loans mainly to either the education of their children or to other income-generating activities, our results may be a poor reflection of longer-term impacts of loan take-up when gestation periods of investments are long.

Given the general short follow-up of microcredit impact evaluations, future efforts should be focused on inferring longer-term impacts of microcredit. Meanwhile, the evidence of asset depletion suggests that TLM should re-evaluate their repayment schedules, at least for the poorest segments of their clientele. This could include considering relaxing the rigid requirement of a perfect repayment schedule for gaining access to second and higher loan cycles. Given the low financial literacy in our sample, loan officers evaluating loan applications should consistently have informative discussions about the intended loan use and advice against non-productive loan uses, such as wedding and funeral ceremonies. TLM could also start experimenting with commitment saving products, as the recent findings from RCT's point to their salutary welfare effects relative to microcredit.

On the methodological front, we focused on identifying interesting treatment parameters under a set of credible assumptions and showed that it is possible to obtain a credible impact estimate from a pre-existing, nonrandomized intervention. By opting consistently for nonparametric methods and methods that minimize and test restrictive assumptions in general, we overcome to a high degree the concern that our estimates are driven by the particular assumptions we made, rather than by the data itself. The main methodological innovation is the development of a new test of instrument validity. It tests a full exclusion restriction in a nonseparable recursive model. Its advantages are that it can be applied in the just identfied case, that it is identification robust, and that it does not impose any parametric assumptions.

We finally have some recommendations for researchers involved in quantitative impact evaluation of development interventions. First, given the high bias incurred from non-classical measurement error in the outcome, we recommend to focus on easily recalled measures based on asset counts, as well as objectively measurable health-related measures such as BMI, anthropometric indices of pre-school child nutritional status and serum haemoglobin concentration. Second, when a potential instrument is available, it should be subjected to different tests. Nonparametric IV estimation should be pursued alongside parametric IV estimation; the former are extremely easy to implement using STATA ado-files available on the net. Third, both when unconfoundedness holds and when it does not, estimators not imposing separability of the effects of observables and unobservables (nonparametric matching methods, nonseparable IV estimation) should be pursued more often, especially when the researcher needs to include endogenous control variables for the causal assumption to hold.

References

Anderson, T. W. & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20 (1): 46-63.

Angelsen, A., Larsen, H., Lund, J., Smith-Hall, C. & Wunder, S. (2011). *Measuring livelihoods and environmental dependence: methods for research and fieldwork*: Earthscan.

Angrist, J. D. & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*: Princeton Univ Pr.

Armendariz, A. & Morduch, J. (2010). *The economics of microfinance (2 nd eds)*: Cambridge, Massachusetts: The MIT Press.

Attanasio, O., Augsburg, B., Haas, R. D., Ftzsimons, E. & Harmgart, H. (2011). *Group lending or individual lending? Evidence from a randomised field experiment in Mongolia*: European Bank for Reconstruction and Development Unpublished manuscript.

Augsburg, B., De Haas, R., Harmgart, H. & Meghir, C. (2012). Microfinance at the Margin: Experimental Evidence from Bosnia and Herzegovina.

Banerjee, A. V. & Duflo, E. (2007). The economic lives of the poor. *The journal of economic perspectives: a journal of the American Economic Association*, 21 (1): 141.

Banerjee, A. V., Duflo, E., Glennerster, R. & Kinnan, C. (2009). *The Miracle of Microfinance?: Evidence from a Randomized Evaluation*: IFMR Research, Centre for Micro Finance.

Basu, K. & Wong, M. (2011). Evaluating Seasonal Food Security Programs in East Indonesia.

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

Bauer, M., Chytilová, J. & Morduch, J. (2012). Behavioral Foundations of Microcredit: Experimental and Survey Evidence from Rural India. *The American Economic Review*, 102 (2): 1118-1139.

Bennett, C. J. (2009). Consistent and Asymptotically Unbiased MinP Tests of Multiple Inequality Moment Restrictions. *Working Papers*.

Berhane, G. & Gardebroek, C. (2011). Does microfinance reduce rural poverty? Evidence based on household panel data from northern Ethiopia. *American Journal of Agricultural Economics*, 93 (1): 43-55.

Boente, G., González–Manteiga, W. & Pérez–González, A. (2009). Robust nonparametric estimation with missing data. *Journal of Statistical Planning and Inference*, 139 (2): 571-592.

BPS. (2010a). Gross Regional Domestic Product at Current Market Prices by Provinces, 2004 - 2010 (Million Rupiahs) Available at: http://dds.bps.go.id/eng/tab_sub/view.php?tabel=1&daftar=1&id_subyek=52&notab=1 (accessed: 21-01-2012).

BPS. (2010b). Population of Indonesia by Province 1971, 1980, 1990, 1995 , 2000 and 2010. Available at: http://dds.bps.go.id/eng/tab_sub/view.php?tabel=1&daftar=1&id_subyek=12&notab=1 (accessed: 21-03-2012).

BPS. (2010c). Number and Percentage of Poor People, Poverty Line, Poverty Gap Index, Poverty Severity Index by Province, 2010. Available at: http://dds.bps.go.id/eng/tab_sub/view.php?tabel=1&daftar=1&id_subyek=23&notab=1 (accessed: 21-03-2012).

BPS. (2010d). Nett Enrollment Ratio (N E R) By Province , 2003-2010. Available at: http://dds.bps.go.id/eng/tab_sub/view.php?tabel=1&daftar=1&id_subyek=28&notab=4 (accessed: 21-03-2012).

Bushway, S., Johnson, B. D. & Slocum, L. A. (2007). Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. *Journal of Quantitative Criminology*, 23 (2): 151-178.

Chalak, K. & White, H. (2012). causality, conditional independence, and graphical separation in settable systems. *Neural Computation*, 24 (7): 1611-1668.

Chalak, K. W., H. (2012). causality, conditional independence, and graphical separation in settable systems. *Neural Computation*, 24 (7): 1611-1668.

Crepon, B., Devoto, F., Duflo, E. & Pariente, W. (2011). *Impact of Microcredit in Rural Areas of Morocco: Evidence from a Randomized Evaluation*. Cambridge, Mass: MIT, March. Unpublished manuscript.

Daley-Harris, S. (2009). State of the microcredit summit campaign report 2009. *Microcredit Summit Campaign: Washington, DC*: 4-76.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*: 1-31.

de Aghion, B. A., Armendariz, B. & Morduch, J. (2007). *The economics of microfinance*: The MIT Press.

De Mel, S., McKenzie, D. & Woodruff, C. (2008). Returns to capital in microenterprises: evidence from a field experiment. *The Quarterly Journal of Economics*, 123 (4): 1329-1372.

Desai, J., Johnson, K. & Tarozzi, A. (2011). On the Impact of Microcredit: Evidence from a Randomized Intervention Rural Ethiopia.

Diamond, A. & Sekhon, J. S. (2006). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies.

Doko, F. & Dufour, J. M. (2008). Instrument endogeneity and identification-robust tests: some analytical results. *Journal of Statistical Planning and Inference*, 138 (9): 2649-2661.

Dupas, P. & Robinson, J. (2009). Savings constraints and microenterprise development: Evidence from a field experiment in Kenya: National Bureau of Economic Research.

FAO. (2005). Livestock sector brief: Indonesia. Livestock Information, Sector Analysis and Policy Branch. *Livestock sector brief*. Rome: FAO.

Ferguson, B., Tandon, A., Gakidou, E. & Murray, C. (2003). Estimating permanent income using indicator variables. *Health systems performance assessment: debates, methods and empiricism. Geneva: World Health Organization*: 747-760.

Fischer, I. & Buchenrieder, G. (2010). *Risk management of vulnerable rural households in southeast Asia*. 4-7 pp.

Frölich, M. (2006). Non-parametric regression for binary dependent variables. *The Econometrics Journal*, 9 (3): 511-540.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139 (1): 35-75.

Frölich, M. (2008). Parametric and nonparametric regression in the presence of endogenous control variables. *International Statistical Review*, 76 (2): 214-227.

Frölich, M. & Melly, B. (2008). Unconditional quantile treatment effects under endogeneity. *IZA DP* (3288).

GB, O. (2009). *A Brief Review on The Persistent of Food Insecurity and Malnutrition Problems in East Nusa Tenggara Province, Indonesia*: Institute of Indonesia Tenggara Timur Studies. Unpublished manuscript.

Grootaert, C. (2004). *Measuring social capital: an integrated questionnaire*: World Bank Publications.

Gujarati, D. N. (2003). *Basic Econometrics. 4th*: New York: McGraw-Hill.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*: 1029-1054.

Hare, D. (1999). 'Push'versus 'pull'factors in migration outflows and returns: Determinants of migration status and spell duration among China's rural population. *The Journal of Development Studies*, 35 (3): 45-72.

Hayfield, T. & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27 (5): 1-32.

Heckman, J. J., LaLonde, R. J. & Smith, J. A. (1999). The economics and econometrics of active labor market programs. *Handbook of labor economics*, 3: 1865-2097.

Heckman, J. J., Urzua, S. & Vytlacil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity: National Bureau of Economic Research.

Hens, N. (2005). *Non-and semi-parametric techniques for handling missing data*: University of Hasselt, Center for Statistics.

Hoderlein, S. & Winter, J. (2010). Structural measurement errors in nonseparable models. *Journal of Econometrics*, 157 (2): 432-440.

Huber, M. & Mellace, G. (2011). Testing instrument validity for LATE identification based on inequality moment constraints. *Economics Working Paper Series*.

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. & Zinman, J. (2011). Microcredit in theory and practice: Using randomized credit scoring for impact evaluation. *Science*, 332 (6035): 1278.

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

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70 (5): 1781-1803.

Kolenikov, S. & Angeles, G. (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth*, 55 (1): 128-165.

Larrea, C. (2002). Desigualdad social, salud materno–infantil y nutrición en ocho países de América Latina: Análisis comparativo de las encuestas DHS III. *línea]< http://www. paho. org/Spanish/HPP/HPN/larrea–encuesta DHS. htm*.

Lawrance, E. C. (1991). Poverty and the Rate of Time Preference: Evidence from Panel Data. *Journal of Political Economy*, 99 (1): 54-77.

Lee, M. J. (2005). *Micro-econometrics for policy, program, and treatment effects*: Oxford University Press, USA.

Li, Q., Maasoumi, E. & Racine, J. S. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics*, 148 (2): 186-200.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80 (2): 319-323.

McKenzie, D. & Woodruff, C. (2008). Experimental evidence on returns to capital and access to finance in Mexico. *The World Bank Economic Review*, 22 (3): 457-482.

Meulman, J. (1998). Optimal scaling methods for multivariate categorical data analysis. *SPSS white paper: http://www. spss. com*.

Millimet, D. L. (2011). The elephant in the corner: a cautionary tale about measurement error in treatment effects models.

Millimet, D. L. & Tchernis, R. (2012). Estimation of Treatment Effects without an Exclusion Restriction: with an Application to the Analysis of the School Breakfast Program. *Journal of Applied Econometrics*: n/a-n/a.

Morduch, J. (1999). The Microfinance Promise. *Journal of Economic Literature*, 37 (4): 1569-1614.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *The journal of economic perspectives*, 20 (4): 111-132.

Nobelprize.org. (2012). *The Noble Peace Prize 2006*. Available at: http://www.nobelprize.org/nobel_prizes/peace/laureates/2006/ (accessed: 7 July).

Pearlman, S. (2012). *Can Low Returns to Capital Explain Low Formal Credit Use? Evidence from Ecuador.* Unpublished manuscript.

Pine, A., Seymour, B., Roiser, J. P., Bossaerts, P., Friston, K. J., Curran, H. V. & Dolan, R. J. (2009). Encoding of marginal utility across time in the human brain. *The Journal of Neuroscience*, 29 (30): 9575-9581.

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

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of economic surveys*, 14 (1): 53-68.

Reiss, J. (2005). Causal instrumental variables and interventions. *Philosophy of science*, 72 (5): 964-976.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*: 41-48.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, vol. 519: Wiley Online Library.

Rudd, J. (2000). Empirical evidence on human capital spillovers.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*: 393-415.

Schicks, J. (2010). Microfinance Over-Indebtedness: Understanding its drivers and challenging the common myths. *Bruxelles: Centre Emilee Bergheim, Solvay School of Business, CEB Working Paper* (10/048).

Schroeder, E. (2010). The Impact of Microcredit Borrowing on Household Consumption in Bangladesh: Mimeo, Georgetown University Economics Department.

Sekhon, J. S. & Mebane Jr, W. R. (1998). Genetic optimization using derivatives. *Political Analysis*, 7 (1): 187-210.

Stiglitz, J. E. & Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. *The American Economic Review*, 71 (3): 393-410.

TLM. (2010). *Who are our clients?* Available at: http://www.ytlm.org/main/who%20are%20our%20clients.html (accessed: 05-09).

UNICEF, O., HKI. (2008). Nutrition survey in East Nusa Tenggara (NTT) 1 October 2007 - 31 March 2008.

Verbeek, M. (2008). *A guide to modern econometrics*: Wiley.

Vyas, S. & Kumaranayake, L. (2006). Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning*, 21 (6): 459-468.

Wooldridge, J. (2010). Econometric Analysis of Cross Section and Panel Data. *MIT Press Books*, 1.

Wooldridge, J. M. (2005). Violating ignorability of treatment by controlling for too many factors. *Econometric Theory*, 21 (05): 1026-1028.

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach*: South-Western Pub.

World Bank. (2010a). PPP conversion factor, GDP (LCU per international $) Available at: http://data.worldbank.org/indicator/PA.NUS.PPP (accessed: 21-03-2012).

World Bank. (2010b). School enrollment, primary (% net). Available at: http://data.worldbank.org/indicator/SE.PRM.NENR (accessed: 21-03-2012).

World Bank. (2010c). School enrollment, secondary (% net) Available at: http://data.worldbank.org/indicator/SE.SEC.NENR (accessed: 21-03-2012).

Appendix

Dropping outliers: exclude observations with *foodconsweek* > 500,000; exclude observations with *bmi_woman2* > 45.

Covariates in the predictive model for single imputation of *wealthindex2* and *livestockindex2*: *wealthindex1, friends, hhsize, age, gender, primaryschool, smp, sma, smk, dip1orhigher, borrowhistory, catholic, muslim, organiz2, spm, friendsinfo, friendsgoodinfo, trustgeneral, socialcapital1, socialcapital2, iv2, healthstatus, treatment.*

Covariates used in the LATE models with *wealthindex2_imputed, livestock2_imputed* and *foodconsweek* as outcome: *wealthindex1, finlit, spm, hhsize, age, landarea, primaryschool, smp, sma, smk, dip1orhigher, gender, savings, muslim, hasbusiness, incomediv.*

Covariates used in the LATE model with *bmi_woman2* as outcome: *finlit, friends, hhsize, landarea, primaryschool, smp, sma, smk, dip1orhigher, gender, savings, muslim, hasbusiness, incomediv.*

Covariates used in the QTE model with *wealthindex2_imputed* as outcome: *wealthindex1, hhsize, age, ownland, primaryschool, smp, sma, smk, dip1orhigher, savings, muslim.*

Covariates used for bmte (BVN, MB and MB-BC estimators): *wealthindex1, finlit, spm, hhsize, age, landarea, primaryschool, smp, sma, smk, dip1orhigher, gender, savings, muslim, hasbusiness, incomediv.*

Covariates conditioned on in the full exclusion restriction test with *wealthindex2_imputed* and *foodconsweek* as outcome: *wealthindex1, finlit, primaryschoolorhigher*.

Covariates conditioned on in the full exclusion restriction test with *livestockindex2_imputed* as outcome: *wealthindex1, finlit, gender*.

Covariates conditioned on in the full exclusion restriction test with *bmi_woman2* as outcome: *bmi_woman1, finlit, primaryschoolorhigher*.

Table 16: Variable names and descriptions

Variable name	**Description**
(Outcomes)	
wealthindex2	Index constructed from asset and livestock holding using CATPCA (round 2)
SES_index2	Same as wealth_index2 but including dummy childeduc (round 2)
BMI_woman2	Body Mass Index of woman in hh as described under data (round 2)
consmonth	Monetary value of monthly household consumption
(Treatment)	

treatment	1=household borrows from TLM, 0=otherwise
(Instrument)	
iv1	The number of people the hh head or other adult respondents of the hh know who borrow or ever borrowed from TLM, not including the members of the current borrowing group for treated hhs
iv2	1=iv1 bigger than 0, 0=otherwise
(control vars.)	All measured in round 1
hhsize	Number of hh members
age	Age of the hh head
gender	Gender of the hh head
primaryschool	1=hh head completed primaryschool, 0=otherwise
smp	1=hh head completed smp (junior high school), 0=otherwise
smasmk	1=hh head completed sma or smk (senior high school), 0=otherwise
dip1orhigher	1=hh head completed a post-secondary school degree, 0=otherwise
religion	1=protestant, 2=catholic, 3=muslim
rspm	Raven's Standard Progressive Matrices, a proxy for reasoning ability
finlit	The number of correct answers to a test consisting of 5 questions, proxying financial literacy
friends	Total number of friends, as answer to the question "About how many *close friends* do you have these days? These are people you feel at ease with, can talk to about private matters, or call on for help."
friendsinfo	How many people give you advice on how to run a business or farm?
friendsgoodinfo	How many of them give you information that is useful?
wealthindex1	The wealth index excluding livestock in round 1.
landarea	landarea in hectare
ownland	1=landarea>0, 0=otherwise
businessplans1	1=household answers "yes" to the question "Do you have business plans?", 0=otherwise
organiz2	The respondent is asked for list of different organization types (religious, farmer association, etc.) whether he/she is member. 1=member of at least one of them; 0=otherwise.
savings	1=has savings account, 0=otherwise
muslim	1=muslim, 0=otherwise
catholic	1=catholic, 0=otherwise
hasbusiness	1=hh runs a business, 0=otherwise
incomediv	1=more than 1 income generating activity, 0=otherwise
educyears	Years of education of the households head
sma	1=hh head completed sma (general senior high school), 0=otherwise
smk	1=hh head completed smk (technical senior high school), 0=otherwise
floor(2)	0=earth floor, 1=otherwise
roof(2)	1=zinc, 0=otherwise
trustgeneral	In general, can people be trusted?" 1=yes, 0=no
socialcapital1	Other people, will they 1=try to take advantage of you if they get the chance to or 0=try to help
socialcapital2	How about other people. Are they 1=likely to help or 0=they only care about

	themselves
borrowhistory	1=ever borrowed in the past, 0=otherwise
healthstatus	In general, how would you describe your health in the past? 1=very good, 2=good, 3=in between good and bad, 4=bad, 5=very bad
bmi_woman1(2)	body mass index of woman in 1st (2nd) survey round
consmonth	weekly food consumption in round 2
mainloanuse2	Main (according to share of the loansize) loan use as reported in round 2. 1=production; 2=consumption, 3=education.
poultry	number of chicken/ducks owned by the household

Table 17: *finlit* is the number of correct answers to the following 5 questions, based on the Baseline Survey of the Adolescent Development Programme – Adolescent Module, used with permission from Selim Gulesci.

	Question	**Codes**

1	What are the things you need to know in order to make a budget for the HH?	Income [1];
		Expenditure [2];
		Savings [3];
		Don't know [4];
		others (specify)

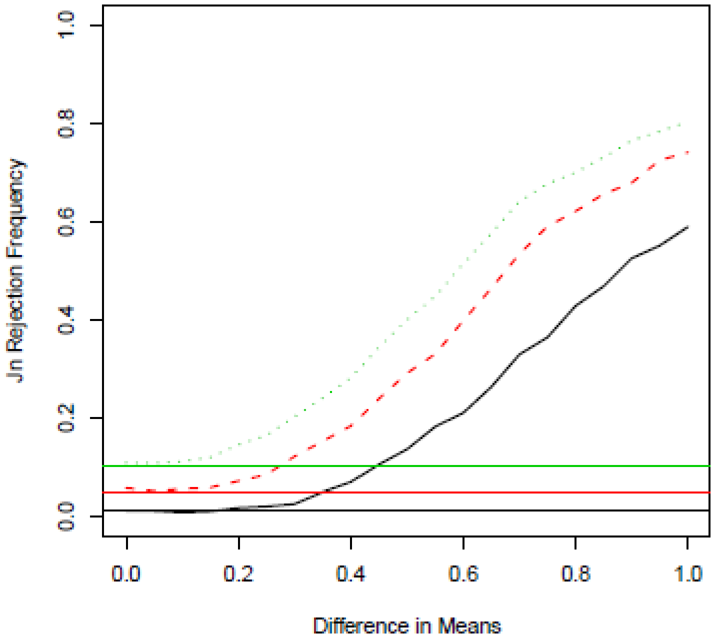
2	Is there any difference in the interest rate of a current account and savings account in a bank? If so, which one gives a higher interest rate?	savings account [1];
		current account [2];
		same interest rate [3];
		don't know [4]

3	Suppose you have deposited Rp. 100,000 in the bank for an interest of 10,000 per year. If you withdraw all the money after 2 years, how much will you get?	write down the amount of money in

4	Suppose you need to take a loan of Rp. 1,000,000 and you have two choices. In one is you pay an interest of Rp. 10,000 every month and in the other you pay an interest of Rp. 120,000 at the end of the year. Which one has a higher interest rate?	1st option (monthly) [1];
		2nd option (yearly) [2];
		Same interest rate for both [3]

5	What will happen to the price of charcoal if the price of kerosene increases?	Increase [1];
		Decrease [2];
		Unchanged [3];
		Don't know [4]

Results of the Monte Carlo simulations performed by Jeffrey Racine. The simulations are of continuos, normally distributed outcomes, two continuous normally distributed covariates, a binary treatment, a binary covariate and a binary instrument. The two samples for which the instrument takes on the value of zero and one differ in a mean shift of the conditional outcome.



The empirical rejection frequencies for different values of the mean shift $\Delta_{\bar{a}}$of the conditional densities.

$\Delta_{\bar{a}}$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
0.00	0.011	0.057	0.108
0.05	0.011	0.051	0.108
0.10	0.008	0.055	0.111
0.15	0.010	0.058	0.119
0.20	0.017	0.071	0.145
0.25	0.020	0.084	0.164
0.30	0.025	0.121	0.202
0.35	0.049	0.152	0.240
0.40	0.069	0.183	0.281
0.45	0.105	0.239	0.344
0.50	0.136	0.291	0.400
0.55	0.182	0.330	0.448
0.60	0.210	0.398	0.514
0.65	0.263	0.464	0.576
0.70	0.329	0.534	0.641
0.75	0.364	0.590	0.677
0.80	0.428	0.621	0.700
0.85	0.468	0.656	0.731
0.90	0.525	0.679	0.764
0.95	0.551	0.725	0.783
1.00	0.589	0.741	0.802

Endnotes

[i] See www.bri.co.id/about_visi

[ii] Available at http://www.rand.org/content/dam/rand/www/external/labor/bps/datadocpdf/susenas/susenas99m.pdf

[iii] Taken from the World Bank Social Capital Questionnaire (Grootaert 2004)

[iv] For reasons unknown to us, the command failed to predict outcome values for 9 observations (hhid 41, 80-84, 86, 87, 89).

[v] Available at Millimet's website: http://faculty.smu.edu/millimet/code.html.

[vi] Indeed, the pre-treatment wealth index (*wealthindex1*) was found to be significantly higher for households in our sample that have more than one income source (*incomediv*) than for those that do not. A Wilcoxon rank sum (Mann-Whitney) test, a nonparametric test of the difference of group means (defined by the value of the binary variable *incomediv*) returned $Z = -3.3$, $p = 0.001$, rejecting the null of no equal group means of *wealthindex1* at 5% level of significance.

[vii] Where we reasonably assume that changes in (permanent) household income is the primary determinant of changes in asset holdings and an important determinant of nutritional status.