Working Paper Number 174
June 2009

The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence

David Roodman and Jonathan Morduch

Abstract

The most-noted studies on the impact of microcredit on households are based on a survey fielded in Bangladesh in the 1990s. Contradictions among them have produced lasting controversy and confusion. Pitt and Khandker (PK, 1998) apply a quasi-experimental design to 1991–92 data; they conclude that microcredit raises household consumption, especially when lent to women. Khandker (2005) applies panel methods using a 1999 resurvey; he concurs and extrapolates to conclude that microcredit helps the extremely poor even more than the moderately poor. But using simpler estimators than PK, Morduch (1999) finds no impact on the level of consumption in the 1991–92 data, even as he questions PK's identifying assumptions. He does find evidence that microcredit reduces consumption volatility. Partly because of the sophistication of PK's Maximum Likelihood estimator, the conflicting results were never directly confronted and reconciled. We end the impasse. A replication exercise shows that all these studies' evidence for impact is weak. As for PK's headline results, we obtain opposite signs. But we do not conclude that lending to women does harm. Rather, all three studies appear to fail in expunging endogeneity. We conclude that for non-experimental methods to retain a place in the program evalu-ator's portfolio, the quality of the claimed natural experiments must be high and demonstrated.

The Center for Global Development is an independent, nonprofit policy research organization that is dedicated to reducing global poverty and inequality and to making globalization work for the poor.

Use and dissemination of this Working Paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License. The views expressed in this paper are those of the author and should not be attributed to the board of directors or funders of the Center for Global Development.

www.cgdev.org

David Roodman and Jonathan Morduch. 2009. "The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence."
CGD Working Paper 174. Washington, D.C.: Center for Global Development.
http://www.cgdev.org/content/publications/detail/1422302

The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence[1]

David Roodman
Center for Global Development

Jonathan Morduch
New York University
Financial Access Initiative

June 2009

Abstract: The most-noted studies on the impact of microcredit on households are based on a survey fielded in Bangladesh in the 1990s. Contradictions among them have produced lasting controversy and confusion. Pitt and Khandker (PK, 1998) apply a quasi-experimental design to 1991–92 data; they conclude that microcredit raises household consumption, especially when lent to women. Khandker (2005) applies panel methods using a 1999 resurvey; he concurs and extrapolates to conclude that microcredit helps the extremely poor even more than the moderately poor. But using simpler estimators than PK, Morduch (1999) finds no impact on the level of consumption in the 1991–92 data, even as he questions PK's identifying assumptions. He does find evidence that microcredit reduces consumption *volatility*. Partly because of the sophistication of PK's Maximum Likelihood estimator, the conflicting results were never directly confronted and reconciled. We end the impasse. A replication exercise shows that all these studies' evidence for impact is weak. As for PK's headline results, we obtain opposite signs. But we do not conclude that lending to women does harm. Rather, all three studies appear to fail in expunging endogeneity. We conclude that for non-experimental methods to retain a place in the program evaluator's portfolio, the quality of the claimed natural experiments must be high and demonstrated.

[1] We thank Mark Pitt and the Research Committee of the World Bank for assistance with data, Maren Duvendack and Richard Palmer Jones for scrutiny of our data set construction, and Xavier Giné and Dean Karlan for reviews. Correspondence: David Roodman, droodman@cgdev.org.

Microcredit is a phenomenon that needs little introduction. From its beginnings in the late 1970s, the idea that access to small loans can help poor families build businesses, increase incomes, and exit poverty has blossomed into a global movement. The movement has captured the public imagination, drawn billions of dollars in financing, reached millions of customers, and garnered a Nobel Peace Prize. Microfinance is manifold in its appeal. It is radical in its suggestion that the poor are creditworthy and conservative in its insistence on individual responsibility. It offers, as the cliché goes, a hand up, not a handout. Because its currency is currency itself, microcredit makes supporters feel that *their* hands are reaching out directly to the poor. And microfinance, especially when channeled to women, is seen as demonstrably lifting people out of poverty. Mohammad Yunus, the visionary founder of the Grameen Bank, often cites the figure that "5 percent of the Grameen borrowers get out of poverty every year."[1]

Yet against this strong appeal, a natural question has long been asked: how robust is the evidence that microcredit works? The question only gains in importance as microcredit touches more lives and attracts more (but scarce) government and private funding. Of course, "works" can mean many things. By one definition, the existence of thriving, competing microfinance organizations and the voluntary patronage of millions of poor people is success in itself. After all, no one asks whether the thriving mobile phone business in the Congo is "working." But by a definition often used by program evaluators and academic researchers, the test is whether the interventions measurably improved the lives of the poor, such as through higher or more stable household consumption. Many studies have attempted to put microfinance to that test, and a few have merited publications in economics journals. In this paper, we revisit the most influential among those studies, including the source of the figure that Yunus cites.

During its first 20 years, the literature on the impact of microcredit relied almost exclusively on non-experimental methods (Armendáriz de Aghion and Morduch 2005, ch. 8). The challenges of establishing causality in such studies are well-known. They include potential biases from omitted variables as

[1] Interview in 2007 on the PBS show "NOW," at pbs.org/now/enterprisingideas/Muhammad-Yunus.html.

well as non-random program placement, client selection and self-selection, and attrition.[2]

A few studies, however, have made strong claims to causal identification. Most of these are based on household surveys funded by the World Bank and carried out with the Bangladesh Institute of Development Studies in Bangladesh—three rounds in 1991–92 and a fourth in 1999. In particular, Pitt and Khandker (1998, henceforth PK) and Khandker (2005, henceforth simply Khandker) have exercised the most influence within and beyond academia.[3] PK uses the data from the first three seasonal rounds and claims quasi-experimental identification; Khandker does not assert a quasi-experiment but takes advantage of the panel dimension introduced by the 1999 follow-up round to strengthen identification. These studies have gained credence and interest from their focus on Bangladesh, a hotbed of microfinance; from the dimensions of the data set (some 1800 households with 7–8-year follow-up); and from understandings, demonstrated in sophisticated economic and econometric analysis, of the challenges to identification.

The studies also exercise great influence beyond the research community. PK's headline result is that "annual household consumption expenditure increases 18 taka for every 100 additional taka borrowed by women…compared with 11 taka for men." In a book, Khandker (1998, p. 56) extrapolates from this finding, derived from the early 1990s data, to the poverty impact Yunus has cited. Meanwhile, a literature survey commissioned by the U.S.-based Grameen Foundation judges that Khandker's 2005 paper, the one incorporating all the 1990s data, "may…be the most reliable impact evaluation of a microfinance program to date" (Goldberg 2005). The president of Freedom from Hunger, a global microfinance group, follows suit, describing Khandker as the "one major study of microfinance impact on po-

[2] One prominent encounter with these difficulties: in the late 1990s, the U.S. Agency for International Development commissioned studies using new members as controls for old ones in evaluation. But that method can bias results to the extent that cohorts differ systematically, e.g., because of attrition (Karlan 2001).
[3] Also based on this data set are Khandker (1996, 2000); Pitt et al. (1999); Pitt (2000); McKernan (2002); Pitt and Khandker (2002); Pitt et al. (2003); Menon (2005); Pitt, Khandker, and Cartwright (2006); and Chemin (2008). Kaboski and Townsend (2005) use similar econometrics but different instruments to study the impacts of microfinance in Thailand.

verty that stands out" (Dunford 2006).

We think these Bangladesh-based papers are worth revisiting for two reasons. First, they have not gone without criticism. The most prominent are in Morduch (1998, henceforth Morduch), which questions assumptions at the heart of PK's asserted quasi-experiment and fails to match their main results with a simpler estimator. (Morduch does find evidence that microcredit reduces consumption *volatility*.) Neither Morduch nor Pitt's (1999) response were published, and their separate estimates were never reconciled, so the debate over this research effort remains unresolved.[4] Second, as the economics profession and major donors shift toward randomized studies, the value of non-randomized approaches is a live question.[5] Our intuition is that randomized and non-randomized approaches have different strengths and weaknesses—non-randomized ones, for example, can opportunistically exploit natural experiments—and that the optimal research portfolio from the point of view of policy should blend the two. Less clear is exactly when non-experimental studies are worth performing.

After going through a replication exercise—applying the same methods to the same data as in PK, Morduch, and Khandker—and performing related Two-Stage Least-Squares (2SLS) regressions, we come to doubt the positive results in all three. With regard to the headline PK finding, our replication generates results opposite in sign. But we do not conclude that microcredit harms; rather, specification tests suggest that the instrumentation strategy is failing, that reverse or omitted-variable causation is driving the results, and that the endogenous credit-consumption relationship varies substantially by sub-sample, as well as borrower sex, which can explain the seeming gender differential in impact. We offer data that question the basis for the quasi-experimental identification in PK (and by extension in Morduch) and show how, in Khandker, exploiting the panel dimension does not compensate for the lack of clearly exogenous variation in the treatment variable. As a result, strikingly, 30 years into the microfin-

[4] Morduch discusses PK in Morduch (1999) and discusses Khandker in Armendáriz de Aghion and Morduch (2005), neither of which were refereed nor provide alternative estimates.
[5] See, for example, the back and forth between Banerjee and Duflo (2008) and Deaton (2009).

ance movement we have little solid evidence that it improves the lives of clients in measurable ways.

At the risk of over-generalizing from one data point, this experience leads us to conclude that when studying causality in social systems with strong endogeneity, claims of non-experimental identification need to be held to demanding standards. The experience also casts doubt on the power of sophisticated parametric techniques to compensate for the lack of such.

The next three sections of this paper describe the identification strategies and results of the three papers of interest and the findings from our replications. The conclusion summarizes and explores the broader lessons.

Pitt and Khandker (1998)

The study

PK analyze surveys of 1,798 households in 87 villages within 29 randomly selected *upazillas* of Bangladesh in 1991–92. (At the time, the country was divided into 391 *upazillas*.) The surveyors visited the households after each of the three main rice seasons—*Aman* (December–January), *Boro* (April–May), and *Aus* (July–August)—losing only 29 households from the sample over the period. The surveyors oversampled households participating in one of the three credit programs evaluated—those of the Grameen Bank, a large NGO called BRAC, and the official Bangladesh Rural Development Board (BRDB)—and oversampled eligible nonparticipants. Since sampling on the basis of eligibility can bias results, PK incorporate sampling weights that are constructed from censuses taken in each study village. All three credit programs formally defined eligibility in terms of land ownership: only functionally landless households, defined as those owning half an acre or less, could borrow.[6] Although most group-based microcredit in Bangladesh now goes to women, the earliest experiments carried out by Yunus and his students in the 1970s targeted men. The shift toward women occurred during the 1980s. Thus in the

[6] Among the three creditors, Grameen at least also applied an alternative eligibility criterion: ownership of assets worth less than one acre of medium-quality land (Hossain 1988, p. 25). However, PK emphasize the half-acre rule in their analysis by, for example, using it to code the "target" status of control village households.

1991–92 surveys, 10 villages had only male borrowing groups, 22 had only female groups, and 40 had both. All groups were single-sex.

In the PK estimation set-up, the three-way split by credit supplier and the two-way split by gender lead to six parameters of interest for a given outcome. A central feature of the estimation problem is that credit variables, by supplier and gender, are at once potentially endogenous and censored (Tobit). Meanwhile, some of the outcomes, such as labor supply and girl's school enrollment, are themselves censored or binary. PK therefore estimate the key impact parameters using a limited-information maximum likelihood (LIML) framework that effectively allows for instrumental variables and appropriately handles censoring. The model contains equations for the outcome variable of interest, for female borrowing, and for male borrowing. The outcome is variously modeled as continuous and unbounded (for log weekly household consumption), Tobit (female non-land assets, female and male labor supply per month), or probit (school enrollment of school-age boys or girls). To state the model precisely, let p_f and p_m be dummies indicating whether credit groups composed of females or males are operating in a given village; and let e be a dummy for whether a household meets the eligibility criteria of such programs, regardless of whether any operate in the village. Then the *credit choice* variables indicating whether women and men in a household can borrow are

$$c_f = p_f e$$

$$c_m = p_m e.$$

Let y_o be the outcome. For some outcomes y_o is modeled as Tobit or probit. But since we focus on household consumption, we will assume y_o is continuous and unbounded. Let y_f and y_m be total borrowings of all female and all male household members, let $\mathbf{y}_{fm} = (y_{f1}, y_{f2}, y_{f3}, y_{m1}, y_{m2}, y_{m3})'$ be the six credit variables disaggregated by program as well as gender, and **x** be a vector of exogenous controls. Then the PK model is

$$\begin{aligned}
y_o &= \mathbf{y}_{fm}{}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\beta}_o + \epsilon_o \\
y_f^* &= \mathbf{x}'\boldsymbol{\beta}_f + \epsilon_f \text{ if } c_f = 1 \\
y_m^* &= \mathbf{x}'\boldsymbol{\beta}_m + \epsilon_m \text{ if } c_m = 1 \\
y_f &= 1\{y_f^* \geq C\} \cdot y_f^* \\
y_m &= 1\{y_m^* \geq C\} \cdot y_m^* \\
(\epsilon_o, \epsilon_f, \epsilon_m)' &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).
\end{aligned} \tag{1}$$

where C is the credit censoring level, $\boldsymbol{\Sigma}$ is a 3×3 positive-definite symmetric matrix, and $1\{\}$ indicates a dummy.

The PK econometric model is innovative and can be counterintuitive for those unfamiliar with the methods. All three equations include exactly the same set of regressors on the right-hand-side, except of course that the outcome equation also includes credit variables. Superficially, there appear to be no excluded instruments.[7] Meanwhile, the credit equations' samples are restricted, which means that the number of equations in the model varies by observation. A final counterintuitive feature is that the outcome equation contains six endogenous credit variables—one for each gender and program—but the

[7] In fact, exclusion restrictions become less necessary for identification in the presence of censoring. Wilde (2000) shows that none is generally needed in multi-equation probit systems.

model includes just two instrumenting equations (those for y_f^* and y_m^*).

Despite this combination of unusual features, the intuition behind the model is analogous to a conventional two-stage instrumental variables set-up in which all equations apply to all observations but all right-hand side variables in the instrumenting equations are entered after being interacted with dummies for those equations' samples in the LIML set-up:

$$\begin{aligned} y_f^* &= c_f \mathbf{x}' \boldsymbol{\beta}_f + C + \epsilon_f \\ y_m^* &= c_m \mathbf{x}' \boldsymbol{\beta}_m + C + \epsilon_m \end{aligned} \tag{2}$$

(The inclusion of *C* sets y_f^* and y_m^* to the censoring level when credit is not available.) Thus PK effectively instrument for the borrowing variable with interactions between the credit choice dummies and all the included exogenous variables. In PK, these exogenous variables include age, sex, and education of the household head; other household characteristics; a set of village characteristics or dummies; and, in the case of regressions on individual-level data, individual characteristics. They also include the constant term, so that c_f and c_m are themselves instruments. To understand how it is possible to have six credit variables in the final stage while instrumenting two more aggregated ones in the first stage, we can imagine instrumenting all six distinctly and imposing constraints that equate first-stage coefficients across the three lending programs.

As multi-equation systems that mix Tobit, probit, and classical continuous and unbounded variables, the PK models for various outcomes are *conditional, recursive, fully observed*, *mixed-process systems*. They are *recursive* in that they contain clear stages, in this case two, and do not model simultaneous causation.[8] They are *fully observed* (Roodman 2009b) in that the observed y_f and y_m, not the latent y_f^* and y_m^*, appear in the y_o equation.[9] The models are *mixed-process* in that they combine equations

[8] That simultaneous causation is hypothesized in reality is what makes the models LIML rather than full-information maximum likelihood (FIML).

[9] Maddala (1983, pp. 117–25) describes models that mix latent and observed variables.

that have various types of censoring. And the models are *conditional* in that their specifics, such as the number of equations, vary by observation, being conditional on the data. A naïve approach to estimating the PK system is to use a Seemingly Unrelated Regressions (SUR) likelihood.[10] Within the y_o equation, this treats $\mathbf{y}_{fm}$ and $\mathbf{x}$ the same way mathematically, to that extent ignoring the endogenous nature of the credit variables. An underappreciated fact, which PK implicitly exploit, is that the naïve SUR is actually correct for fully-observed recursive systems (Roodman 2009b). Thus, for example, the standard SUR bivariate probit estimator is consistent and efficient for a two-stage, two-equation instrumental variable model in which both stages are probit (Greene 1998).[11] The econometric literature on recursive mixed-process models historically focused on multi-stage estimation procedures that are less computationally demanding than Maximum Likelihood (ML), if less efficient (e.g., Amemiya 1974; Heckman 1976; Maddala 1983, chs. 7–8; Smith and Blundell 1986; Rivers and Vuong 1988). Faster computers have made direct ML estimation more practical, and PK is a leading example.

As stated, the PK model assumes spherical errors. Of interest is how much this assumption can be relaxed. In fact, heteroskedasticity can render Tobit-type models inconsistent. To this important extent, PK implicitly assume homoskedasticity. They do, however, explicitly allow for correlations across observations within households—across seasons or, in individual-level regressions, across individuals—by computing clustered standard errors. In other words, they assume identically but not independently distributed errors.

Since c_f and c_m are the bases for all instruments in (2) and are instruments themselves, a key to this identification strategy, as PK emphasize, is that c_f and c_m are exogenous after conditioning on controls. Specifically, the factors driving credit choice—the formation of credit groups by village and gend-

[10] This is complicated because the likelihood for a given observation depends on the number of equations that are relevant and on which credit variables, if any, are censored. See PK's appendix and Roodman (2009b).

[11] Even in this simple case, Greene uses the phrases "surprisingly" and "seem not to be widely known" in asserting consistency.

er, and whether individual households are eligible—must be exogenous. Analyzing these assumptions economically and testing them econometrically are therefore important. PK do not appear to offer a reasoned defense of the exogeneity of the first factor. They do make one for the second, the exogeneity of landholdings: "Market turnover of land is well known to be low in South Asia. The absence of an active land market is the rationale given for the treatment of landownership as an exogenous regressor in almost all the empirical work on household behavior in South Asia" (p. 970). However, this appears to be a case for landholdings being *external* to the model (Heckman 2000). *Exogeneity* is a distinct notion (Brock and Durlauf 2001; Deaton 2009), requiring that landholdings are related to outcomes *only* through microcredit after linearly conditioning on controls. Meanwhile, one disadvantage of the LIML estimator is that it does not offer an easy way to test the assertion of instrument validity. In the Generalized Method of Moments framework (including 2SLS), the Hansen test is available for over-identified models such as these.

As Morduch notes, both of the key PK identifying assumptions are open to important questions. As for the first, regarding the formation of the credit groups by gender and village, PK recognize that unobserved factors could affect both group formation and outcomes, creating endogeneity. Their strongest response is to include village dummies to control for any such factors at the village level. Morduch's concern is about sub-village effects— that village effects are not fixed within villages. For example, in villages where the portion of *eligible* households is relatively well-off, credit group formation may be more likely and outcomes systematically better. In reply, Pitt (1999) acknowledges these potential nonlinearities by adding interaction terms between landholdings and all the **x** variables to PK's instrument set. If anything, it strengthens their results.

As for the exogeneity of the second factor inside the credit choice dummy, household landholdings, Morduch points out that (i) in the PK data land markets are in fact active and (ii) there is substantial and presumably endogenous mistargeting. We find that 203 of the 905 households in the 1991–92 sample that borrowed owned more than 0.5 acres before borrowing—1.5 acres on average. Evidently, loan officers were pragmatically bending the eligibility rule to extend credit to borrowers who seemed reliable and who were poor by global standards. Thus the *de facto* rule at work in the PK estimates is that any household that was *de jure* eligible *or* that borrowed was "eligible." Some of over-half-acre households that borrowed may have been met an alternative eligibility criterion (see footnote 6), but Lowess plots of borrowing probability against the area or value of landholdings among households only reinforce the impression of substantial mistargeting that runs counter to the banks' stated ideals. (See Figure 1 and Figure 2. The sample for each line is restricted to households in villages where microcredit is offered to people of the given sex.) Pitt's (1999) reply to Morduch points out that identification with LIML requires not that the rule be perfectly observed but that it drive an exogenous component of variation in borrowing. In a sense, Pitt casts the identification strategy as a Fuzzy Regression Discontinuity (FRD) design, albeit an unusual one that uses all observations, not just those near the threshold.[12] The upshot, though, is that both of the key claims behind the PK quasi-experimental design are asserted rather than being clear in the data.[13]

[12] PK footnote 16: "The quasi-experimental identification strategy used here is an example of the regression discontinuity design."

[13] Ito (1999) describes a mid-1990s Grameen Bank village in her doctoral dissertation: "One bank member I met outside my study area made no efforts to hide the fact that her husband had always owned 1.5 acres of land, which was three times as much as the Bank's targeting line. The woman explained it simply: '*The Bank informed us that we had to be 'bhumi-hin' (landless) to become a bank member. So we decided to call ourselves bhumi-hin ever since.*' Thus the Bank seems to be accepting almost any applicant whom current group members bring in, as long as one does not have a bad record with the Bank in the past."

Figure 1. Probability of borrowing vs. *area* of household land before borrowing (Lowess) , households with access to credit for given gender

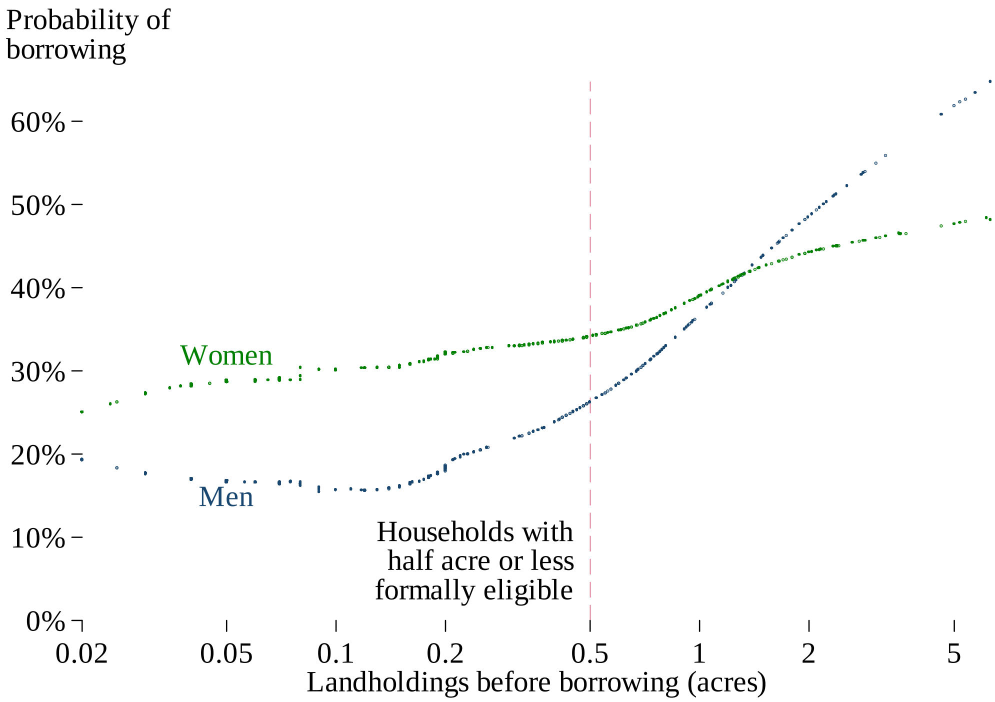
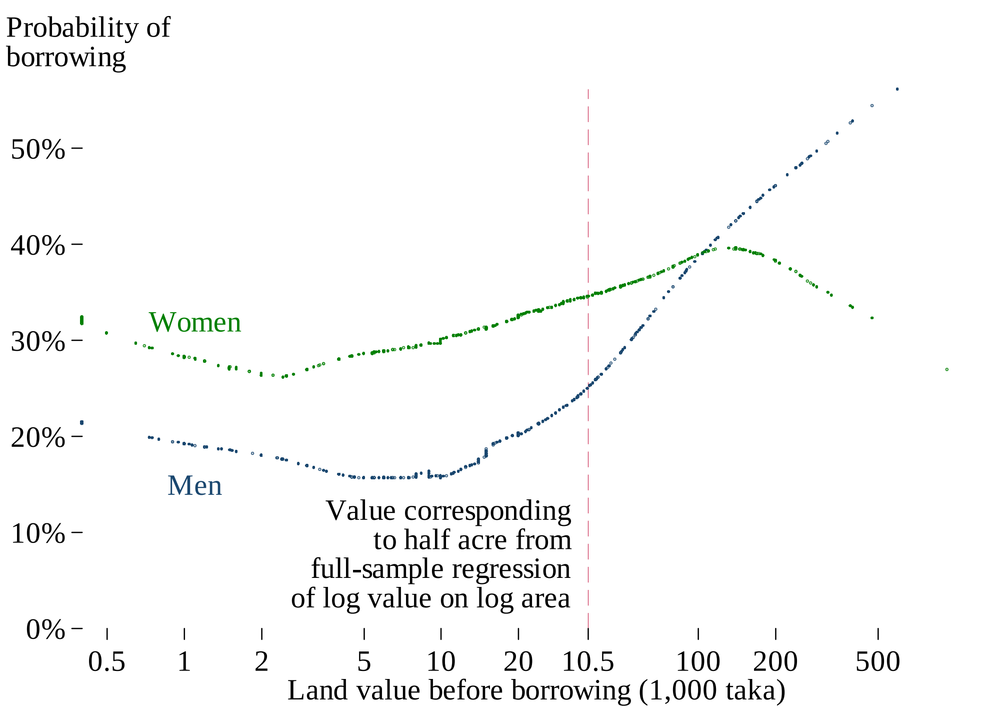


Figure 2. Probability of borrowing vs. *value* of household land before borrowing (Lowess), households with access to credit for given gender

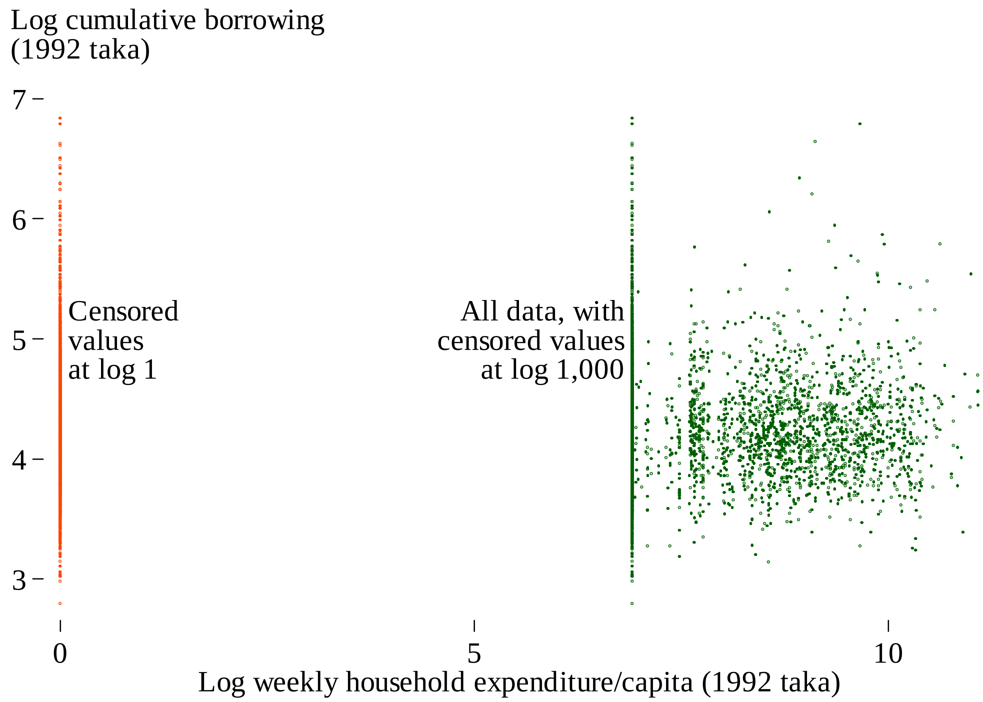


The PK credit variables are simple sums of all borrowing from the three microcredit programs since December 1986, adjusting for inflation; they are taken in logs and modeled as censored from below. This definition raises two subtle methodological questions. First, taking the simple sum of past borrowings implicitly imposes the assumption that borrowings in 1987, borrowings in 1988, etc., all affect consumption in 1991–92 with the same coefficient. In fact, we would expect the effects to vary over time. However, because borrowings in successive years tend to be collinear—typically, after paying off one one-year loan, a client immediately takes out a larger one—identifying the time profile within a five-year period would be difficult.

Second, modeling the log of cumulative borrowing as censored forces a choice about what small value the assumed censoring level should take. The difference between 1 and 10 *taka*, say, is minor in levels since most loans are thousands of *taka*, but major in logs. Although this issue is ultimately secondary to our conclusions, it may help explain large differences between the original regressions and our replications in the magnitudes of coefficients of interest (though not in the signs or significance). The lowest observed non-zero value for a credit variable is 1,000, and PK use 1,000 in a simplified example without logarithms in their appendix. For these reasons, we censor with log 1,000 ≈ 6.9. We have not ascertained what level the PK regressions use, but have reasons to think that it is log 1 = 0, the chief being that we get a better match in OLS using that value.[14] Figure 3 illustrates the issue with a scatter of cumulative female borrowing versus weekly household per-capita consumption using the full PK sample for all three survey rounds. The columns of dots at 0 and 6.9 correspond to the same data points and reflect different censoring values. One can see the reasonableness of log 1,000 as a censoring value; and how using log 1 would substantially flatten lines fit to the data, reducing coefficients even if not affecting signs or statistical distance from 0.

[14] A dataset provided by Mark Pitt includes some credit variables censored at log 1,000 and others at log 1. Pitt cautioned that this data set may not be exactly the same as PK's.

Figure 3. Household borrowing by women vs. household consumption, with censoring levels of log 1 or log 1,000



The replication

Using a new program written for Stata, called "cmp" for "conditional mixed process" (Roodman 2009b), we replicate all of the PK regressions, in the sense of applying the same methods to the same data. In the case of the household consumption outcome variable, which is continuous and unbounded, we also run 2SLS analogs motivated by the intuitions above. We first confirmed that our estimation software works properly on a simulated data set constructed by a program (sim7.do) included in Pitt (1999) (see Appendix). And we use the "cmp" program that performs the LIML to exactly match the output of half a dozen multi-equation commands written by the Stata Corporation, such as for Heckman selection models (Roodman 2009b).

We then begin the replication of the PK regressions by returning to the original survey data and reconstructing the data table used for estimation.[15] Predictably, benchmarking against the means and standard deviations in the PK appendix and the partial data set shared by Mark Pitt surfaces a few apparent errors on both sides. Coming second affords us the luxury of correcting ours before publication.[16] On the PK side, it appears that their female non-land assets variable actually includes land; the years-of-education variables treat current students has having completed no grades; the enrollment variable for children aged 5–17 is computed for 18-year-olds too; and the "cumulative borrowing since December 1986" variables include a few older loans. These problems do not turn out to be major concerns. Accounting for these differences, the match between the data sets appears to be very good. (See Table 1 and Table 2.) For right-hand-side variables, including credit variables, the means and standard deviations are close. Where we can compare at the observation level, almost all correlation coefficients exceed 0.97. (Not shown in either table is that the correlation for the dependent variable of central interest, log household per-capita consumption, is 0.995.) The same goes for left-hand-side variables; Table 2 shows only the aggregates from the new data set but can be compared to directly PK's Table A2. Subsample sizes match exactly and aggregates are close.

[15] The survey data for all rounds are now at go.worldbank.org/E9WWFZIXJ0.
[16] 329 of the 12,679.loans in the data were taken before December 1986. We correct the first two errors listed in text. But for simplicity and accuracy of replication we imitate the last two, which are minor. Correcting them does not materially affect our results.

Table 1. Weighted means and standard deviations of PK right-side variables, first survey round

	Reported in PK		New data set		
	Mean	Standard deviation	Mean	Standard deviation	Correlation[1]
Age of all individuals	23	18	23	18	
Schooling of individual aged 5 or above (years)	1.377	2.773	2.066	3.136	
Schooling of individual 5 or above (years, current students=0)			1.391	2.784	
Parents of household head own land?	0.256	0.564	0.254	0.563	0.992
# of brothers of household head owning land	0.815	1.308	0.810	1.305	0.978
# of sisters of household head owning land	0.755	1.208	0.750	1.206	0.988
Parents of household head's spouse own land?	0.529	0.784	0.529	0.783	0.986
# of brothers of household head's spouse owning land	0.919	1.427	0.919	1.427	0.980
# of sisters of household head's spouse owning land	0.753	1.202	0.753	1.202	0.985
Household land (in decimals)	76.142	108.540	76.145	108.052	0.999
Highest grade completed by household head	2.486	3.501	2.523	3.525	0.987
Sex of household head (1 = male)	0.948	0.223	0.948	0.223	0.998
Age of household head (years)	40.821	12.795	40.874	12.789	1.000
Highest grade completed by any female household member	1.606	2.853	1.664	2.999	
Highest grade completed by any male household member	3.082	3.081	3.277	4.016	
Highest grade by any female HH member (current students=0)			1.539	2.829	0.972[2]
Highest grade by any male HH member (current students=0)			3.046	3.805	0.991[2]
Adult female not present in household?	0.017	0.129	0.017	0.129	1.000
Adult male not present in household?	0.035	0.185	0.035	0.185	1.000
Spouse not present in household?	0.126	0.332	0.123	0.329	0.950
Amount borrowed by female from BRAC (taka)	350	1,574	349	1,564	0.988
Amount borrowed by male from BRAC (taka)	172	1,565	173	1,575	0.980
Amount borrowed by female from BRDB (taka)	114	747	114	746	0.978
Amount borrowed by male from BRDB (taka)	203	1,573	204	1,576	0.995
Amount borrowed by female from Grameen (taka)	956	4,293	972	4,324	0.986
Amount borrowed by male from Grameen Bank (taka)	374	2,923	360	2,895	0.957
Nontarget household	0.295	0.456	0.295	0.456	
Has any primary school?	0.686	0.464	0.686	0.464	
Has rural health center?	0.300	0.458	0.064	0.246	
Has family planning center?	0.097	0.296	0.097	0.296	
Is dai/midwife available?	0.673	0.469	0.673	0.469	
Price of rice	11.15	0.85	11.15	0.85	
Price of wheat flour	9.59	1.00	9.59	1.00	
Price of mustard oil	52.65	5.96	52.65	5.96	
Price of hen egg	2.46	1.81	2.46	1.81	
Price of milk	12.54	3.04	12.54	3.04	
Price of potato	3.74	1.59	3.74	1.49	
Average female wage	16.154	9.613	16.154	9.613	
No female wage dummy	0.193	0.395	0.193	0.395	
Average male wage	37.893	9.4	37.893	9.4	
Distance to bank (km)	3.49	2.85	3.49	2.85	

[1]Based on all three rounds from a household-level data set shared by Mark Pitt. [2]Correlations are with PK variables shown in previous pair of rows.

Table 2. Weighted means and standard deviations of PK endogenous variables, new data set

	Program villages				
	Participants	Non-participants	Total	Nonprogram villages	All
Cumulative female borrowing (1992 taka)	5,619.540		2,661.615		2,661.615
	(7,608.565)		(5,940.411)		(5,940.411)
	N = 779	N = 326	N = 1,105		N = 1,105
Cumulative male borrowing (1992 taka)	3,854.775		1,771.669		1,771.669
	(7,482.515)		(5,423.560)		(5,423.560)
	N = 631	N = 263	N = 894		N = 894
Current school enrollment of girls aged 5–18 years (yes = 1)[1]	0.535	0.528	0.531	0.552	0.534
	(0.499)	(0.500)	(0.499)	(0.498)	(0.499)
	N = 802	N = 434	N = 1,236	N = 225	N = 1,461
Current school enrollment of boys aged 5–18 years (yes = 1)[1]	0.566	0.555	0.558	0.553	0.557
	(0.496)	(0.498)	(0.497)	(0.498)	(0.497)
	N = 856	N = 468	N = 1,324	N = 267	N = 1,591
Women's labor supply (hours/month, aged 16–59 years)	40.390	32.438	35.068	31.238	34.446
	(70.532)	(64.283)	(66.512)	(60.202)	(65.540)
	N = 3,420	N = 2,108	N = 5,528	N = 1,074	N = 6,602
Men's labor supply (hours/month, aged 16–59 years)	202.747	185.779	191.252	180.604	189.371
	(100.817)	(104.870)	(103.872)	(99.400)	(103.168)
	N = 3,534	N = 2,254	N = 5,788	N = 1,126	N = 6,914
Per capita household total expenditure (taka/week)	76.537	85.250	82.376	88.993	83.475
	(44.862)	(64.986)	(59.241)	(66.212)	(60.498)
	N = 2,696	N = 1,650	N = 4,346	N = 872	N = 5,218
Female nonland assets (taka)	2,365.546	1,736.295	1,945.805	838.152	1,759.426
	(6,695.634)	(5,048.828)	(5,656.181)	(2,212.449)	(5,253.494)
	N = 899	N = 542	N = 1,441	N = 292	N = 1,733
Female assets (taka)	7,503.448	4,831.695	5,721.258	1,997.424	5,094.669
	(31,557.500)	(19,994.800)	(24,482.600)	(6,480.442)	(22,527.100)
	N = 899	N = 542	N = 1,441	N = 292	N = 1,733

Based on round 1 data. [1]PK report using school enrollment for ages 5–17, but 5–18 produces a near-perfect match with the enrollment variable aggregates in their Table A2.

As a first step in understanding the relationship between microcredit and household consumption in the 1991–92 data, Figure 4 and Figure 5 exhibit simple bivariate linear and Lowess regressions of past cumulative male and female microcredit borrowing against current household consumption per capita, using all three rounds of data and the PK weights and samples. We perform the Lowess plots to reveal some of the texture of the underlying data, not to make formal inferences. And in this spirit of data exploration, we reverse the roles that PK assign the credit and consumption variables—treating credit as dependent and putting it on the vertical axis—because it gives a clearer picture of potential selection biases.[17] Importantly, this reversal does not affect what interests us most, the signs of the slopes of certain best-fit lines that represent impact estimates. OLS regressions of y on x and x on y yield the same sign.

The first graph, Figure 4, covers "target" households only: all those owning less than half an acre, whether in program or non-program villages, and those with more than half an acre that borrowed anyway. The second graph covers the full sample.[18] Several facts become clear. First, the observed credit-consumption relationship differs by gender. Second, it is highly nonlinear. For the full sample of women, it is inverted-"U" shaped. This pattern is compatible with the frequently observed reality that the poorest are excluded (or self-excluded) from microcredit programs. Habibah, the powerful captain of a "center" of some 30 Grameen borrowers in the Tangail district of Bangladesh (and a borrower herself), explained how she thinks about member selection: "They should not be [too] landed, but they should own some land—some house land and some vegetable land. They should not be extremely poor. Most important, they should be hard working, not just the wife but also the husband" (Todd 1996, p. 173). The curve for men also tends toward an inverted "U," except that borrowing picks up at the high end. Finally, in moving from Figure 4 to Figure 5, adding the non-borrowing and generally affluent non-target households pulls down the right ends of all the contours. This is as it should be; but we note for future reference that the causal link here is almost certainly endogenous from the point of view of impact evaluation, running from being a household with a high consumption level to having a low (zero) probability of being a microcredit borrower.

[17] Plots with the axes reversed are available from the authors.
[18] All the analysis of PK copies them in excluding households with more than 5 acres—41 households in round 1 and 43 in rounds 2 and 3.

Figure 4. Household borrowing by women and men vs. household consumption, target households only (Lowess and linear)

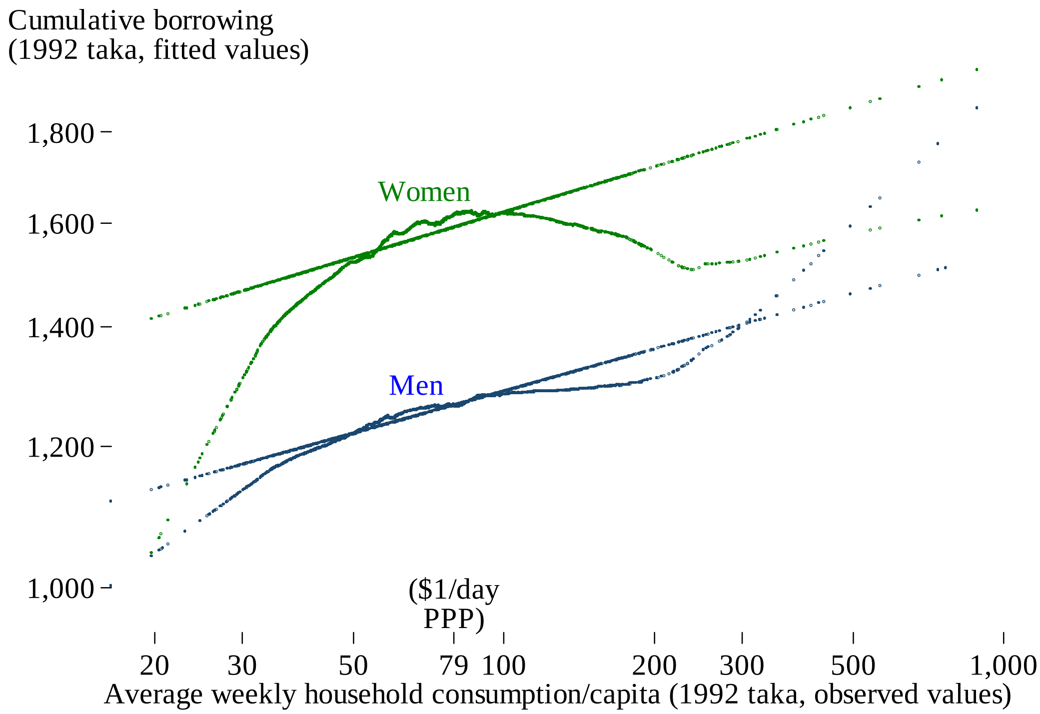


Figure 5. Household borrowing by women and men vs. household consumption, full PK sample (Lowess and linear)

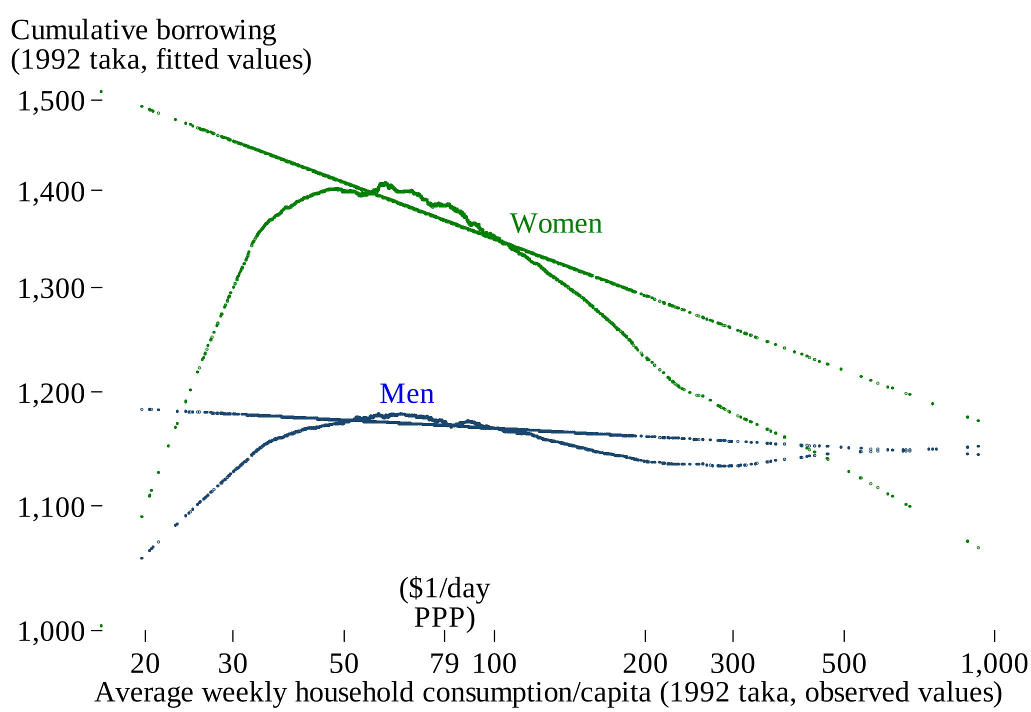


Figure 6 and Figure 7 have the same format, but are constructed to execute a 2SLS analog of the PK estimator.[19] Before graphing, the credit variables are linearly projected onto their instruments within the appropriate subsamples, according to (1). Then the controls—household characteristics, survey round and village dummies—are partialled out from the projected credit variables and household consumption. 2SLS is consistent (Kelejian 1971) but less efficient because it neglects the censored nature of credit. (On the other hand, it is superior in being robust to heteroskedasticity.) If the PK identifying assumptions hold, weighted linear fits to these residuals are consistent estimates of the impacts of female and male borrowing on household spending. These residuals are the bases for the graphs. For consistency with previous graphs, we regress the credit residuals on the consumption residuals rather than vice versa, so the lines reveal only the sign of the estimated impact. (In our formal analysis below we regress in the other direction, as an impact analysis demands.) In examining the two new figures, note first the continuities with the previous two. In all four, the best-fit lines for men and women seem distinct—though whether statistically so remains to be seen. And in the both pairs, adding non-target households pulls down the right ends of the best-fit lines. Finally, the slopes of the full-sample best-fit lines for women's credit (in Figure 5 and Figure 7) are both negative.

[19] The appendix of Pitt (1999) performs 2SLS in this way.

Figure 6. Household borrowing by women and men vs. household consumption, controlling for all covariates, target households only (Lowess and linear)

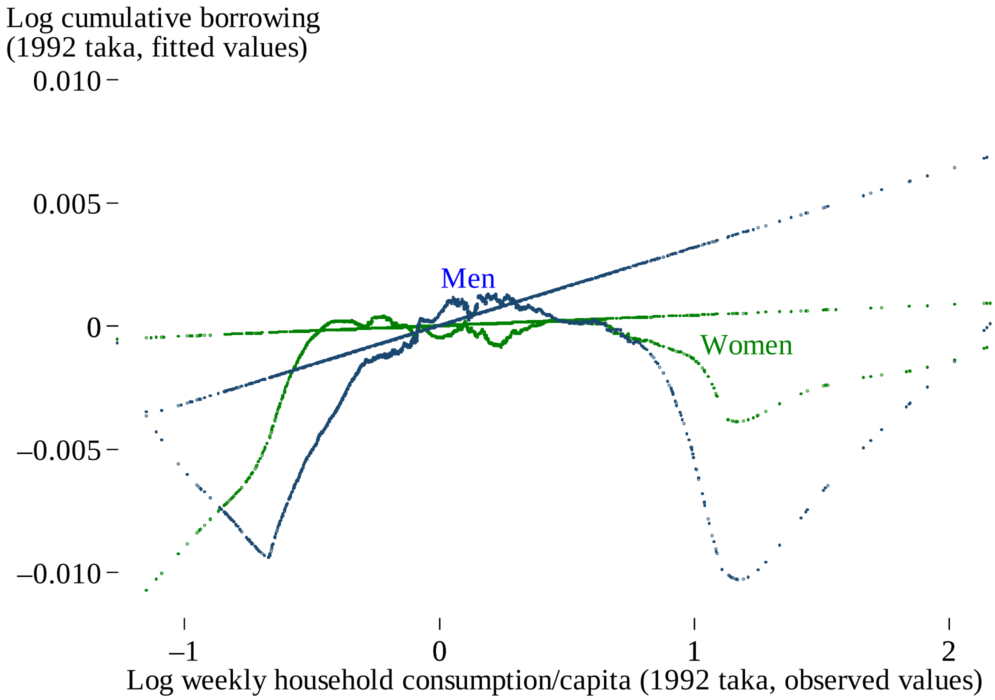
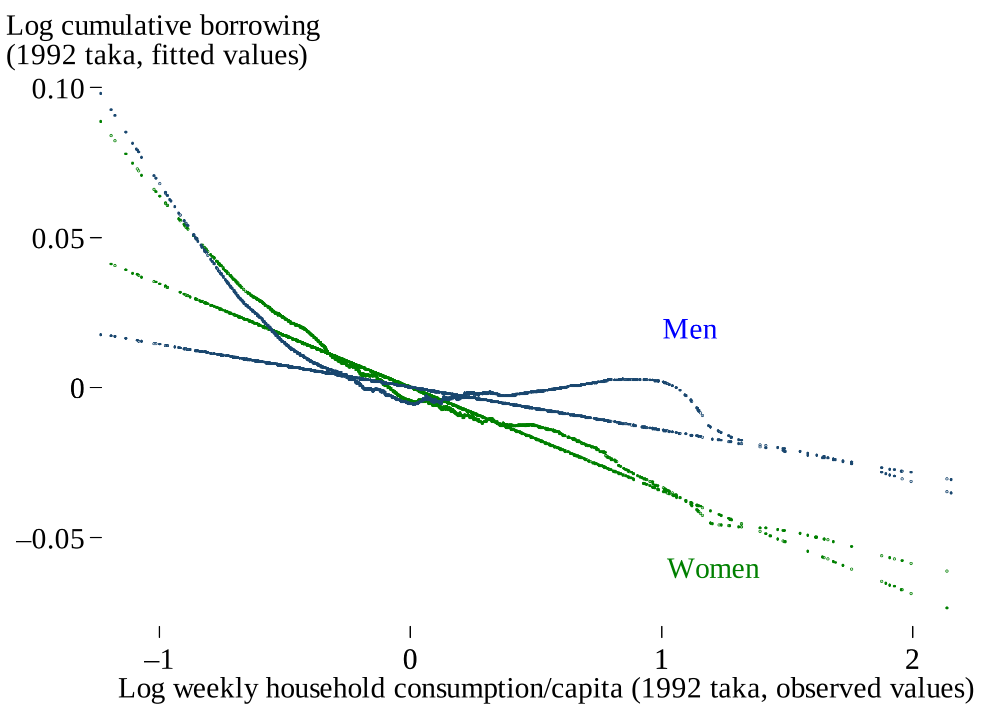


Figure 7. Household borrowing by women and men vs. household consumption, instrumenting and controlling all for covariates, full PK sample (Lowess and linear)



These graphs hint at several conclusions about the PK results. First, the negative slope on women's credit in Figure 7 contradicts the headline PK result to which it directly corresponds, namely their finding that lending a woman 100 *taka* raises household consumption by 18 *taka*/year. Meanwhile, adding non-target households—moving from Figure 6 to Figure 7—appears to perturb the parameter estimates implied by the best fit lines. To the extent that the instruments are valid and the causal relationship between credit and consumption within this added sample is endogenous, as argued earlier, this should not happen. That it does raises worries about the effectiveness of the instrumentation strategy. In the same vein, the differences throughout between male and female regression curves resonate with the tendency in the PK results for coefficients on the three male and the three female credit variables, as groups, to differ systematically from each other. It too hints that these differences reflect endogeneity.

The non- and semi-parametric regressions are meant to provide intuition and motivation. For more rigorous tests, we run all the PK household consumption specifications on both the target household and full samples, paralleling the graphs. (Where PK run "naïve" Ordinary Least Squares regressions on the target subsample and LIML on the full sample, we do both on both.) The highlights are in Table 3, which reports results from OLS; from LIML with controls for the 14 village characteristics listed at the bottom of Table 1; and from LIML with village dummies. Note that the LIML regression are identified even on the target subsample, because while e does not vary over the subsample, p_f and p_m, thus c_f and c_m still do; in other words, the PK assumption about the exogeneity of the gender status of credit availability suffices to identify the model.

Table 3. Estimates of impact of cumulative borrowing on log per capita household consumption, PK estimators, target households only

	Target households only			All households		
		LIML, controlling for HH characteristics and…			LIML, controlling for HH characteristics and…	
	OLS	Village characteristics	Village fixed effects	OLS	Village characteristics	Village fixed effects
Log female borrowing from BRAC	0.034	0.008	−0.021	0.016	−0.107	−0.103
	(2.468)**	(0.156)	(0.398)	(1.163)	(2.586)***	(2.696)***
Log male borrowing from BRAC	0.042	−0.010	−0.007	0.024	−0.022	−0.000
	(2.131)**	(0.219)	(0.165)	(1.194)	(0.437)	(0.007)
Log female borrowing from BRDB	0.016	−0.011	−0.029	−0.007	−0.142	−0.146
	(0.928)	(0.187)	(0.433)	(0.397)	(2.926)***	(2.940)***
Log male borrowing from BRDB	0.036	−0.016	0.032	0.022	−0.035	0.005
	(3.139)***	(0.501)	(0.915)	(1.962)**	(0.871)	(0.100)
Log female borrowing from Grameen	0.017	−0.006	−0.015	0.001	−0.099	−0.087
	(2.324)**	(0.151)	(0.356)	(0.101)	(3.200)***	(3.116)***
Log male borrowing from Grameen	0.000	−0.041	−0.008	−0.017	−0.052	−0.012
	(0.017)	(1.384)	(0.240)	(1.452)	(1.491)	(0.313)
Observations	4,567	4,567	4,567	5,218	5,218	5,218
Log pseudolikelihood	−2054.73	−6261.57	−5842.03	−2683.09	−7227.65	−6711.62

"Target households" includes those that would be eligible if credit programs operated in their villages. HH characteristics are: sex, age, and education level of household head; log landholdings before borrowing; how many parents, brothers, and sisters of household head or spouse own land (spearately); highest grade completed by any female or (separately) male household member; highest grade by any female or (separately) male HH member; dummies for survey rounds, whether no adult female or (separately) male is present in household, whether the HH head's spouse is not present, and whether the HH borrowed. Village characteristics are: separate dummies for whether a primary school, rural health center, family planning center, or midwife are available; prices of rice, wheat flour, mustard oil, hen's eggs, milk, potatoes; average female and (separately) male wage; dummy for female wage data availability; distance to nearest bank. Absolute *t* statistics (columns 1 and 4) and *z* statistics clustered by household (other columns) in parenthesis. All regressions are weighted. *significant at 10%. **significant at 5%. ***significant at 1%.

The regression results match the graphs. While the OLS point estimates for the target subsample do not match PK's, the pattern of significance is similar, putting statistically positive coefficients on all but the female-BRDB and male-Grameen credit variables. But in the regression that is meant to replicate the headline results (last column of Table 3), the coefficients on all three female credit variables are strongly negative. This is true too of the preceding regression with 14 village controls instead of village dummies. Comparing regressions in the two halves of the table, in every case adding non-target households reduces the coefficients on the credit variables, and for all but a single coefficient the reduction is statistically significant.

The sharp contradiction of PK's headline result poses a mystery. To check our results, we run the same estimation program on the data set provided by Mark Pitt.[20] The coefficients on female credit remain strongly negative. 2SLS regressions reported below also produce results of the same profile on Pitt's data set and ours. In an additional variant, we constrain the fit to match PK's published results; this reduces the maximum likelihood achieved. We also re-estimate using log 1 instead of log 1,000 for zero-observations of credit variables; this reduces coefficient magnitudes but by and large does not affect signs and significance.[21]

If the PK identifying assumptions hold, then both LIML fixed effects estimates in Table 3 are consistent; yet they are statistically different, the first essentially putting a 0 on female credit, the second a strong negative sign. This difference admits at least two explanations. One is that the effect of female credit on consumption is heterogeneous: its impact on target households is minimal, explaining the flat LIML results in the left half of Table 3 (and likewise in Figure 6), but the exclusion or self-exclusion of affluent non-target households is good for them, enough so that it makes the average "benefit" negative in the full sample. A second story, which we find more plausible, is that household decisions to borrow, as functions of household prosperity, are nonlinear and heterogeneous and differ by gender. This endogenous-causation theory would imply that the PK instrumentation strategy is not working as well as one would hope.

To examine the instrumentation, we run 2SLS analogs of the headline LIML fixed effects regression. Modeling on (2), we instrument with all the $c_f\mathbf{x}$ and $c_f\mathbf{x}$ interaction terms, where $\mathbf{x}$ includes village dummies. As noted earlier, 2SLS is consistent but less efficient under the PK assumptions. Using 2SLS opens the door to well-developed tests of instrumentation. The first column of Table 4 shows that the

[20] See note 14.
[21] Results available from the authors.

closest 2SLS analog provides a rough match to PK's headline LIML fixed-effects regression. The absolute *t* statistics on the female credit variables weaken as expected, to 1.4–1.9, but the coefficients are all negative and generally lower than the male credit coefficients. What is wholly new is the Hansen *J* test, which takes advantage of the overidentification to test instrument exogeneity. The test rejects the null hypothesis that the instruments are jointly valid at a *p* value of 0.038. In order to investigate which instruments are causing the trouble, we run difference-Hansen tests on various subsets and experiment with dropping them. The difference-Hansen tests reported in column 1 show where our suspicion settles: on the instruments that are interaction of the female and male credit choice dummies with a) the survey round dummies and b) the village dummies. In the right half of the table, we heed this cue about non-excludability by including these two groups of interaction terms as controls. Focusing on the first column in the right half, we see that both groups are reasonably, jointly significant according to *F* tests. On the one hand, this finding justifies Morduch's concern that village (as well as season) effects are not fixed between eligible and ineligible households: they are omitted variables in the PK specification. On the other, we find as Pitt does that including them actually strengthens our most significant results. In our case those results are negative coefficients on female credit.

The Hansen test, performed here with household-clustered standard errors, is robust to heteroskedasticity and autocorrelation. However, this generality also weakens the test. If we run the regressions separately for each round, which one observation per household, we can exploit PK's assumptions of homoskedasticity and error correlation only within households, to apply the more-powerful Sargan test. It is valid where errors are i.i.d. Columns 2–4 of both halves of Table 4 show these regressions and the associated Sargan tests. In the right half, we see that the regression that passes the Hansen test actually fails the Sargan tests for two out of the three survey rounds.

Table 4. PK-analogous 2SLS estimates of impact of cumulative borrowing on log per capita household consumption, all households

	Rounds 1–3	Round 1	Round 2	Round 3	Rounds 1–3	Round 1	Round 2	Round 3
Log female borrowing from BRAC	−0.122	−0.109	−0.119	−0.075	−0.191	−0.070	−0.209	−0.450
	(1.439)	(0.917)	(1.089)	(0.755)	(0.905)	(0.252)	(0.825)	(1.584)
Log male borrowing from BRAC	0.213	0.308	0.062	0.208	0.482	0.291	−0.126	0.748
	(1.825)*	(2.186)**	(0.365)	(1.554)	(1.434)	(0.841)	(0.300)	(1.642)
Log female borrowing from BRDB	−0.304	−0.037	−0.543	−0.234	−1.209	−0.545	−0.859	−1.118
	(1.890)*	(0.192)	(2.409)**	(1.403)	(2.419)**	(0.866)	(1.515)	(2.014)**
Log male borrowing from BRDB	−0.135	−0.244	0.009	−0.237	−0.462	−0.335	0.463	−0.616
	(1.030)	(1.399)	(0.060)	(1.522)	(2.247)**	(1.356)	(1.479)	(1.659)*
Log female borrowing from Grameen	−0.057	−0.103	−0.043	−0.004	0.171	0.393	0.208	0.157
	(1.474)	(1.639)	(0.893)	(0.092)	(1.145)	(2.184)**	(1.189)	(0.882)
Log male borrowing from Grameen	−0.063	−0.141	−0.015	−0.055	−0.032	−0.001	0.035	0.030
	(0.920)	(1.730)*	(0.161)	(0.567)	(0.318)	(0.006)	(0.213)	(0.205)
Interaction terms using								
Survey round dummies (F test p value)					0.136			
Village dummies (F test p value)					0.000	0.000	0.000	0.000
Observations	5,218	1,757	1,735	1,726	5,218	1,757	1,735	1,726
Tests of joint validity of instruments								
Sargan, all instruments (p value)		0.000	0.000	0.000		0.000	0.006	0.688
Hansen, all instruments (p value)	0.038	0.012	0.058	0.046	0.927	0.532	0.895	0.968
Diff-Hansen, interaction terms using								
Survey round dummies (p value)	0.090							
Village dummies (p value)	0.106	0.160	0.035	0.124				

Analogously with the PK LIML fixed effects regression, all regressions instrument with interactions of male and female credit choice dummies with household characteristics, survey round dummies, and village dummies. The second set includes the interactions with round and village dummies as controls. The PK regression requires homoskedasticity for consistency, but allows serial correlation in the errors; under these assumptions, errors within each survey round are i.i.d., making Sargan tests valid for the regressions on single-round samples. The Hansen test does not require sphericity, making it consistent for the three-round regressions as well, but is weaker. The Difference-Sargan/Hansen test for validity of instrument subsets is based on Hansen tests for the first column and Sargan tests for the remainder. Unreported controls are as in previous table. All regressions are weighted. Absolute t statistics clustered by household in parenthesis. *significant at 10%. **significant at 5%.

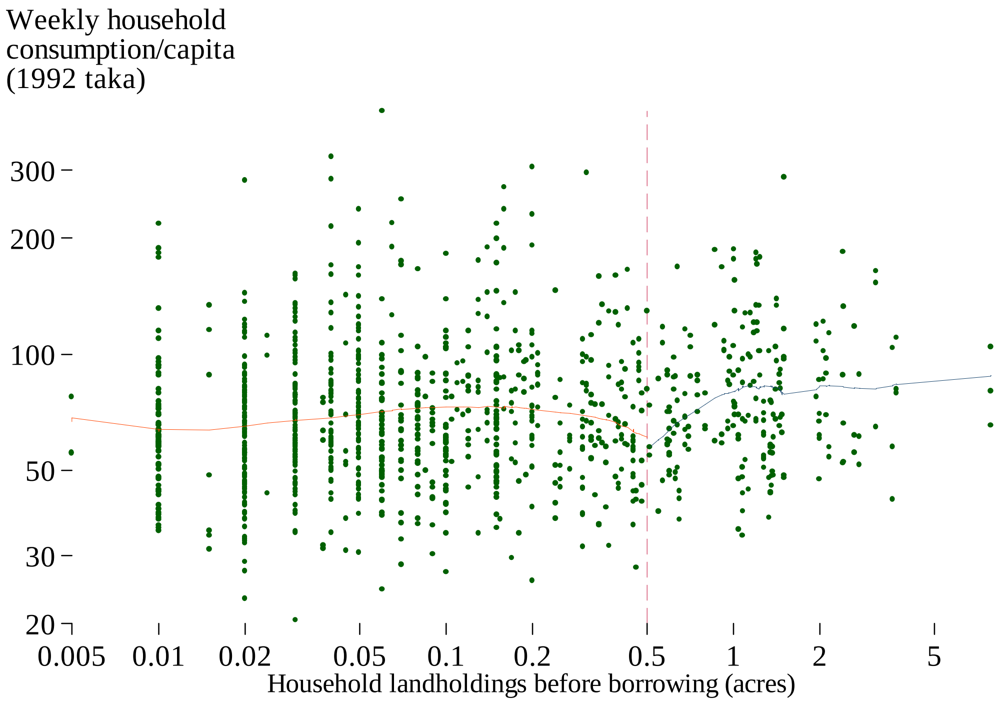
The failures on the Sargan tests can be interpreted in two ways. The assumption of homoskedasticity does not hold, in which the Sargan test should not be trusted, or it does hold and the excluded instruments are invalid. Either possibility would undermine the PK estimator. Only the latter would undermine the 2SLS estimate in column 5 of Table 4. So perhaps that regression is evidence that microlending to women reduces household consumption. Given all the doubts raised, though, we are not ready to conclude that microcredit does harm.

A more standard Fuzzy Regression Discontinuity design might side-step the endogeneity concerns by restricting to households closer to the formal threshold eligibility value of 0.5 acres of land. But the more we focus around the threshold the more the mistargeting identified by Morduch comes to the fore. Following the advice of Imbens and Lemieux (2008), we start an FRD analysis by plotting the outcome of interest, household consumption per capita, against the continuous forcing variable in the model, household landholdings before borrowing. We add Lowess smoothed plots, but separately for the below- and above-threshold subsamples in order to allow for a discontinuity at the half-acre mark. We construct this graph first for all villages with a microcredit program; then, in order to narrow the focus by gender, for those where only women could borrow and for those where only men could borrow. Figure 8 is the plot for the female-only villages. The vertical line at log 0.5 ≈ –0.69 marks the threshold. The discontinuity in the outcome at the threshold is small compared to the variation in the data. (We expect some discontinuity by chance since the two Lowess curves are fit to different data.) Imbens and Lemieux warn that "if the basic plot does not show any evidence of a discontinuity, there is relatively little chance that the more sophisticated analyses will lead to robust and credible estimates with statistically and substantially significant magnitudes." Indeed, when we perform a formal FRD analysis using 2SLS, as suggested by Hahn, Todd, and Van der Klaauw (2001), we find little evidence of significance for the coefficient on microcredit in female-only villages.[22] Varying the sample retained between 1% and 50% of available observations, the largest absolute *t* statistic is 0.86—or 1.27 if PK's controls, including village dummies, are added. Graphical and 2SLS results for male-only villages and for all program villages are very similar.[23]

[22] Hahn, Todd, and Van Der Klaauw show that when the same observations are retained for the outcome and forcing variables, and when the weighting on them is uniform, the FRD estimate can be computed by a 2SLS regression of the outcome on x, the forcing variable, instrumenting with the dummy $1\{x \geq c\}$, where c is the threshold, and controlling for $1\{x < c\}\cdot(x - c)$ and $1\{x \geq c\}\cdot(x - c)$.

[23] Results are available from the authors.

Figure 8. Household consumption versus landholdings before borrowing in female-only credit program villages, with separate Lowess plots for subsamples above and below half-acre



We replicate the PK regressions for other outcomes too. (See Table 5, which reports results from PK's preferred weighted LIML fixed-effects specification.) We concur in finding little effect on school enrollment of girls or boys. The same goes for the value of female-owned assets, which PK may have unintentionally studied rather than female *non-land* assets. On the other hand, our replications differ in finding a strong positive association between female (not male) borrowing and female-owned non-land assets; a strong negative association between male (but not female) borrowing and female labor supply; and no association with male labor supply, where PK found a strong negative effect. We have not investigated these regressions in the same depth. Certainly, the difficulties with the consumption regressions make us cautious about inferring causality from the other ones. And endogenous-causation stories can easily explain our results. For instance, Figure 5 suggests that male borrowing is lowest in the poorest households, where women may work more as a matter of survival.

Table 5. Weighted LIML fixed-effect estimates of impact of microcredit on various outcomes, following PK

	Log female non-land assets	Log female hours worked per month	Log male hours worked per month	School enrollment of girls, 5–18	School enrollment of boys, 5–18
Log female borrowing from BRAC	0.604	–0.130	0.288	–0.193	–0.229
	(2.074)**	(0.418)	(0.942)	(0.945)	(1.341)
Log male borrowing from BRAC	0.019	–0.662	–0.237	–0.038	–0.138
	(0.050)	(1.924)*	(0.290)	(0.129)	(0.727)
Log female borrowing from BRDB	1.024	–0.114	0.125	–0.178	0.088
	(1.975)**	(0.293)	(0.333)	(0.615)	(0.344)
Log male borrowing from BRDB	–0.386	–0.581	–0.233	–0.083	0.065
	(1.178)	(1.973)**	(0.274)	(0.417)	(0.300)
Log female borrowing from Grameen	0.679	0.129	0.120	–0.105	–0.028
	(3.077)***	(0.570)	(0.515)	(0.737)	(0.207)
Log male borrowing from Grameen	–0.243	–0.548	–0.273	–0.029	0.150
	(1.026)	(2.385)**	(0.489)	(0.183)	(0.946)
Observations	1,757	6,537	6,835	1,453	1,573
Log pseudolikelihood	–4039.19	–14889.30	–18267.10	–1836.16	–2033.35

Regressions run on household-level data for first column and individual-level for remainder. All use round 1 data only and are weighted. PK report regressing school enrollment for ages 5–17, but 5–18 produces a near-perfect match with the enrollment variable aggregates in their Table A2. Absolute *z* statistics clustered by household in parenthesis. *significant at 10%. **significant at 5%. ***significant at 1%.

In sum, we come away from the PK study with doubts about the magnitude, sign, and direction of the reported effects of microcredit. We do not necessarily doubt microcredit itself, but we doubt the result that emerged from analyzing the 1991–92 Bangladesh survey.

Morduch (1998)

The Study

Morduch critiques PK and offers new evidence, notably on the connection between credit and the volatility of household consumption and labor supply. Having just critiqued PK, we focus here on replicating the novel results in Morduch.

Morduch’s estimation strategy is simpler and less efficient than PK’s, but analogous. He uses sampling weights and nearly the same control sets. The major departure is that rather than instrumenting credit in a LIML framework, he regresses directly on the primary instruments for credit, dummies for credit choice. Rather than distinguishing borrowing by gender, he splits by the lending program, leading to three variables of interest: dummies for the availability of credit from Grameen, BRAC, and the BRDB to at least one gender in a given village. Morduch first performs simple difference-in-difference estimates, then adds controls.

Morduch fails to confirm the PK results on household consumption. His OLS regression with the full control set including village effects puts *t* statistics of –1.48 on Grameen credit access, +0.41 on BRAC access, and –1.71 on BRDB access. The hint of negativity is consistent with the results in our Table 4, especially considering that Morduch's program-wise division mixes the coefficients on credit to women, which we find to be negative, with those for men, which we cannot distinguish from zero. Morduch, however, finds hopeful evidence that microcredit is affecting the second moment of consumption over the three seasonal rounds of the 1991–92 surveys, with *t* statistics of –1.95, –1.42, and –1.96 in a specification with village dummies. Consumption volatility is extremely important for the poor since how often children go to bed hungry matters at least as much as whether they are well-fed on average (Morduch 1994, 1995). Morduch also finds somewhat weaker evidence (with *t* statistics of –1.78, –1.35, and –1.85) that households with access to credit are actively managing and smoothing their labor income, not just their spending. He asserts, without direct evidence, that it is the ability to smooth income over the year which drives smoother within-year consumption.

The Replication

Our replication data set matches Morduch's original quite well, not surprisingly. Still, the rebuilding a data set again exposed a few errors in the original, mostly affecting the labor supply variables.[24] In our replication, the minor changes turn out to strengthen two of the three negative signs on credit for *average* consumption, reinforcing our analysis of PK, but weakening what were arguably marginal results on

[24] For instance, Morduch's construction of the enrollment and labor supply variables omitted individuals reaching school age (5) or adulthood for purposes of labor supply (16) after survey round 1 but before round 3.

labor supply. (See Table 6, which can be compared directly to Morduch's Table 13.)

Table 6. Replication of Morduch regressions with controls

	Target households, controlling for household characteristics			Target households, controlling for household & village characteristics			All households, controlling for household characteristics and village fixed effects		
	Grameen	BRAC	BRDB	Grameen	BRAC	BRDB	Grameen	BRAC	BRDB
Log consumption/capita	–0.042	–0.026	–0.078	–0.062	–0.031	–0.065	–0.097	0.024	–0.144
	(0.89)	(0.56)	(1.92)*	(1.16)	(0.63)	(1.60)	(1.54)	(0.42)	(2.02)**
Variance of log consumption/capita	–0.008	–0.005	–0.010	–0.013	–0.006	–0.009	–0.035	–0.045	–0.038
	(0.71)	(0.46)	(0.89)	(1.01)	(0.50)	(0.85)	(1.46)	(1.19)	(1.54)
Log labor per adult in past month	0.056	–0.075	0.020	0.091	–0.068	0.017	–0.068	–0.139	0.119
	(1.21)	(1.52)	(0.43)	(1.36)	(1.14)	(0.38)	(0.70)	(1.37)	(1.12)
Variance of per adult log labor	–0.024	0.008	–0.006	–0.072	–0.006	–0.014	–0.046	–0.081	–0.031
	(0.83)	(0.30)	(0.21)	(2.09)**	(0.21)	(0.50)	(0.81)	(1.12)	(0.56)
Adult male labor hours in past month	16.23	3.99	10.30	10.11	–6.50	5.67	–9.23	–9.91	9.90
	(2.19)**	(0.49)	(1.35)	(0.95)	(0.59)	(0.71)	(0.62)	(0.71)	(0.73)
Adult female labor hours in past month	1.78	–19.49	–0.74	13.69	5.06	8.91	–9.25	–17.39	12.10
	(0.12)	(1.50)	(0.06)	(0.85)	(0.42)	(0.97)	(0.71)	(1.23)	(0.80)
% males in school (age 5–17)	0.88	–2.12	–5.55	7.18	7.14	–2.12	–1.35	–2.05	–12.23
	(0.14)	(0.34)	(0.90)	(0.95)	(1.00)	(0.35)	(0.13)	(0.19)	(1.34)
% females in school (age 5–17)	–5.59	–2.24	–15.85	–12.53	–11.14	–20.78	2.57	5.51	–0.81
	(0.84)	(0.33)	(2.32)**	(1.93)*	(1.77)*	(3.59)***	(0.23)	(0.56)	(0.08)

Unit of observation is the household for the top half of the table and the individual for bottom half. All regressions are OLS, except for the male and female labor hours ones, which are Tobit. All regressions are weighted. Absolute *t* statistics robust to intra-household correlation in parenthesis. *significant at 10%. **significant at 5%. ***significant at 1%.

The changes also weaken the findings on consumption volatility, reducing the *t* statistics on Grameen and BRDB credit from –1.95 and –1.96 in the original to –1.46 and –1.54 (right pane of Table 6). This result, however, still appears to be more than noise, though we caution against interpreting it as evidence of causation from credit to volatility. Table 7 shows why: it replicates Morduch's difference-in-difference analysis (without controls) of the relationship between credit availability and the variance of log household per-capita consumption over the three seasons, excluding mistargeted households. The estimates in the bottom right of the table are strongly negative. But they are driven not by lower volatility in treatment households but higher volatility in *ineligible households in program villages*. Concretely, five of the eight core numbers in the upper left are about the same, with the three for non-target households in Grameen, BRAC, and BRDB villages the odd ones out. To interpret the difference-in-difference as impact measures, we must believe that the households with access to credit would, lacking that access, have experienced the same volatility as their affluent neighbors, and well more than their target brethren in non-program villages. The fact that volatility for these households dropped to about the level experienced by target and non-target households in non-program villages (about 0.6–0.7) would then have to be a coincidence. A competing and arguably more parsimonious explanation is that non-target households in villages where credit programs had chosen to operate are systematically different both from target households in those villages and from all households in villages where the programs did not operate. That would fit with our findings above about the non-excludability of credit choice–village dummy interactions. Buttressing this interpretation is the fact that the volatility comes mainly from rare but large expenditures on land, home improvement, and social and religious ceremonies, perhaps including dowry, which for some reason are reportedly rarer among the non-target households in the five control villages than among non-target (and non-borrowing) households in program villages. Control villages were home to 45 of the sample's 256 non-target, non-borrowing households yet account for only one of the 25 largest individual purchases reported in this class (See Table 8.) Fundamentally, the volatility of expenditures among the better off is a debatable benchmark for volatility among the poorest.

Table 7. Replication of Morduch difference-in-difference estimates of variance of log household consumption/capita

	Grameen	BRAC	BRDB	Control	Difference: Grameen	Difference: BRAC	Difference: BRDB
Under 0.5 acre	0.061	0.062	0.055	0.069	–0.008	–0.007	–0.013
					(1.00)	(0.85)	(1.52)
Over 0.5 acre	0.112	0.131	0.117	0.069	0.043	0.062	0.048
					(1.86)*	(1.70)*	(2.15)**
Difference	–0.052	–0.069	–0.062	–0.000	–0.05	–0.07	–0.06
	(2.41)**	(1.96)*	(2.98)***	(0.03)	(2.08)**	(1.84)*	(2.55)**

Absolute *t* statistics robust to intra-household correlation in parenthesis. *significant at 10%. **significant at 5%.

Table 8. Top 25 individual expenditures all non-target, non-borrowing households, all 1991–92 survey rounds

Expenditure	Amount	Survey round	Village credit program	Consumption/ capita (taka/week)	Land (acres)	Upazilla	District	Division
Home Improvent	130,000	3	Grameen	281	1.3	Sonargaon	Narayanganj	Dhaka
Land/Property Purchase	125,000	1	BRAC	1,735	6.5	Habiganj Sadar	Habiganj	Sylhet
Miscellaneous	72,000	2	BRDB	194	57.5	Birganj	Dinajpur	Rajshahi
Land/Property Purchase	65,000	1	Grameen	438	13.3	Sreepur	Gazipur	Dhaka
Land/Property Purchase	53,000	3	BRAC	86	1.5	Sreebardi	Sherpur	Dhaka
Land/Property Purchase	53,000	3	BRDB	116	2.1	Fakirhat	Bagerhat	Khulna
Home Improvent	50,000	3	Grameen	235	3.1	Sonargaon	Narayanganj	Dhaka
Home Improvent	50,000	2	Grameen	281	1.3	Sonargaon	Narayanganj	Dhaka
Home Improvent	50,000	1	BRAC	666	1.4	Rangpur Sadar	Rangpur	Rajshahi
Public Transport	45,000	2	BRAC	168	1.3	Kalaroa	Satkhira	Khulna
Home Improvent	40,000	2	Grameen	235	3.1	Sonargaon	Narayanganj	Dhaka
Servants Wage	37,000	1	BRDB	614	0.9	Muktagachha	Mymensingh	Dhaka
Land/Property Purchase	35,000	2	BRDB	614	0.9	Muktagachha	Mymensingh	Dhaka
Social/Religious Ceremony	35,000	2	BRAC	189	6.8	Habiganj Sadar	Habiganj	Sylhet
Medicine	35,000	2	BRDB	142	2.1	Fakirhat	Bagerhat	Khulna
Home Improvent	30,000	2	Grameen	103	2.4	Sakhipur	Tangail	Dhaka
Servant's Wages	28,000	1	BRAC	254	1.2	Kalaroa	Satkhira	Khulna
Land/Property Purchase	27,000	2	BRAC	123	1.7	Sreebardi	Sherpur	Dhaka
Land/Property Purchase	26,000	2	Grameen	121	6.0	Raiganj	Sirajganj	Rajshahi
Marriage/Birth/Death	25,000	2	BRDB	66	3.1	Shibganj	Bogra	Rajshahi
Ceremony	25,000	1	Grameen	381	5.0	Jaldhaka	Nilphamari	Rajshahi
Land/Property Purchase	24,000	2	None	132	0.9	Jhenaidah Sadar	Jhenaidah	Khulna
Servant's Wages	24,000	1	BRAC	288	1.5	Manikganj Sadar	Manikganj	Dhaka
Land/Property Purchase	24,000	3	BRDB	162	8.3	Birganj	Dinajpur	Rajshahi
Land/Property Purchase	24,000	3	BRDB	66	3.1	Shibganj	Bogra	Rajshahi

1 taka ≈ $0.10 in 1992. Sample-average weekly household consumption/capita is 83 taka.

Overall, although we share the puzzlement in Morduch over the inability to replicate PK's positive findings for the effects of microcredit on the level of household consumption, we do not find powerful evidence for effects on its variability either.[25] Because Morduch's regression are exactly identified, not overidentified like PK's, we cannot apply the Hansen *J* test. But the same doubts about the validity of the three implicit instruments—availability of credit from each of the programs studied—pertain.

Khandker (2005)

The Study

In 1999, surveyors in Bangladesh sought to revisit the 1,769 households that persisted through all three 1991–92 data collection rounds. For 1,638, they found the original household or one or more successors, yielding an attrition rate of just 7.4%. Of the original households, 237 households had split, yielding 546 new ones. Confronted with the conceptually complex problems of attrition and dissolution of the unit of observation, Khandker's response is straightforward: amalgamate split households for purposes of analysis and drop attritors from all rounds.

The potential for endogenous attrition raises worries about bias. On the one hand, Thomas, Frankenberg, and Smith (2001) argue from Indonesian household survey data that attritors who move long distances differ statistically from those they leave behind, and are worth trying to follow. On the other, in tests on longitudinal household data from Bolivia, Kenya, and South Africa, Alderman *et al*. (2001) find little bias in practice. Khandker reports formally testing, in an uncirculated paper by Khandker and Pitt, for attrition and amalgamation biases and finding that both issues are largely ignorable.

As Khandker notes, the availability of panel data raises the hope of eliminating one potential source of bias in the PK and Morduch cross-section analyses, namely that unobserved but fixed household and individual characteristics simultaneously affect microcredit borrowing and outcomes of inter-

[25] PK-style LIML FE regressions for household consumption variability find no effect for male borrowing but a *positive* effect for female borrowing. Results available from the authors.

est. In particular, differencing can respond to the concern that village effects are not "fixed" within villages—to the extent that they *are* fixed over time, which is plausibly more true. Khandker explicitly distinguishes the panel approach he takes from PK's quasi-experimental design. Indeed, any claims for an exogenous component in the allocation of credit weakened over the 1990s. By 1999, every study village had access to microcredit, at least for women, so variation in the choice variables declined; and, as Khandker documents, formal mistargeting of credit to above-half-acre households actually increased.

Khandker points out, however, that individual- and village-level effects may not be fixed, and that other sources of endogeneity may remain, so he starts with 2SLS regressions that instrument like ours with interaction terms between choice dummies and the **x** variables. Khandker treats the 1991–92 data as a single time period. Adding the 1999 data and including individual fixed effects gives a cross-section in differences. Reflecting the new time dimension, the regressions feature four credit variables: "current" female and male borrowing (i.e., cumulative borrowing since the first survey rounds) and "past" female and male borrowing (i.e., cumulative borrowing between late 1986 and 1991, as in PK).[26] And whereas in our 2SLS regressions (above) we interacted with the female and male choice dummies for instruments, Khandker interacts with a pair of dummies differentiated along the time dimension: one for whether household members of either gender could borrow in 1991–92, and the same for 1999.[27]

Khandker studies three outcomes: household food consumption, non-food consumption, and total consumption, all in inflation-adjusted *taka* per year. The control set is nearly identical to that in PK's non–fixed effects specifications, including time-varying village-level variables. Unlike PK, Khandker includes households owning more than 5 acres. The 1991–92 sampling weights are used throughout.

[26] These too formally enter in differences, but in practice they can also be seen as entering undifferenced. The value for twice-lagged cumulative borrowing is not observed—it would cover a period in the first half of the 1980s, it is assumed to be zero, perhaps not unreasonably, since microcredit was less common then. The lagged difference of cumulative borrowing is then just the lagged level. And, conditioning on this past level, regressing on the current difference is tantamount to regressing on the current level.

[27] By 1999, all villages had credit programs, so the later dummy merely indicates whether households are eligible. Khandker appears to treat mistargeted households as eligible.

Khandker also performs OLS regressions in parallel with the 2SLS ones. A Wu-Hausman test fails to reject the hypothesis that the results from the two estimators differ, so he reports only OLS.

Khandker then builds on the foundation of his core OLS regressions. First, he adds average borrowing in a village as a regressor in order to test for spill-over effects, which he finds for women's borrowing. Then he feeds the results on the benefits of female borrowing for households and villages into a retroactive simulation to study the effects of microcredit on households by poverty level. Here, he distinguishes between the "moderately" and "extremely poor."[28] Starting from observed consumption and borrowing levels, he calculates that in aggregate microcredit reduced the moderate poverty rate by 1.0 percentage point per year, equivalent to 40% of the total decline in Bangladesh over the 1990s; and extreme poverty by 1.3 percentage points a year. This extreme-moderate differential arises mainly from the fact that different households borrowed different amounts. It does *not* arise from an econometric estimate that allows separate impact elasticities for the two groups. Nor does it come from the fact that the elasticities that are estimated imply different marginal effects at different consumption and borrowing levels, because Khandker assumes a fixed average impact for the simulation.

Buried in the shift to the panel set-up are at least two issues relating to the recurring theme of whether there is a credible source of exogenous variation in credit. First, the shift to a panel estimator only reduces the need for an exogenous source of variation in borrowing to the extent that endogeneity of all types is removed by differencing. Khandker's 2SLS regressions are premised on the assumption that the particular family of interaction terms used as instruments embodies such variation—and no more. But this assumption is not grounded in economic reasoning: the Khandker paper distances itself from any claim to quasi-experimental variation. And, as in all the papers replicated here, the assumption is not tested.

[28] Khandker (1998, p. 55) defines moderate poverty as household consumption below 5,270 *taka*/person/year and extreme poverty as 80% of that, 3,330 *taka*/person/year.

The second issue relates to time-varying effects. In PK's cross-section analysis, using 14 village-level controls is less conservative than entering 86 village dummies, which is why PK prefer the latter. Dummies express our ignorance about the many village-level factors that affect both credit and outcomes. As PK explain: "These attributes include prices, infrastructure, village attitudes, and the nature of the environment, including climate and propensity to natural disaster. For example, the proximity of villages to urban areas may influence the demand for credit to undertake small-scale activities but may also affect household behavior by altering attitudes." Yet when we move from the cross-section to the time series as the locus of identification, we meet a paradox: controlling for a handful of concrete but time-varying village controls is *more* conservative than using a much larger set of village fixed effects. In the case at hand, time-varying village variables such as the rice price usefully remain in the model after differencing. Village fixed effects disappear. The core problem is that few if any of the factors rightly cited by PK in arguing for modeling with village fixed effects are in fact fixed. Sadly, climate changes. Practical proximity to cities depends on road quality.

There is a way out of the paradox: where entering village dummies is conservative in the PK estimation set-up, entering them in the Khandker set-up *after* other variables are differenced is the conservative analog. In the model, this would allow all unknown village-level factors to vary in impact over the 1990s. In fact, Khandker essentially does this in the first stage of his 2SLS regressions since the instrument sets include interactions with village dummies. The question is whether it is proper to exclude village dummies from the second stage.

The Replication

In replicating Khandker, we run into a problem opposite that we had with PK: our summary statistics for key variables do not match the original nearly so precisely (see Table 9) but we easily replicate the pattern of core results, with strong positive coefficients on current and past borrowing by women for total household consumption (first column of Table 10).[29] A 2SLS regression replicating the one Khandker describes but does not report produces even stronger results (column 3 of Table 10). We then proceed to check some of the methodological issues just raised. First, we introduce village dummy controls after differencing (columns 2 and 4). This substantially weakens the results for female credit, but perhaps does not destroy them. In 2SLS at least, the coefficient on women's past loans becomes very large and remains statistically strong (column 4).

But here we encounter a new concern: the first Hansen test is clearly rejecting the hypothesis that the Khandker instrument set is valid. Thus the fact that the OLS results fit with the 2SLS ones, the crux of Khandker's argument, is not so reassuring. The premise of the Wu-Hausman test, that 2SLS is consistent, appears violated. So, much as with PK (Table 4), we enter the instruments based on village dummies as controls (final two columns). Their joint significance is clear, and the 2SLS no-FE regression (column 5) does better on the Hansen test. On the other hand, the 2SLS FE regression (column 6) produces a perfect Hansen *p* value of 1.000, a sure and unsurprising sign that overinstrumentation is weakening the test (Roodman 2009a). As with the PK replication, this step does not change the pattern of signs much; but nor does it leave us with great confidence in the instrumentation strategy. And the point estimate for the significant coefficient in the last regression, 0.317 on past women's loans, is ten times larger than that from OLS. Plugging this number into Khandker's simulation might lead to the estimate that microcredit accounted for more than 100% of the poverty reduction in Bangladesh in the 1990s.

[29] Our coefficients of 0.026 and 0.034 on current and past women's borrowing are much larger than Khandker's 0.009 and 0.010. Perhaps Khandker is censoring the log of credit at log 1 rather than log 1,000.

Table 9. Summary statistics of consumption and credit variables, Khandker and new data set

	Reported in Khandker				New data set			
		Non-participants				Non-participants		
	Partici-pants	Target	Non-target	All	Partici-pants	Target	Non-target	All
1991/92, round 1								
Cumulative borrowing by men (taka)	3,472			797	2,787			720
	(6,829)			(3,678,4)	(6,659)			(3,598)
Cumulative borrowing by women (taka)	5,853			1,583	5,442			1,407
	(8,038)			(4,974)	(7,654)			(4,562)
Total expenditure/capita (taka per year)	3,910	3,791	5,635	4,452	3,993	3,819	5,693	4,549
	(1,586)	(1,678)	(3,666)	(2,555)	(1,644)	(1,724)	(3,615)	(2,708)
Food expenditure/capita (taka per year)	3,051	2,966	3,705	3,237	3,075	2,990	3,662	3,258
	(795)	(879)	(1,123)	(987)	(795)	(895)	(1,079)	(992)
Non-food expenditure/capita (taka per year)	859	825	1,931	1,215	919	829	2,031	1,291
	(1,102)	(1,061)	(3,086)	(1,958)	(1,187)	(1,091)	(3,031)	(2,115)
Obervations	824	535	279	1,638	824	541	273	1,638
1999								
Cumulative borrowing by men (taka)	2,483			1,088	2,150			1,149
	(9,013)			(5,791)	(7,958)			(5,898)
Cumulative borrowing by women (taka)	11,348			5,581	11,795			6,266
	(17,592)			(13,392)	(18,141)			(14,472)
Total expenditure/capita (taka per year)	5,264	4,504	7,214	5,810	4,977	4,465	7,059	5,431
	(3,580)	(2,664)	(5,789)	(4,503)	(3,263)	(2,704)	(5,526)	(4,034)
Food expenditure/capita (taka per year)	3,550	3,305	4,374	3,753	3,284	3,175	3,971	3,446
	(1,335)	(1,506)	(2,189)	(1,688)	(1,214)	(1,377)	(2,018)	(1,533)
Non-food expenditure/capita (taka per year)	1,714	1,198	2,840	2,057	1,693	1,290	3,088	1,985
	(2,848)	(1,579)	(4,571)	(3,575)	(2,593)	(1,687)	(4,591)	(3,199)
Obervations	1,104	292	242	1,638	1,123	288	227	1,638

Standard deviations in parentheses.

Table 10. Estimates of the impact of microcredit on household (HH) consumption, following Khandker

	OLS		2SLS		2SLS	
	No FE	FE	No FE	FE	No FE	FE
Log women's current loans	0.026	0.017	0.046	0.007	–0.027	–0.001
	(2.228)**	(1.580)	(1.818)*	(0.218)	(0.378)	(0.019)
Log women's past loans	0.034	0.020	0.107	0.135	0.348	0.317
	(2.324)**	(1.257)	(3.372)***	(2.700)***	(2.310)**	(3.522)***
Log men's current loans	0.036	0.001	0.145	0.078	0.057	0.105
	(1.824)*	(0.029)	(2.358)**	(1.280)	(0.397)	(0.804)
Log men's past loans	0.004	–0.024	0.026	–0.052	0.008	–0.092
	(0.177)	(1.057)	(0.520)	(0.703)	(0.057)	(0.934)
Interaction terms using village dummies						
(F test p value)					0.000	0.000
Observations	1,638	1,638	1,638	1,638	1,638	1,638
Hansen J test (p value)			0.000	0.112	0.288	1.000

All regressions run in differences except that fixed-effect (FE) regressions include village dummy controls undifferenced. All 2SLS regressions instrument with lagged and current interactions of the credit choice dummy with village dummies and (unreported) controls. Final pair includes interaction terms involving village dummies as controls. "Current loans" is cumulative borrowing over the last 6–7-year period; "past loans" is that for the previous period and is set to 0 for 1991–92. Controls are: sex, age, and education level of household head; whether parents, brothers, and sisters of household head or spouse own land (for 1991–92) or own at least 0.5 acres (1999); availability of co-education; and, for non-FE regressions, prices of rice, wheat flour, mustard oil, hen's eggs, milk, and potatoes, as well as male and female wage levels. Absolute t statistics robust to heteroskadasticity in parenthesis. *significant at 10%. **significant at 5%. ***significant at 1%.

Overall, these findings reduce our confidence that Khandker's results reflect causality from credit to household consumption. Since we doubt the OLS foundation of the Khandker paper, we also doubt that which is built upon it, in particular the claim that microcredit has disproportionately helped extremely poor people. Fundamentally, the move to the panel framework does not seem to compensate for the lack of clearly exogenous variation in the use of microcredit.

Conclusion

Pitt and Khandker (1998) and Khandker (2005) prominently reinforced three broad ideas about microcredit: that it is effective in reducing poverty generally, that this is especially so when women do the borrowing, and that the extremely poor benefit most. Morduch (1998) disseminated the idea that microcredit helps families smooth their expenditures, lessening the pinch of hunger and need in lean times. In our view, nothing in the present paper contradicts those ideas. We assert, however, that decisive statistical evidence in *favor* of them is absent from these studies and extraordinarily scarce in the literature as a whole. The principle difficulties for studying the effects of microfinance have been a lack of clean quasi-experiments and an absence until recently of randomized trials.

Our short list of exceptions includes Coleman (1999, 2006), Fernald *et al*. (2008), Banerjee et al (2009), and Karlan and Zinman (2009 and forthcoming). Coleman performs an experiment in the form of random and unannounced delays in implementing a credit program in some villages in Northeast Thailand. He finds measurable benefits for relatively affluent and well-connected villagers. Fernald *et al*., as well as Karlan and Zinman (forthcoming), study a cash loan business in South Africa, not unlike a payday lender in the United States, which agreed to randomly relax its computerized risk assessment rules for marginal candidates. Fernald *et al*. find that loans increase psychological stress among women, but not men. But Karlan and Zinman find benefits across genders and a variety of outcomes, including for household consumption. Notably, the South African loans are perhaps not "microcredit" as usually conceived: they are high-cost consumer finance and the key mechanism may have been that the loans let people obtain jobs that required them to pay for training up-front, whereas poor people targeted by microcredit typically have little hope of such employment (Banerjee and Duflo 2007). Karlan and Zinman (2009) take a similar method to the Philippines, with a focus there on traditional microcredit for small business investment. Profits rise, but largely for men and particularly for men with higher incomes. Moreover, the increases in profits appear to arise from business contractions that yielded smaller, lower-cost (and more profitable) enterprises. Banerjee e*t al.*, (2009) run a traditional randomized trial of microcredit in urban India. After a year, they report a mix of economic results but no strong average impacts; measured impacts on health, education, and women's empowerment were negligible. As we write, Pitt and Khandker (1998) and Khandker (2005) thus remain the only high-profile economic papers asserting large, sustained impacts of microcredit.

At least three more randomized controlled trials of microfinance are underway or in prospect in Mexico, Morocco, and Peru. The sudden swell of randomized trials 30 years after the birth of microcredit of course reflects a broader trend in the social sciences. As such, it also leads to a broader question, about the value of non-randomized studies. Our prior is that exclusive reliance on one type of study is not optimal. But the present analysis suggests that for non-randomized studies to contribute to the study of causation in social systems where endogeneity is pervasive, the quality of the natural experiments must be very high. And it must be demonstrated. We also believe that longitudinal surveys like the ones in Bangladesh are worthwhile even when they fail to enlighten us about the impacts of outside interventions. In the Lowess plots in this paper, for instance, one can glimpse a trove of information about how poor households manage money and use financial services. Because of the eagerness to study important questions of impact, this trove remains substantially unexplored.

If our conclusions stand the test of time, they will also raise a question about how researchers and practitioners can more easily determine the robustness of important findings. One partial solution is for more journals to encourage replication studies like this one, for example by requiring authors to share data and code (Hamermesh 2007). Another step is to develop norms for graphically demonstrating identifying assumptions in non-experimental studies of causal mechanisms. More can be done to improve how research reaches policymakers.

References

Alderman, Harold, Jere R. Behrman, Hans-Peter Kohler, John A. Maluccio, and Susan Cotts Watkins. 2001. Attrition in Longitudinal Household Survey Data: Some Tests for Three Developing-Country Samples. *Demographic Research* 5(4): 79–124.

Amemiya, Takeshi. 1974. Multivariate Regression and Simultaneous Equation Models When the Dependent Variables Are Truncated Normal. *Econometrica* 42(6): 999–1012

Armendáriz de Aghion, Beatriz, and Jonathan Morduch. 2005. *The Economics of Microfinance*. Cambridge, MA: The MIT Press.

Banerjee, Abhijit V., and Esther Duflo. 2007. The Economic Lives of the Poor. *Journal of Economic Perspectives* 21(1): 141–67.

Banerjee, Abhijit V., and Esther Duflo. 2008. The Experimental Approach to Development Economics. Working Paper. Cambridge, MA: MIT Department of Economics and Abdul Latif Jameel Poverty Action Lab.

Banerjee, Abhijit V., Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2009. "The Miracle of Microfinance? Evidence from a Randomized Evaluation. Working Paper. Cambridge, MA: MIT Department of Economics and Abdul Latif Jameel Poverty Action Lab.

Brock, William A., and Steven N. Durlauf. 2001. Growth Empirics and Reality. *World Bank Economic Review* 15(2): 229–71.

Chemin, Matthieu. 2008. The Benefits and Costs of Microfinance: Evidence from Bangladesh. *Journal of Development Studies* 44(4): 463–84.

Coleman, Brett E. 1999. The Impact of Group Lending in Northeast Thailand. *Journal of Development Economics* 60: 105–41.

Coleman, Brett E. Microfinance in Northeast Thailand: Who Benefits and How Much? *World Development* 34(9): 1612–38.

Deaton, Angus. 2009. Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. Working Paper 14690. Cambridge, MA: National Bureau of Economic Research.

Dunford, Christopher. 2006. *Evidence of Microfinance's Contribution to Achieving the Millennium Development Goals*. Davis, CA: Freedom from Hunger.

Fernald, Lia C.H., Rita Hamad, Dean Karlan, Emily J. Ozer, and Jonathan Zinman. 2008. Small Individual Loans and Mental Health: A Randomized Controlled Trial among South African Adults. *BMC Public Health* 8: 409.

Goldberg, Nathanael. 2005. *Measuring the Impact of Microfinance: Taking Stock of What We Know.* Washington, DC: Grameen Foundation USA.

Greene, Willam H. 1998. Gender Economics Courses in Liberal Arts Colleges: Further Results. *Research in Economic Education* 29(4): 291–300.

Hamermesh, Daniel S. 2007. Replication in Economics. *Canadian Journal of Economics* 40(3): 715–33.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. Identification and Estimation of Treatment Effects with a Regression Discontinuity Design. *Econometrica* 69(1): 201–09.

Heckman, James J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement* 5: 475–492.

Heckman, James J. 2000. Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective. *Quarterly Journal of Economics* 115(1): 45–97.

Hossain, Mahabub. 1988. *Credit for Alleviation of Rural Poverty: The Grameen Bank in Bangladesh*. Washington, DC: International Food Policy Research Institute.

Imbens, Guido W., and Thomas Lemieux. 2008 . Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics* 142(2): 615–35.

Ito, Sanae. 1999. *The Grameen Bank: Rhetoric and Reality*. Doctoral dissertation. Brighton: University of Sussex.

Kaboski, Joseph P., and Robert M. Townsend. 2005. Policies and Impact: An Analysis of Village-Level Microfinance Institutions. *Journal of the European Economic Association* 3(1): 1–50.

Karlan, Dean. 2001. Microfinance Impact Assessments: The Perils of Using New Members as a Control Group. *Journal of Microfinance* 3(2): 75–85.

Karlan, D., Zinman, J. (2009). Expanding Microenterprise Credit Access: Using Randomized Supply Decisions to Estimate the Impacts in Manila. Working Paper. Yale University, Dartmouth College, and Innovations for Poverty Action.

Karlan, Dean, and Jonathan Zinman. Forthcoming. Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts. *Review of Financial Studies*.

Kelejian, Harry H. 1971. Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables. *Journal of the American Statistical Association* 66(334): 373–74.

Khandker, Shahidur R. 1996. Role of Targeted Credit in Rural Non-farm Growth. *Bangladesh Development Studies* 24(3 & 4).

Khandker, Shahidur R. 1998. *Fighting Poverty with Microcredit: Experience in Bangladesh* (New York: Oxford University Press).

Khandker, Shahidur R. 2000. Savings, Informal Borrowing and Microfinance. *Bangladesh Development Studies* 26(2 & 3).

Khandker, Shahidur R. 2005. Microfinance and Poverty: Evidence Using Panel Data from Bangladesh. *World Bank Economic Review* 19(2): 263–86.

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.

McKernan, Signe-Mary (2002). The Impact of Microcredit Programs on Self-employment Profits: Do Non-credit Program Aspects Matter? *The Review of Economic and Statistics* 84(1): 93–115.

Menon, Nidhiya. 2005. Non-linearities in Returns to Participation in Grameen Bank Programs. *Journal of Development Studies* 42(8): 1379–1400.

Morduch, Jonathan. 1994. Poverty and Vulnerability. *American Economic Review* 84(2): 221–25.

Morduch, Jonathan. 1995. Income Smoothing and Consumption Smoothing. *Journal of Economic Perspectives* 9(3): 103–14.

Morduch, Jonathan. 1998. Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh. New York University. Department of Economics. Available at nyu.edu/projects/morduch/documents/microfinance/Does_Microfinance_Really_Help.pdf.

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

Pitt, Mark M. 1999. Reply to Jonathan Morduch's "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh." Department of Economics. Brown University. Available at www.pstc.brown.edu/~mp/reply.pdf.

Pitt, Mark M. 2000. The Effect of Nonagricultural Self-Employment Credit on Contractual Relations and Employment in Agriculture: The Case of Microcredit Programs in Bangladesh. *Bangladesh Development Studies* 26(2 & 3): 15–48.

Pitt, Mark M., and Shahidur R. Khandker. 1998. The Impact of Group-Based Credit on Poor Households in Bangladesh: Does the Gender of Participants Matter? *Journal of Political Economy* 106(5): 958–96.

Pitt, Mark M., and Shahidur R. Khandker. 2002. Credit Programs for the Poor and Seasonality in Rural Bangladesh. *Journal of Development Studies* 39(2): 1–24.

Pitt, Mark M., Shahidur R. Khandker, and Jennifer Cartwright. 2006. Empowering Women with Micro Finance: Evidence from Bangladesh. *Economic Development and Cultural Change* 54(4): 791–831.

Pitt, Mark M., Shahidur R. Khandker, Omar Haider Chowdhury, and Daniel L. Millimet 2003. Credit Programs for the Poor and the Health Status of Children in Rural Bangladesh. *International Economic Review* 44(1): 87–118.

Pitt, Mark M., Shahidur R. Khandker, Signe-Mary McKernan, and M. Abdul Latif. 1999. Credit Programs for the Poor and Reproductive Behavior in Low Income Countries: Are the Reported Causal Relationships the Result of Heterogeneity Bias? *Demography* 36(1): 1–21.

Rivers, Douglas, and Quang H. Vuong. 1988. Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models. *Journal of Econometrics* 39: 347–66.

Roodman, David. 2009a. A Note on the Theme of Too Many Instruments. *Oxford Bulletin of Economics and Statistics* 71 (1): 135–158.

Roodman, David. 2009b. Estimating fully observed recursive mixed-process models with cmp. Working Paper 168. Washington, DC: Center for Global Development.

Smith, Richard J., and Richard W. Blundell. 1986. An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply. *Econometrica* 54(3): 679–85.

Thomas, Duncan, Elizabeth Frankenberg, and James P. Smith. 2001. Lost but not Forgotten: Attrition and Follow-up in the Indonesia Family Life Survey. *Journal of Human Resources* 36(3): 556–92.

Todd, Helen. 1996. *Women at the Center: Grameen Bank Borrowers after One Decade*. Dhaka: University Press Limited.

Wilde, Joachim. 2000. Identification of Multiple Equation Probit Models with Endogenous Dummy Regressors. *Economics Letters* 69: 309–12.

Appendix. Testing the estimation software on a simulated dataset

Pitt's (1999) reply to Morduch (1999) includes Stata command files that simulate data sets illustrating various aspects of the estimation problem as well as the consistency of the 2SLS analog of the PK estimator. Because we cannot explain the contradiction between PK's headline results and our replication, we report here on a set of simulations performed with code adapted from Pitt. We borrow from his command file "sim7.do," which is the most elaborate simulation that embodies most of the key features of the PK model.[31]

The simulated data sets can be described as follows. The outcome, female borrowing, and male borrowing equations contain correlated village-level fixed effects, according to:

$$(\mu_o, \mu_f, \mu_m)' \sim \text{i. i. d.}\, \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \sqrt{0.1} & -\sqrt{0.1} \\ \sqrt{0.1} & 1 & -0.5 \\ -\sqrt{0.1} & -0.5 & 1 \end{bmatrix}\right)$$

At the household level, idiosyncratic errors are structured similarly and combine with the village effects for overall error terms:

$$(v_o, v_f, v_m)' \sim \text{i. i. d.}\, \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \sqrt{0.5} & \sqrt{0.5} \\ \sqrt{0.5} & 1 & 0.5 \\ \sqrt{0.5} & 0.5 & 1 \end{bmatrix}\right),$$

$$\epsilon_o = \mu_o + v_o$$

$$\epsilon_f = \mu_f + v_f$$

$$\epsilon_m = \mu_m + v_m$$

Exogenous regressors are generated using the uniform distribution on the unit interval, $U[0,1]$:

$$w_1, w_2, w_3 \sim \text{i. i. d.}\, U[0,1]$$

$$x_1 = w_1 - 0.5$$

$$x_2 = w_2 - 0.5$$

$$land = 0.7w_3$$

[31] Pitt's later simulations illustrate consistency of the LIML estimator in the face of various deviations from the basic assumptions, such as a fuzzy rather than sharp discontinuity at the half-acre line.

Female credit programs are more common than male ones:

$$p_f = 1\{\mu_f < 1\}$$

$$p_m = 1\{\mu_m < 0\}$$

And households owning less than half an acre are eligible:

$$e = 1\{land < 0.5\}$$

Using a credit censoring level of $C = 0$, and following the nomenclature in (1), the system of equations is:

$$c_f = p_f e$$

$$c_m = p_m e$$

$$y_f^* = 2x_1 + 3x_2 + land - 1 + \epsilon_f \text{ if } c_f = 1$$

$$y_m^* = x_1 + 2x_2 + 2 \times land - 1 + \epsilon_m \text{ if } c_m = 1$$

$$y_f = 1\{y_f^* \geq C\} \cdot y_f^*$$

$$y_m = 1\{y_m^* \geq C\} \cdot y_m^*$$

$$y_o = 2x_1 + 2x_2 + land + 1.5y_f + 0.5y_m + \epsilon_o$$

where the coefficients on y_f and y_m are of primary interest.

Table 11 characterizes the distributions of the coefficients estimates using three different estimators: WESML-LIML-FE; the analogous 2SLS estimator with a large set of interaction terms, as in (2); and exactly identified 2SLS, which instruments only with c_f and c_m. The main point is in the first row: our implementation of PK's LIML estimator clearly works correctly in this case. The analogous 2SLS estimates (second row) are also reasonable, but less efficient and, as the last row suggests, somewhat upward-biased by overfitting in the first stage onto the large instrument set.

Table 11. Estimated coefficients on y_f and y_m, 100 draws

	y_f (true value = 1.5)		y_m (true value = 0.5)	
Estimator	Mean	St. dev.	Mean	St. dev.
WESML-LIML-FE	1.500	0.040	0.497	0.055
2SLS	1.587	0.122	0.594	0.172
2SLS, instrumenting with c_f,c_m only	1.527	0.166	0.479	0.239