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

David Roodman and Jonathan Morduch

Abstract

With corrections from Pitt (2011), we improve our replication of Pitt and Khandker (PK, 1998), the most influential study of the long-term impacts of microcredit. Better replication strengthens our doubts about PK's causal claims. Problems include: an arbitrary and influential imputation for the logarithm of the treatment when zero; the absence of a discontinuity asserted as a basis for identification; likely endogeneity in the quasi-experiment as a result; econometric evidence of such endogeneity; and instrument weakness arising from PK's disaggregation of borrowings by gender. The PK results are robust to fixes for most of these problems. But the instrument weakness is apparently making the estimator sensitive to outliers as well as bimodal: there is a negative-impact mode in addition to the positive one reported in PK. Dropping the 16 highest observations of the outcome of greatest interest, household consumption, eliminates the bimodality and any finding of positive impact. Our experience demonstrates the value of replication, and of sharing of data and code to facilitate it. It shows that sophisticated estimators can obscure specification problems, thus why simpler, typically linear, techniques should be used as checks. And it highlights difficulties of non-experimental evaluations.

JEL Codes: E52, O16, O53
Keywords: microcredit, bangladesh, poverty reduction

Center for Global Development
www.cgdev.org

Working Paper 174
May 2009 (revised 12-14-11)

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

David Roodman
Center for Global Development

Jonathan Morduch
New York University
Financial Access Initiative

June 2009
(revised 12-14-11)

We thank Mark Pitt and the Research Committee of the World Bank for assistance with data, and Xavier Giné and Dean Karlan for reviews.
Correspondence: David Roodman, droodman@cgdev.org.

CGD is grateful for contributions from the Australian Agency for International Development in support of this work.

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/1422654

Center for Global Development
1800 Massachusetts Ave., NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

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

Revised version

David Roodman
Center for Global Development

Jonathan Morduch
New York University
Financial Access Initiative

December 2011

Abstract: With corrections from Pitt (2011), we improve our replication of Pitt and Khandker (PK, 1998), the most influential study of the long-term impacts of microcredit. Better replication strengthens our doubts about PK's causal claims. Problems include: an arbitrary and influential imputation for the logarithm of the treatment when zero; the absence of a discontinuity asserted as a basis for identification; likely endogeneity in the quasi-experiment as a result; econometric evidence of such endogeneity; and instrument weakness arising from PK's disaggregation of borrowings by gender. The PK results are robust to fixes for most of these problems. But the instrument weakness is apparently making the estimator sensitive to outliers as well as bimodal: there is a negative-impact mode in addition to the positive one reported in PK. Dropping the 16 highest observations of the outcome of greatest interest, household consumption, eliminates the bimodality and any finding of positive impact. Our experience demonstrates the value of replication, and of sharing of data and code to facilitate it. It shows that sophisticated estimators can obscure specification problems, thus why simpler, typically linear, techniques should be used as checks. And it highlights difficulties of non-experimental evaluations.

[1] Correspondence: David Roodman, droodman@cgdev.org.

Over the last few decades, microcredit has captured millions of customers, billions of dollars in financing, a Nobel Peace Prize, and the imagination of the global public. Many have seen microcredit as lifting families out of poverty, especially when lent to women. The movement owes its strength in part to an early literature based on observational data that shows strong positive impacts. The leading studies in this literature took place in the leading nation of microcredit, Bangladesh. More recently, muted results from randomized trials in India, the Philippines, and Morocco are prompting second thoughts.[2] The contradiction between the old and new studies naturally leads to a question: what explains the differences? Has the impact of microcredit varied over time and place? Is the key that the Bangladesh studies were longer-term? Or is it the difference in methods?

To probe this question, we replicate the most-cited evaluation of the long-term impacts of microcredit, Pitt and Khandker (1998). Our in-depth inquiry is warranted by the great weight historically placed on that study, by its relatively long time frame, and by the continuing public controversy over the efficacy of microfinance. Muhammad Yunus, who founded the Grameen Bank and shared a Nobel with it, once regularly invoked PK as showing that "In a typical year 5 percent of Grameen borrowers…rise above the poverty level."[3] In contrast with the randomized studies, which have measured impacts over at most two years, PK assesses the impact of up to five years of microcredit use. As a result, it remains a premier analysis of long-term microcredit impacts—and still a leading basis for the claim that microcredit strongly reduces poverty.

The path to replicating PK has been long and bumpy. We use the Stata estimation command

---

[2] Randomized studies have found that access to capital increases average profitability of male-run microenterprises, but challenged the central claim that it does so for female-run businesses (see McKenzie and Woodruff 2008 on male-run businesses in Mexico and de Mel, McKenzie, and Woodruff 2008 on male- and female-run businesses in Sri Lanka). Other randomized studies find no support for the claim that microcredit increases household consumption in the short-run (Banerjee et al. 2009; Karlan and Zinman 2011; Crépon et al. 2011).

[3] E.g., Appelbaum (2008), Yunus (1999, 2007, 2008), Yunus and Abed (2004). The 5% figure comes from Khandker (1998, p. 56), which extrapolates from PK.

cmp (Roodman 2011) for estimating conditional mixed-processed models like PK's (such as ones with a first-stage tobit and second-stage linear function).[4] Our failure in the first edition of this paper to match key coefficients, and our public sharing of data and code, led and allowed Mark Pitt (2011) to unearth key discrepancies in our specifications. Thanks to his corrections, we now replicate well.

PK remains striking for its attack on selection bias through an innovative and complex limited-information maximum likelihood (LIML) framework. Our replication and robustness testing leads to a deeper appreciation of the strengths and weaknesses of these methods. The PK results turn out to be robust to fixes for some specification problems previously noted (Morduch 1998), such as the apparent lack of an asserted discontinuity in treatment. But, ultimately, PK's decision to disaggregate borrowings by gender stretches their instruments to the breaking point. Instrument weakness appears to make the maximized likelihood and the estimator bimodal. Key results evaporate when a small number of outliers are removed.

Although some of the econometric tools we bring to bear were not developed or were less practical in the late 1990s than now, certain specification checks were practical then, and would have led the PK study to be interpreted differently. True, the dangers of instrument weakness were not as well understood. And bootstrapping, which led us to the bimodality of the estimator, was less feasible for complex maximum likelihood estimators on the computers of the day. But other checks could have been performed: for the presence of asserted discontinuities; for the normality of the errors; for robustness to outlier removal; for robustness to aggregation by gender; and for robustness to switching to linear LIML.

[4] The first edition of this paper also replicated two other prominent studies of the same data (Morduch 1998; Khandker 2005). But because of PK's stature, as well as the complexity of our new analysis, PK is the sole focus in this edition. Our analysis of the other two studies has not changed.

Our analysis has several broad implications. It demonstrates inherent difficulties in judging the credibility of results from complex estimators while highlighting the value of specific diagnostics for such estimators, such as bootstrapping and the use of simple, typically linear, alternatives more robust to violations of distributional assumptions. It also fits into the debate over the merits of experimental and non-experimental evaluations.[5] Our intuition is that each approach has strengths and weaknesses. Non-randomized studies, for example, can exploit natural experiments, as PK aimed to do by organizing around a discontinuity created by an eligibility rule for microcredit access. But the analysis here puts up warning flags about the difficulties of compensating for a lack of experimental variation.

The paper runs as follows. Section 1 reintroduces the PK estimator. Section 2 presents our replication of the regression of central interest, with household consumption as the outcome. It uses Pitt's (2011) data set. Section 3 demonstrates five concerns about the estimator and tests fixes where possible. Section 4 attempts to replicate PK regressions for other outcomes using outcome variables we construct from the underlying survey data. Section 5 concludes.[6]

1. The econometrics of PK

1.1 The estimation problem

PK analyze data from surveys of 1,798 households in 87 randomly selected villages within a randomly selected 29 of Bangladesh's 391 *upazillas*. The surveyors visited the households in 1991–92 after each of the three main rice seasons: *aman* (December–January), *boro* (April–May), and *aus* (July–August). Only 29 households attrited by the third round. Ten of the 87 villages had male borrowing groups, 22

[5] For more, see, for example, the debate between Banerjee and Duflo (2009) and Deaton (2010).
[6] Separately, we respond more directly to Pitt's criticisms and explain how our initial attempt at replication came to differ from the original (Roodman and Morduch 2011). There has been a long debate over the paper (PK, Morduch 1998, Pitt 1999, Roodman and Morduch 2009, Pitt 2011), and we believe the present analysis resolves it to a great extent. To the understandably doubtful reader, we point out that this is the first analysis by someone other than Pitt or Khandker of the actual PK headline regression, or a close facsimile thereof. In that sense, it is a turning point.

had female groups, and 40 had both. All groups were single-sex. Credit programs of three institutions were evaluated: the Grameen Bank; a large non-governmental group called BRAC; and the official Bangladesh Rural Development Board (BRDB). According to PK (p. 959), all three credit programs essentially set eligibility in terms of land ownership: only functionally landless households, defined as those owning half an acre or less, could borrow.[7] For statistical precision, the surveyors oversampled households poor enough to be targeted for microcredit. Since sampling on the basis of eligibility can bias results, PK incorporate sampling weights constructed from village censuses.

PK study impacts on six outcomes. Two—per-capita consumption and female-owned non-land assets—are household-level variables. The rest—male and female labor supply and school enrollment of girls and boys—are individual-level. For each outcome, the three-way split by credit supplier and the two-way split by sex lead to six parameters of interest, the impact coefficients on credit by lender and gender. A central feature of the estimation problem is that the credit variables are at once presumed endogenous and bounded from below. Meanwhile, all of the outcomes except log household consumption are themselves bounded or binary. PK therefore estimate the impact parameters using a limited-information maximum likelihood (LIML) framework that models the limited nature of all the endogenous variables. Each estimation model contains equations for the outcome variable of interest as well as for female borrowing and male borrowing. The outcome is variously modeled as linear and unbounded, Tobit, or probit.

In order to study longer-term impacts, PK take advantage of microcredit borrowing histories collected back to late 1986. They measure credit as the simple sum of borrowings since that time. If a

[7] Among the three creditors, at least Grameen also officially applied an alternative eligibility criterion: ownership of assets worth less than one acre of medium-quality land (The Grameen Bank Ordinance, as amended through 2008, §2(h).) However, PK rely exclusively on the half-acre rule in their analysis, using it, for example, to code the eligibility status of households in villages without credit programs.

woman borrowed 1,000 Bangladeshi taka, repaid it over the course of a year, then repeated with cycles of 2,000, 3,000, 4,000, and 5,000, that would count as 15,000 in borrowings.

1.2 The estimator

To state the PK model precisely, we first need to formally describe access to credit. Let p_f and p_m be dummies indicating whether credit groups composed of females or males are operating in the village of a given household or household member. Let e be a dummy for whether the household is deemed eligible for a microcredit program, regardless of whether any borrowing groups operate in the village. Then the credit choice variables, indicating whether members of each sex can borrow, are

$$c_f = p_f e$$

$$c_m = p_m e$$

A central contention in PK is that c_f and c_m are exogenous and excludable. This allows the availability of microcredit to instrument for uptake. As we will discuss, this contention is based in part on the idea that eligibility for microcredit, e, is set by a discontinuous program rule: only households owning half an acre or less of land can borrow.

As mentioned, PK study impacts on several outcomes. Since we focus on log per-capita household consumption, the outcome relating to PK's influential finding that microcredit reduces poverty, we take the outcome in the model, y_o, to be continuous and unbounded. For the same reason, we take the unit of observation as the household-season combination rather than the individual-season.

Let y_f be the logarithm of total microcredit borrowings of all females in a household and y_m be the corresponding variable for males. Let $\mathbf{y}_{fm} = (y_{f1}, y_{f2}, y_{f3}, y_{m1}, y_{m2}, y_{m3})'$ be the six credit variables disaggregated by lender as well as gender. And let $\mathbf{x}$ be a vector of controls that includes log landholdings, the eligibility dummy e, household characteristics, village and survey round dummies, and a constant.[8] Let C_t be the credit censoring *threshold*, the minimum observable log borrowing amount among borrowers. If there is no borrowing, the household gets C_v, the censoring *value* for log borrowing assigned by the researcher (necessary since log 0 is undefined). Then the PK estimation model, fit with maximum likelihood (ML), can be written as:

$$
\begin{aligned}
y_o &= \mathbf{y}_{fm}'\boldsymbol{\delta} + \mathbf{x}'\boldsymbol{\beta}_o + \epsilon_o \\
y_f^* &= \mathbf{x}'\boldsymbol{\beta}_f + \epsilon_f \text{ if } c_f = 1 \\
y_m^* &= \mathbf{x}'\boldsymbol{\beta}_m + \epsilon_m \text{ if } c_m = 1 \\
y_f &= \begin{cases} y_f^* & \text{if } c_f = 1 \text{ and } y_f^* \geq C_t \\ C_v & \text{otherwise} \end{cases} \\
y_m &= \begin{cases} y_m^* & \text{if } c_m = 1 \text{ and } y_m^* \geq C_t \\ C_v & \text{otherwise} \end{cases} \\
\boldsymbol{\varepsilon} &\equiv \left(\epsilon_o, \epsilon_f, \epsilon_m\right)' \\
\boldsymbol{\varepsilon}|\mathbf{x} &\sim \text{i.i.d.}\,\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\Sigma}$ is a 3×3 positive-definite symmetric matrix. After estimation, PK "cluster" standard errors by household to allow for error correlations across the three survey rounds.

The PK model is unusual in several respects. The three main equations include the same exogenous regressors, $\mathbf{x}$: seemingly, no instruments are excluded. The exogeneity of c_f and c_m is the asserted basis for identification, yet those dummies do not seem to serve as instruments. The credit equations' samples are restricted, which means that the number of equations in the model varies by observation. The outcome equation contains six endogenous credit variables, $\mathbf{y}_{fm}$, but the model includes just two instrumenting equations. The instrumenting stage is modeled as censored, which forces the unusual distinction between the censoring threshold, which is relevant for the Tobit

[8] PK include specifications that control for a set of village characteristics instead of a full set of village dummies. But the fixed-effect specifications are preferred, so we focus exclusively on them.

modeling in the credit equations, and the censoring value, which is relevant for the treatment of credit on the right side of the outcome equation.[9] And while PK set out to exploit a discontinuity in access to credit, theirs is not a typical regression discontinuity design: the sample is not concentrated around the half-acre mark, but spans households from a *de minimus* 0.1 acres up to 5 acres of landholdings.[10] This wide bandwidth necessitates a parametric approach.

1.3 A closer look at assumptions

A key to understanding some of these unusual characteristics is to note that the last line of (1) elides a complexity. The y_f and y_m equations are not defined over the full sample, so ϵ_f, ϵ_m, and the joint distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ are not either. Thus the last line is not strictly meaningful. To state the distributional assumption more precisely, we distinguish the four possible cases of credit availability by gender. We use combinations of o, f, and m subscripts to denote subvectors of $\boldsymbol{\varepsilon}$ and submatrices of $\mathbf{\Sigma}$ corresponding to combinations of the equations for the outcome, female credit, and male credit. A precise statement of the distributional assumption is then:

$$\varepsilon_o|\mathbf{x} \sim \mathcal{N}(0, \Sigma_o) \text{ when } c_f = 0, c_m = 0$$

$$\boldsymbol{\varepsilon}_{of}|\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{of}\right) \text{ when } c_f = 1, c_m = 0$$

$$\boldsymbol{\varepsilon}_{om}|\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{om}) \text{ when } c_f = 0, c_m = 1$$

$$\boldsymbol{\varepsilon}_{\boldsymbol{ofm}}|\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{ofm}}\right) \text{ when } c_f = 1, c_m = 1$$

where $\boldsymbol{\varepsilon}_{\boldsymbol{ofm}} \equiv \boldsymbol{\varepsilon}$ and $\mathbf{\Sigma}_{\boldsymbol{ofm}} \equiv \mathbf{\Sigma}$. Every case implies $\varepsilon_o|\mathbf{x} \sim \mathcal{N}(0, \Sigma_o)$. Thus

$$\varepsilon_o|\mathbf{x}, c_f, c_m \sim \mathcal{N}(0, \Sigma_o) \qquad (2)$$

That is, the error in the outcome equation is independent not only of the controls but credit availability too. This is how the exogeneity of credit choice enters the identification strategy.

[9] The distinction between censoring threshold and censoring value is a major point in Pitt (2011). For more discussion, see Roodman and Morduch (2011).

[10] PK exclude 41 households owning more than 5 acres.

If ε_o is independent of $\mathbf{x}$, c_f, and c_m, it is uncorrelated with any function of them. In particular, ε_o is uncorrelated with $\mathbf{x}$, c_f, c_m, $c_f\mathbf{x}$, and $c_m\mathbf{x}$. As a result, one can gain intuition about the PK model by innocuously inserting c_f and c_m into the latent credit equations in (1):

$$\begin{aligned} y_f^* &= c_f\mathbf{x}'\boldsymbol{\beta}_f + \epsilon_f \text{ if } c_f = 1 \\ y_m^* &= c_m\mathbf{x}'\boldsymbol{\beta}_m + \epsilon_m \text{ if } c_m = 1 \end{aligned} \tag{3}$$

This version communicates the idea that $c_f\mathbf{x}$ and $c_m\mathbf{x}$ are the instruments, being excluded from the y_o equation. And since $\mathbf{x}$ includes a constant, c_f and c_m are seen as instruments too.

One other peculiarity of the PK specification can perhaps be given intuition, namely, how two first-stage equations instrument 6 endogenous credit variables. This can be thought of as equivalent to instrumenting all six distinctly while imposing constraints that equate first-stage coefficients across the three lending programs. Few households appear in the data to have borrowed from more than one lender.

The PK fixed-effects specifications—which include the headline regression showing that microcredit reduces poverty—are akin to the difference-in-differences (DID) estimator with controls (Morduch 1998). The two dimensions of difference are the eligibility of a household for microcredit and the availability of microcredit in a village. All PK regressions control for the first dimension via the household eligibility dummy e in $\mathbf{x}$; and the village dummies in $\mathbf{x}$ include the program dummies p_f and p_m in their span. As a result, all the factors of $c_f(= p_f e)$ and $c_m(= p_m e)$ are controlled for. Rather as in DID, identification comes primarily from exogenous variation in the excluded products $p_f e$ and $p_m e$ conditional on the included factors p_f, p_m, and e.[11]

[11] The comparison to standard DID is not exact because $\mathbf{x}$ includes additional controls, and $c_f\mathbf{x}$ and $c_m\mathbf{x}$ contain additional instruments.

As with standard DID, in which the dimensions of difference are treatment group membership and time, the assumptions in the PK estimator are not trivial. For DID with controls to yield consistent estimates, being in the treatment group must, aside from any treatment effect, have the same impact on the outcome in the before and after periods (after conditioning linearly on controls). Otherwise, apparent impacts of being in the treatment or control group cannot be entirely attributed to the treatment. Equivalently, the passage of time must, aside from the treatment effect, have the same impact in both treatment and control groups. Analogously, PK's identification strategy requires us to believe that being in a village where microcredit is available to a given sex has the same relationship with mean household consumption for eligible and ineligible households; equivalently, that, aside from the impacts of borrowing, being eligible for microcredit has the same implications for mean household consumption in program and non-program villages.[12]

The validity of these assumptions is of course open to question. For example, in villages where eligible households are relatively well-off, credit group formation may be more likely. Turning that around, conditioning on being in a village with credit groups, outcomes for eligible households may be systematically better relative to those for ineligible households for reasons other than microcredit (Morduch 1998). Village effects may interact with eligibility to predict outcomes. Since the village dummies are in $\mathbf{x}$, this is to say that $c_f\mathbf{x}$ and $c_m\mathbf{x}$ may not be excludable from the outcome equation after all.

1.4 Computing the likelihood

The PK estimation models for various outcomes are *recursive, fully observed, mixed-process*, and *conditional*. They are recursive in that they contain clear stages—two—and do not explicitly model

[12] More precisely, these statements must hold after linearly conditioning on household-level controls.

simultaneous causation. They are fully observed (Roodman 2011) in that the observed y_f and y_m, not the latent y_f^* and y_m^*, appear in the y_o equation.[13] They are mixed-process in that they combine equations that have various types of censoring. And they are conditional in that the number of equations varies by observation, depending on the data.

A naïve approach to estimating the fully observed, recursive systems is to use a seemingly unrelated regression (SUR) likelihood.[14] In the case of the PK model, within the y_o equation, this would treat $\mathbf{y}_{fm}$ and $\mathbf{x}$ the same way mathematically, to that extent ignoring the endogeneity of $\mathbf{y}_{fm}$. An underappreciated fact, which PK implicitly exploit, is that the SUR likelihood is correct for fully-observed recursive systems (Roodman 2011). Thus, for example, the standard SUR bivariate probit estimator is consistent and efficient for a two-stage IV model in which both stages are probit (Greene 1998).[15] The econometric literature on recursive mixed-process models historically focused on multi-stage estimation procedures that are less computationally demanding than direct ML fitting, if less efficient (Amemiya 1974; Heckman 1976; Maddala 1983, chs. 7–8; Smith and Blundell 1986; Rivers and Vuong 1988). Faster computers have made ML more practical, and PK is a leading example.

2. Replication

The first edition of this paper accurately replicated many aspects of the PK regressions. Working from the survey data, we produced a data matrix for analysis that closely matches the original, going by reported first and second moments of variables. Using a new program for Stata (Roodman 2011), we

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

[14] 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 (2011).

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

also provided the first independent implementation of PK's LIML estimator.[16] However, we contradicted the signs of the key parameters, the coefficients on female microcredit borrowing, y_{f1}, y_{f2}, and y_{f3} for household consumption. This contradiction was not central to our main conclusion that instrumentation is not credible in PK, but it was dramatic and troubling.

Pitt (2011) solves the mystery. He points out that we failed to include e in $\mathbf{x}$ and set $C_v = \log$ 1,000 instead of log 1.[17] Pitt also identifies an issue that does not affect the pattern of results. We allowed the credit variables and the copies of $\mathbf{x}$ in the instrumenting equations to vary over the three rounds of data, whereas PK fixed all these at their first-round values.[18]

Also, Pitt for the first time provided a data set that contains all the variables needed to replicate PK's headline household consumption regression (but almost no other PK regressions), allowing us to switch from our data set to Pitt's.[19] By our calculations, the first and second moments of these variables closely match those reported in PK—though not exactly. (See Table 1 and Table 2.) The Pitt (2011) data set, like PK, has the disadvantage of treating current students as having zero years of schooling. But this does not significantly affect results.

We now closely replicate PK's headline regression. The first column of Table 3 shows the PK estimates for the impact of microcredit on per-capita household consumption by gender and lender. The second shows our best replication. The matches for the female credit parameters are particularly good. Those for male credit are not as close, though statistically indistinguishable. The estimated correlations between $\epsilon_o$on the one hand and ϵ_f and $\epsilon_m$on the other, labeled "ρ female" and "ρ male"

---

[16] Pitt (2011) is mistaken in stating that "*cmp* does not correctly estimate models with a non-zero censoring threshold." (Roodman 2011, p. 185)
[17] Roodman and Morduch (2011) discuss these discrepancies and the history behind them.
[18] This means that survey round dummies and the time-invariant e, though elements of $\mathbf{x}$, drop out of the instrumenting equations.
[19] The data set Pitt sent us in 2008 is missing c_f and c_m.

in the table, also match well. The apparent differences between the PK and Pitt (2011) data sets may explain the remaining differences in the regression results.

Table 1. Weighted means and standard deviations of household-level independent variables, first survey round, as reported in PK and as calculated from Pitt (2011) data

	Mean		Standard deviation	
	PK	Pitt (2011)	PK	Pitt (2011)
Parents of household head own land?	0.256	0.250	0.564	0.559
# of brothers of household head owning land	0.815	0.796	1.308	1.298
# of sisters of household head owning land	0.755	0.737	1.208	1.197
Parents of household head's spouse own land?	0.529	0.521	0.784	0.780
# of brothers of household head's spouse owning land	0.919	0.905	1.427	1.421
# of sisters of household head's spouse owning land	0.753	0.740	1.202	1.195
Household land (in decimals)	76.142	75.883	108.540	107.977
Highest grade completed by household head[1]	2.486	2.479	3.501	3.500
Sex of household head (1 = male)	0.948	0.947	0.223	0.223
Age of household head (years)	40.821	40.803	12.795	12.790
Highest grade completed by any female household member[1]	1.606	1.601	2.853	2.850
Highest grade completed by any male household member[1]	3.082	3.069	3.081	3.796515
Adult female not present in household?	0.017	0.017	0.129	0.130
Adult male not present in household?	0.035	0.036	0.185	0.185
Spouse not present in household?	0.126	0.126	0.332	0.332
Amount borrowed by female from BRAC (taka)	350.345	350.369	1,573.650	1,573.630
Amount borrowed by male from BRAC (taka)	171.993	171.973	1,565.000	1,564.920
Amount borrowed by female from BRDB (taka)	114.348	114.119	747.301	746.722
Amount borrowed by male from BRDB (taka)	203.25	202.793	1,572.660	1,571.620
Amount borrowed by female from Grameen (taka)	956.159	953.581	4,293.360	4,287.960
Amount borrowed by male from Grameen Bank (taka)	374.383	373.940	2,922.790	2,921.460
Nontarget household	0.295	0.293	0.456	0.455

N = 1,757. [1]Treats current students as having no years of schooling.

Table 2. Weighted means and standard deviations of household spending and microcredit borrowing as reported in PK and in Pitt (2011) data set

	Program villages									
	Participants		Nonparticipants		Total		Nonprogram villages		All villages	
	PK	Pitt (2011)	PK	Pitt (2011)	PK	Pitt (2011)	PK	Pitt (2011)	PK	Pitt (2011)
Cumulative female borrowing,	5,498.85	5,554.04	0.00	0.00	2,604.45	2,617.61			2,604.45	2,617.61
first survey round (1992 taka)	(7,229.35)	(7,580.10)	0.00	0.00	(5,682.40)	(5,896.01)			(5,682.40)	(5,896.01)
	N = 779	N = 779	N = 326	N = 326	N = 1,105	N = 1,105			N = 1,105	N = 1,105
Cumulative male borrowing	3,691.99	3,757.37	0.00	0.00	1,729.63	1,748.91			1,729.63	1,748.91
first survey round (1992 taka)	(7,081.58)	(7,409.36)	0.00	0.00	(5,184.67)	(5,390.53)			(5,184.67)	(5,390.53)
	N = 631	N = 631	N = 263	N = 263	N = 894	N = 894			N = 895	N = 894
Per-capita household spending,	77.014	77.014	85.886	85.886	82.959	82.959	89.661	89.661	84.072	84.072
all three survey rounds	(41.496)	(41.496)	(64.820)	(64.820)	(58.309)	(58.308)	(66.823)	(66.825)	(59.851)	(59.851)
(taka/week)	N = 2,696	N = 2,696	N = 1,650	N = 1,650	N = 4,346	N = 4,346	N = 872	N = 872	N = 5,218	N = 5,218

Table 3. Replication and robustness tests of PK fixed-effects LIML household consumption regression

Explanatory variables	PK	Replication	Probit	*De jure*
Log cumulative female borrowing, BRAC	0.0394	0.0389	0.3731	0.0428
	(4.237)***	(3.987)***	(4.649)***	(7.510)***
Log cumulative female borrowing, BRDB	0.0402	0.0407	0.3832	0.0439
	(3.813)***	(3.643)***	(4.273)***	(6.402)***
Log cumulative female borrowing, Grameen	0.0432	0.0425	0.3874	0.0446
	(4.249)***	(4.032)***	(4.844)***	(7.843)***
Log cumulative male borrowing, BRAC	0.0192	0.0156	0.1316	0.0059
	(1.593)	(0.911)	(0.721)	(0.309)
Log cumulative male borrowing, BRDB	0.0233	0.0182	0.1584	0.0095
	(1.936)*	(1.024)	(0.833)	(0.489)
Log cumulative male borrowing, Grameen	0.0179	0.0132	0.1229	0.0041
	(1.431)	(0.755)	(0.686)	(0.210)
ρ female	–0.4809	–0.4739	–0.5533	–0.5793
	(4.657)***	(4.340)***	(5.414)***	(10.363)***
ρ male	–0.2060	–0.1314	–0.1593	–0.1157
	(1.432)	(0.607)	(0.561)	(0.469)
Log likelihood	–6,634	–6,541	–4,725	–7,293
Observations	5,218	5,218	5,218	5,218

Absolute z statistics clustered by household in parenthesis. ***significant at 1%.

3. Specification problems in PK

Morduch (1998) identifies several problems in the headline PK specification. Our replication exposes more. This section inventories the problems and applies fixes where possible.

3.1 The logarithm of zero

The problem

The analysis of the logarithm of credit requires dealing with observations where credit is zero and thus the log is undefined. But here a twist is added. As displayed in (1), the PK estimation model creates a distinction between the censoring threshold for credit, C_t, and censoring value, C_v. PK set $C_t =$

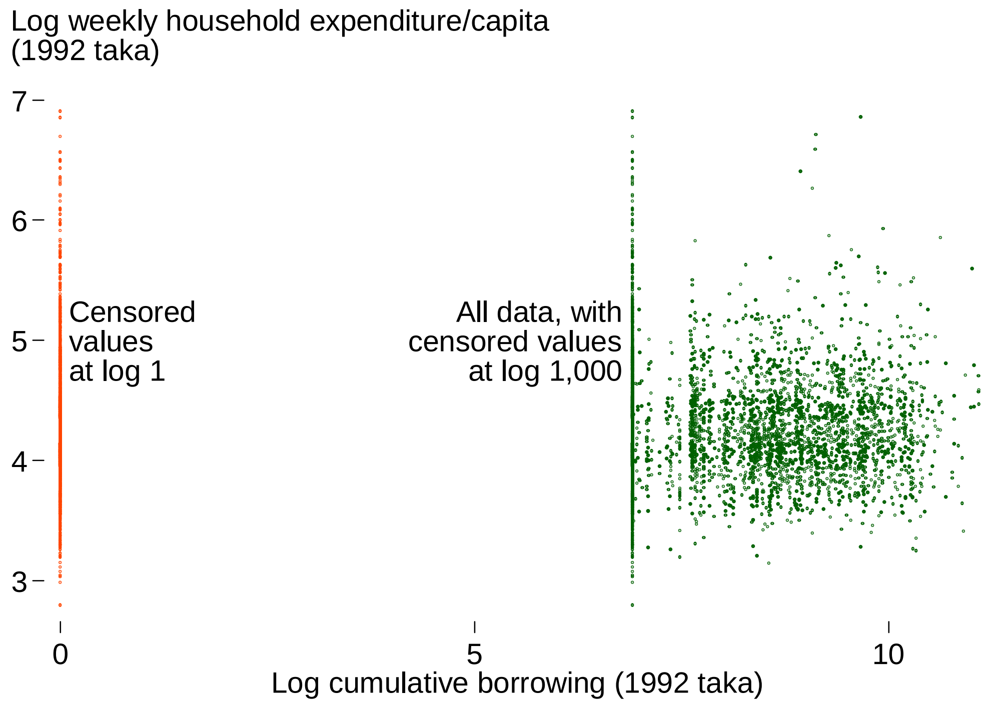
$\log 1{,}000$ since 1,000 taka is the smallest observed amount of cumulative total borrowing. They set $C_v = \log 1 = 0$.

The latter choice was first documented in Pitt (2011).[20] It is not as innocuous as it may appear. Censoring with $\log 1$ models non-borrowers as receiving 1 taka of treatment. Since household consumption is also taken in logs—so that coefficients on credit are elasticities—the implied *ceteris paribus* assumption is that moving from non-borrowing status, proxied by 1 taka, to minimal borrowing status—1,000 taka, or about $25—has the same proportional impact as moving from 1,000 to 1,000,000 taka of borrowing. (The highest observed cumulative borrowing is 58,800 taka.) That is a strong and unexamined assumption.

It is also econometrically influential. PK could have censored with, say, log 10 or log 0.1. The differences among these choices are pennies in levels, but substantial in logs. The lower the censoring value, the greater the variance in log credit, thus the lower the best-fit slope coefficients in a regression of consumption on log credit. Figure 3 illustrates by showing the data with the censoring value at log 1, which PK use, and at log 1,000, which we originally used. One can see why the slope of a best-fit line would vary substantially as the censoring value changes. Since the impact estimates in PK are based on this arbitrary choice, they do not mean what they seem to mean.

[20] An example in the PK appendix uses $C_v = 0$. However, it does not take credit in logs, so it leaves ambiguous how PK censor the log of credit.

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



The problem is that the elasticity construct—regressing logs on logs—does not allow for zeroes. Thus a hypothetical move from non-borrowing to borrowing status lies outside the construct, and can only be linked to it via an assumption about the relative impacts of such a non-marginal move as compared to a marginal increase in borrowing. A better solution to this conundrum would be to model the two kinds of variation separately, regressing on a set of dummies for borrowing by gender and lender as well as continuous variables for borrowing amounts. All would be entered in the y_o equation and all would be instrumented. But we see no good instruments for borrowing amounts separate from borrowing decisions.

In fact, the key instruments in the PK model, c_f and c_m, can be expected to be strong instruments only for the borrowing decision. Thus to the extent that the PK estimator is succeeding in identifying impacts, these are mainly the average impacts of the decision to borrow. The PK conclusion about the marginal impacts of borrowing—"annual household consumption expenditure increases 18 taka for every 100 additional taka borrowed by women from these credit programs, compared with 11 taka for men"—arises from a conversion of the average impact of the borrowing decision into a marginal impact of borrowing, made via the implicit assumption that becoming a minimal borrower has the same proportional impact as increasing borrowings a thousandfold.

A fix

How to fix this problem? More practical than simultaneously modeling the borrowing decision and borrowing amount is to focus on the decision, modeling credit as probit rather than Tobit. This circumvents the question of how to handle the log of 0 while focusing on the variation in borrowing for which credit choice is a potentially strong instrument. Ironically, PK's use of an implausibly low censoring value pushes their model in this more meaningful direction by causing the variation associated with the borrowing decision—the wide gap between log 1 and log 1,000 in Figure 1—to dominate total variation in credit. So it is not surprising that "probitizing" the model this way corroborates PK's results—indeed, strengthens them. (See column 3 of Table 3.) Going by the new point estimates, households in which women took microcredit had about $e^{0.38} - 1 = 46.2\%$ higher per-capita consumption.

3.2 A missing discontinuity

The problem

As noted, the primary identification assumption in PK is that eligibility for microcredit is exogenous after linearly conditioning on controls. PK buttress this claim by pointing to the arbitrariness of the half-acre eligibility cut-off. It is based on landownership, which itself is arguably exogenous: "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 2010), requiring that the characteristic of owning more than half an acre relates to outcomes only through microcredit (after linearly conditioning on controls, including log landholdings).

Thus whether an eligibility dummy based on the half-acre rule is exogenous is a distinct question from whether land turnover was low in the study area. The question is also less relevant than first appears, for PK use no such dummy. Morduch (1998) points out substantial and presumably endogenous mistargeting of microcredit in the PK data. We find that 203 of the 905 households in the 1991–92 sample that borrowed—a weighted 24% of borrowing households—owned more than half an acre before borrowing. They owned 1.6 acres on average. PK reclassify all as "eligible." As a result, the dummy e departs substantially from the *de jure* definition of eligibility. Meanwhile, eligibility of households in villages without credit programs is defined in the PK data strictly on the half-acre rule, raising a question about the comparability of the treatment and control groups.

So there are two caveats for the estimation here: the identifying variation in e lacks discontinuity and is presumably endogenous. To illuminate the matter, we take a recommendation from Imbens and Lemieux (2008) on graphical preliminaries to regression discontinuity estimation. We plot the key regressors, female and male borrowing, and the outcome of interest, household consumption per capita, against the continuous forcing variable, household landholdings before borrowing. (See Figure 2 and Figure 3.) We perform Lowess regressions separately for the below- and above-threshold subsamples in order to allow a discontinuity at the half-acre mark.[21] 95% confidence intervals are shown to help judge the statistical significance of the discontinuities. The plots are restricted to households for which $p_f = 1$ (for female borrowings in Figure 2), $p_m = 1$ (for male

[21] The bandwidth for the Lowess regressions is an (unweighted) 80% of the sample. The local weighting function is tricubic and incorporates PK's sample weights.

borrowings), or either (for consumption in Figure 3). As a result, they explore the effects of the interactions between the presence of credit programs and eligibility for them. This is consistent with the definitions of the key instruments: $c_f = p_f e$ and $c_m = p_m e$. (Per PK, non-borrowers along the bottom of the credit graphs are assigned 1 taka of borrowing.)

Effective enforcement of the half-acre eligibility criterion would cause the mean borrowing curves to plunge to zero at the threshold in Figure 2. Instead they hop a bit in opposite directions, without statistical significance. Meanwhile, household consumption rises, with statistical significance. But the change is in the wrong direction in the sense of contradicting the PK conclusion that lack of access to credit reduces consumption.

Evidently, loan officers were either unaware of the formal half-acre eligibility rule or pragmatically bending it to extend credit to borrowers who seemed reliable and who were, after all, poor by global standards. Some over-half-acre households that borrowed may have met an alternative eligibility criterion (see footnote 7), but this cannot explain such a substantial degree of mistargeting that runs counter to the asserted targeting of the poorest. Regardless, the asserted discontinuity is not visible in the data.

Figure 2. Cumulative borrowing vs. household landholdings before borrowing, first survey round, in villages with access to credit for given gender

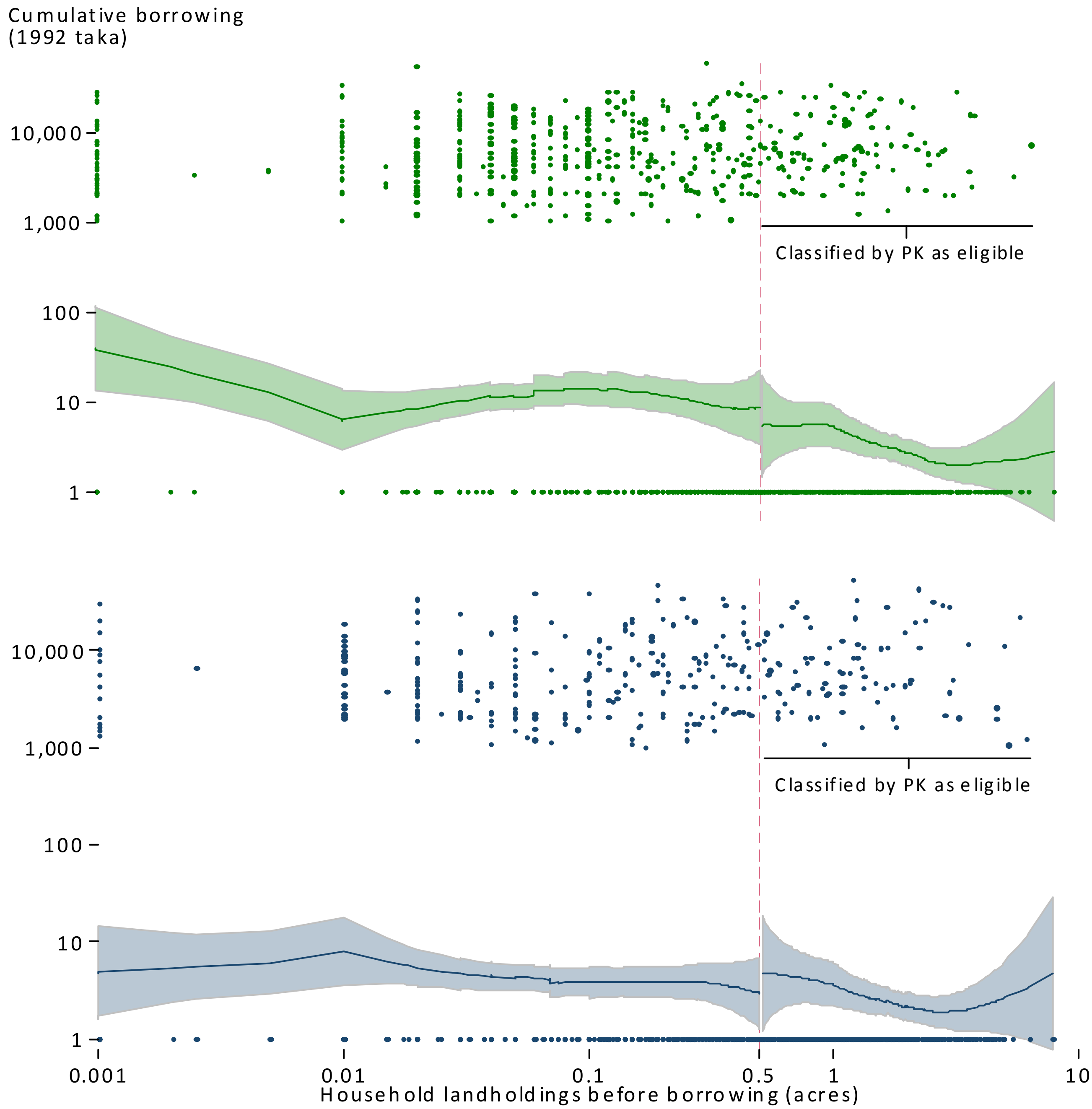
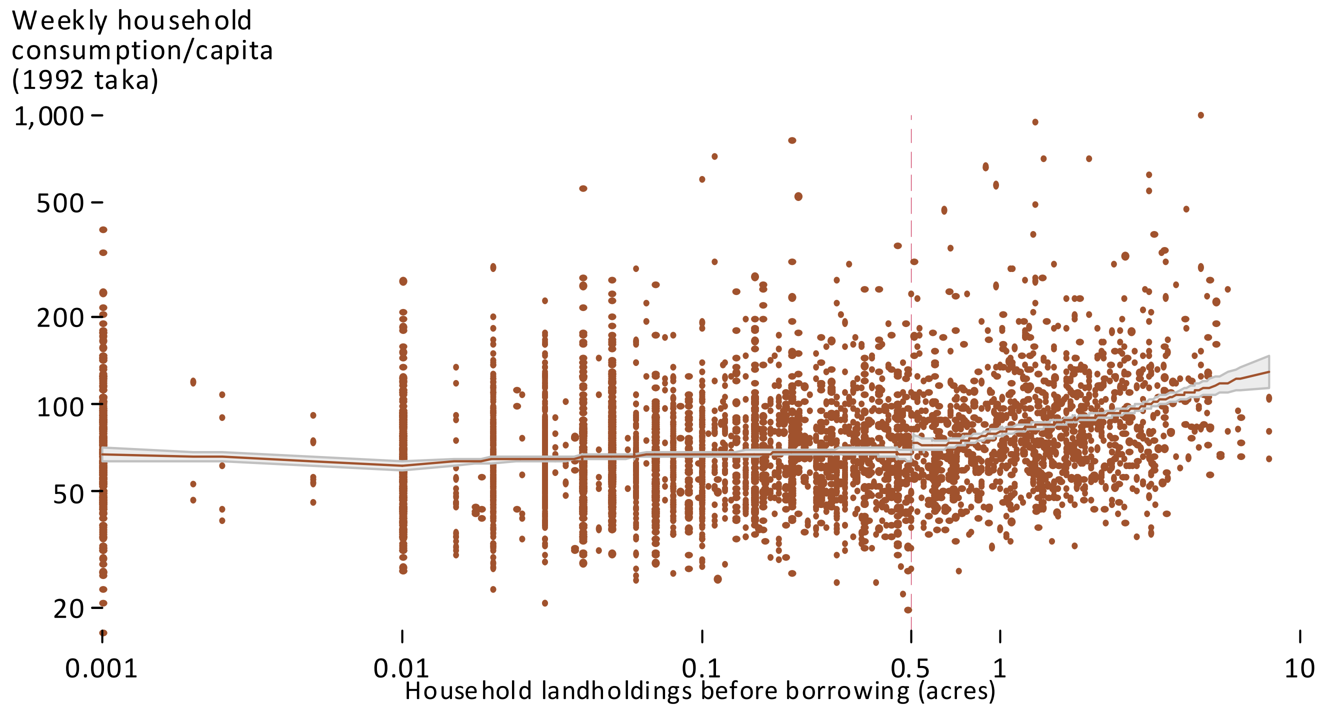


Figure 3. Average household consumption/capita over three survey rounds vs. household landholdings before borrowing in villages with access to credit for at least one gender



Morduch (1998) makes most of these points. Pitt's (1999) reply acknowledges them but argues that the true microcredit eligibility rule is "unknown" and that identification in PK's IV set-up requires only that the exogenous half-acre rule drive a component of variation in borrowing. In effect, Pitt casts the identification strategy as a Fuzzy Regression Discontinuity (FRD) design.[22]

This argument has two problems not previously noted. First, it concedes that the claimed quasi-experiment, central to PK's bid for credibility, is only asserted, not observed. Second, even if the quasi-experiment did occur, the PK model does not exploit it. In light of the pervasive non-enforcement of the rule evident in Figure 2, the eligibility dummy e as defined by PK, and thus the key instruments, $c_f = p_f e$ and $c_m = p_m e$, should be presumed endogenous.

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

A fix

To properly exploit the quasi-experiment, PK's *de facto* eligibility dummy, e, should be replaced by a *de jure* one built strictly on the half-acre rule. We check the PK regression for robustness to this change. How requires explanation. A naïve implementation would be to replace e throughout the model with $\tilde{e} \equiv \mathbf{1}\{\text{landholdings} \leq 0.5 \text{ acres}\}$. This would affect the second-stage equation, which includes e in $\mathbf{x}$, and alter the credit choice dummies, c_f and c_m, that in turn define the samples for the first-stage equations. A problem with this approach is that the mistargeted households that borrowed would now be excluded from the first-stage equations since for them $\tilde{e} = 0$. Their borrowings would not be instrumented and would be treated as exogenous. To include them in the instrumented variation while defining the samples for the instrumenting equations in a way that is more plausibly exogenous (or at least external), it is better to expand the samples for those equations to *all* households in villages with credit programs of the given gender. This treats all households in program villages as having the option to borrow. As Pitt (1999) points out, erring on the side of modeling more households as having choice does not affect consistency. Within these expanded samples, credit can then be instrumented as in (3) with $\tilde{c}_f\mathbf{x} \equiv \tilde{e}p_f\mathbf{x}$ and $\tilde{c}_m\mathbf{x} \equiv \tilde{e}p_m\mathbf{x}$. Roughly speaking, this instruments all treatments, targeted and mistargeted, with intent to treat.

Column 4 of Table 3 reports the results of such an alteration. Remarkably, it *strengthens* the PK results. This does not mean the PK instruments are valid. But it does suggest that one potential source of invalidity, violation of the eligibility rule, is not driving the PK results.

3.3 Correlation between the instruments and the error

We next examine more directly whether the instruments are valid even after the *de jure* redefinition.

ML estimation of misspecified models can be consistent for some parameters (White 1982). For example, the maximized likelihood in classical LIML, like the likelihood that corresponds to OLS, formally assumes that errors are i.i.d. normal and independent of the exogenous variables (Anderson and Rubin 1950).[23] Yet both models are consistent for the coefficients if the error is merely *uncorrelated* with the exogenous variables. The error need not be fully independent of them, nor Gaussian, nor i.i.d. Likewise, the assumptions implied by the PK model are probably weaker than formally appears to be the case. PK do not precisely analyze how the conditions in (1) and (2) can be loosened while preserving consistency for the impact parameters $\boldsymbol{\delta}$, nor formally state their own assumptions. The thrust of their text is that consistency requires that errors be identically, if not independently, distributed; and that ε_o be uncorrelated with $\mathbf{x}$, $c_f\mathbf{x}$, and $c_m\mathbf{x}$.[24]

The latter condition is technically incorrect because of the nonlinearity of the Tobit instrumenting equations; and this fact opens the door to an indirect instrument validity check. To understand our point, imagine a two-stage least squares (2SLS) regression of y on x instrumented by z. For estimation to converge, z must be excluded from the y equation. z^2, however, *could* be included, as could any other nonlinear function of z: the algorithm would still find a solution.[25] The foundation of linear estimation is that any *linear* relationship between z and $\mathrm{E}[y]$ is assumed to arise only from causal channels that run through x. That justifies the exclusion of z from the y equation, but does not require the exclusion of nonlinear functions of z.

The feasibility of estimating while controlling for z^2 is interesting for two reasons. First, it constitutes a robustness test. Second, it can shed light on the identifying assumption. If z^2 proved to have significant explanatory power in the y equation, that would contradict the hypothesis that z is

[23] Lack of correlation, recall, is equivalent to lack of *linear* relationship. As an example, a variate symmetrically distributed around 0 is uncorrelated with its square but entirely related to it.

[24] Identification conditions on pages 964 and 968 of PK state that certain errors must have zero expectation conditional on certain variables. These are conditions of lack of correlation. For binary variables, lack of correlation with and independence from error terms are synonymous. But they are distinct for non-dichotomous variables, such as age of household head, as countenanced on page 964.

[25] We assume that z is not dichotomous, so that it is not typically collinear with functions of itself.

entirely unrelated to the structural error in the model without z^2. The researcher running the regression could then give up on the instrument, conceding that it is not exogenous in the strong sense of being independent of the error. Or the researcher could maintain that while z is related to the error, it is *linearly* unrelated to it, i.e. uncorrelated with it, which is all that is technically needed for valid identification. But that argument would be weak.

PK's model turns the situation on its head. Its first stage is nonlinear, involving Tobit functions of $c_f\mathbf{x}$ and $c_m\mathbf{x}$. The necessary condition for identification is not, as stated in PK, that the instruments $c_f\mathbf{x}$ and $c_m\mathbf{x}$ are uncorrelated with ϵ_o but that any relationship between ϵ_o and $c_f\mathbf{x}$ and $c_m\mathbf{x}$ that obeys the Tobit functional form is attributable solely to credit-based channels. This assumption justifies the exclusion of Tobit functions of $c_f\mathbf{x}$ and $c_m\mathbf{x}$ from the y_o equation.

A test

Just as estimation of a linear model can proceed after including z^2, so can estimation of the nonlinear PK model proceed after entering the instruments $c_f\mathbf{x}$ and $c_m\mathbf{x}$ linearly in the structural equation.[26] This lets us directly test whether the instruments are *linearly* related to ϵ_o. If they are, that should reduce our confidence in the nonlinear conditions for identification.

Table 4 reports the results of such tests. The first column shows the effect of introducing just c_f and c_m into the second stage. The next column does the same for all of $c_f\mathbf{x}$ and $c_m\mathbf{x}$. The second pair of columns parallels the first, but with respect to the *de jure* modifications reported earlier. In all cases, the newly included instruments have clear explanatory power. As shown near the bottom of the table,

[26] This conclusion is akin to Wilde's (2000) proof that a multistage estimation model in which each stage is probit is identified without any linear exclusion restrictions.

the *p* values on the Wald *F* tests for the joint significance of the included instruments are all less than 0.05.[27]

Yet the PK results stay intact. Since including the instruments linearly does not drive out the PK results, it appears that nonlinear relationships between credit and household spending are generating the identification in PK. These relationships could be exogenous. But the linear endogeneity of the instruments makes this seem less likely.

[27] The first edition of this paper tested instrument validity via overidentification tests on analogous 2SLS regressions, and reached the same conclusion. The new approach is an improvement because it is rooted in PK's specification.

Table 4. Entering instruments linearly in the second stage

	De facto		*De jure*	
Explanatory variables	Include choice dummies	Include all instruments	Include choice dummies	Include all instruments
Log cumulative female BRAC	0.0422	0.0346	0.0440	0.0398
	(5.019)***	(3.914)***	(8.053)***	(7.351)***
Log cumulative female BRDB	0.0449	0.0401	0.0447	0.0393
	(4.615)***	(3.888)***	(6.644)***	(5.806)***
Log cumulative female Grameen	0.0459	0.0404	0.0459	0.0385
	(5.045)***	(4.329)***	(8.472)***	(7.406)***
Log cumulative male BRAC	0.0272	0.0101	0.0170	0.0072
	(2.272)**	(0.669)	(0.679)	(0.445)
Log cumulative male BRDB	0.0300	0.0139	0.0214	0.0115
	(2.499)**	(0.957)	(0.836)	(0.726)
Log cumulative male Grameen	0.0253	0.0110	0.0160	0.0037
	(2.106)**	(0.775)	(0.625)	(0.239)
ρ female	–0.4983	–0.4371	–0.5929	–0.5415
	(5.701)***	(4.406)***	(12.075)***	(10.198)***
ρ male	–0.2643	–0.0794	–0.2663	–0.1080
	(1.999)**	(0.468)	(0.842)	(0.541)
Instruments included linearly F test p value	0.041	0.000	0.000	0.000
Log likelihood	–6,530	–6,322	–7,287	–7,098
Observations	5,218	5,218	5,218	5,218

Absolute z statistics clustered by village in parenthesis. **significant at 5%. ***significant at 1%.

3.4 Weak instruments and instability

The results just shown strengthen our doubts about the PK instruments. Yet they are also unsatisfying as a diagnosis. Yes, there is evidence of endogeneity, but that evidence comes from regressions that repair what they expose, linearly controlling for the instruments. And the PK results are robust to the change.

We believe that another econometric issue is at the heart of the strong PK results: instrument weakness. In linear IV, weak instrumentation is known to cause endogeneity bias and size distortions (downward-biased standard errors that lead to over-rejection of the null of no effect) (Stock and Yogo 2005). Less appreciated is that instrument weakness can make IV bimodal.

Varying the sample by gender

Because of the unusual design of the PK estimator, no theory is available to measure instrument weakness in terms of its potential for distorting estimates. We find some evidence, however, that credit eligibility is a weak instrument for credit uptake. Differentiating credit by gender exacerbates the weakness.

That last statement requires some explanation. If the overarching goal is to estimate the impact of microcredit, it is not necessary for PK to disaggregate the estimates by gender. If they did not, they could define a single program placement dummy p; a single credit availability dummy $c \equiv pe$; and a single instrumenting equation for household borrowings, whose sample would be restricted to households for which $c = 1$. Instead, in order to identify separate impacts for female and male borrowing, PK interact e with the dummies p_f and p_m, which, recall, indicate whether microcredit groups of a given gender operated in a given village. The assumption required for this split, that p_f and p_m are exogenous, undergirds the influential finding of differential impacts by sex.

Where the exogeneity of e is defended in PK and has subsequently been challenged, the exogeneity of p_f and p_m is not motivated in PK and has received little attention since. Because the assumption is neither necessary nor defended, as a robustness test, we try dispensing with it, modifying our replication to aggregate credit across gender. (See Table 5, column 3. We report the results in the rows for female credit.) As one might expect, the point estimates lie approximately halfway between those the replication puts on male and female credit. But statistical significance is similar to that for male credit alone, which is to say low.

The loss of significance may merely be a sign of an imperfect model: female and male credit may have different impacts and so are better disaggregated. But the next four specifications challenge this defense, indeed strengthen the attack. Here, we retain PK's split by gender and instead drop parts of the sample. Since the sections dropped are defined by p_f and p_m, maintained exogenous, this step does not introduce bias under PK's assumptions. In the first of these runs, column 4 of Table 5, we drop households in villages where men can borrow. The resulting comparison of female-only villages to no-credit ones generates another estimate of the impact of microcredit for women consistent under PK's assumptions. The next regression does the same for men. The third excludes only villages where *both* women and men can borrow, generating impact estimates for each gender from non-overlapping subsamples. All variants destroy the PK results.[28]

In contrast, the last variant (in the last column) is restricted to the complement of the previous one, villages in which both women and men could borrow. The results resemble those from the full sample. The coefficients on female credit are almost the same as in PK and statistically different from 0 at 0.05 or 0.1. Yet it is precisely here that the instrumentation is weakest. For in this regression's sample, $p_f = p_m = 1$ and $c_f = c_m = e$. So the gender-differentiated dummies cannot do the job given them: differentiating impacts by gender. The instrument set (which consists, in full, of $c_f\mathbf{x} = c_m\mathbf{x}$) explains less of the independent variation in y_f and y_m than in the previous regressions.

Testing the preferred *de jure* specification in the same ways produces the same pattern of coefficients and *z* statistics as in Table 5 except that the first test, pooling across gender, preserves the significance of the female impact coefficients.[29] But that exception comes with a major caveat that will be discussed in the section on instability.

[28] No estimate for BRAC microcredit is available in the regression excluding villages where women could borrow because BRAC did not lend to men.

[29] Results available from the authors and from public data and code.

Table 5. Tests relating to identification of impacts by gender

Explanatory variables	PK	Replication	Pool accross gender[1]	Exclude villages with... Male groups	Female groups	Both kinds	One or no kinds
Log cumulative female BRAC	0.0394	0.0389	0.0239	−0.0239		−0.0168	0.0383
	(4.237)***	(3.987)***	(1.039)	(1.603)		(0.314)	(2.003)**
Log cumulative female BRDB	0.0402	0.0407	0.0283	−0.0155		−0.0073	0.0395
	(3.813)***	(3.643)***	(1.094)	(0.817)		(0.119)	(1.882)*
Log cumulative female Grameen	0.0432	0.0425	0.0255	−0.0154		−0.0075	0.0418
	(4.249)***	(4.032)***	(1.033)	(1.102)		(0.149)	(1.875)*
Log cumulative male BRAC	0.0192	0.0156				−0.0197	0.0431
	(1.593)	(0.911)				(0.763)	(3.782)***
Log cumulative male BRDB	0.0233	0.0182			−0.0011	0.0010	0.0429
	(1.936)*	(1.024)			(0.072)	(0.067)	(3.884)***
Log cumulative male Grameen	0.0179	0.0132			−0.0097	−0.0020	0.0442
	(1.431)	(0.755)			(0.455)	(0.121)	(3.677)***
ρ female	−0.4809	−0.4739				0.1779	−0.4170
	(4.657)***	(4.340)***				(0.245)	(1.872)*
ρ male	−0.2060	−0.1314				0.0964	−0.4470
	(1.432)	(0.607)				(0.605)	(3.782)***
Log likelihood	−6,634	−6,541	−5,840	−1,965	−971	−2,581	−3,860
Observations	5,218	5,218	5,218	2,189	1,478	2,795	2,423

Absolute z statistics clustered by village in parenthesis. [1]Results in female credit rows are for total borrowings by both men and women. *significant at 10%. **significant at 5%. ***significant at 1%.

Linear LIML

For want of a better way to quantify the weakness, we run linear LIML regressions that are consistent for the impact parameters under PK's assumptions, and report associated tests of underidentification and weak identification. Relative to the PK specification, the new ones expand the first-stage equations to the full sample, model credit as linear, and instrument with $c_f\mathbf{x}$ and $c_m\mathbf{x}$, rather as in (3). In the first variant, with six instrumented credit variables, the system fails the underidentification test (column 1 of Table 6.).[30] Aggregating across lenders helps the regression pass that test, but the Kleibergen-Paap

[30] We thank Mark Schaffer for this observation.

(2006) *F* statistic, a heteroscedasticity- and serial correlation–robust generalization of the Cragg-Donald *F* test for weakness (Baum, Schaffer, and Stillman 2007), is 5.606. This is well below Staiger and Stock's (1997) rule-of-thumb minimum of 10 for 2SLS. On the other hand, LIML is more robust to weak instruments.

Restricting the sample to households where both sexes can borrow—the locus of the PK results in the previous table—reduces the Kleibergen-Paap *F* to 2.116 and again causes failure on the underidentification test (column 3). Returning to the full sample and dropping the gender distinction improves matters, raising the Kleibergen-Paap *F* to 11.084. Switching to the preferable *de jure* definition of the instruments actually worsens the situation, producing underidentification even when the credit variables are aggregated across gender (last two columns). This should perhaps be expected since, as Figure 2 conveys, formal eligibility is less correlated than de facto eligibility with borrowing. Overall, the PK instrumentation may be weak even without the gender split and is certainly weakened by it.

Table 6. Tests of underidentification and weak identification in linear LIML

		Split by gender		Pool		
	Split by gender and lender	All villages	Villages with female and male groups	across gender and lender	*De jure*	
Log cumulative female borrowings, BRAC	−3.816 (0.012)					
Log cumulative female borrowings, BRDB	−0.665 (0.013)					
Log cumulative female borrowings, Grameen	−0.783 (0.012)					
Log cumulative male borrowings, BRAC	3.930 (0.012)					
Log cumulative male borrowings, BRDB	−2.229 (0.012)					
Log cumulative male borrowings, Grameen	−0.814 (0.012)					
Log cumulative female borrowings, all lenders		0.445 (0.100)	−0.407 (0.184)		0.039 (0.689)	
Log cumulative male borrowings, all lenders		0.002 (0.004)	0.715 (0.243)		−0.054 (1.027)	
Log cumulative borrowings, all lenders				1.886 (0.006)		−0.018 (1.924)*
Kleibergen-Paap underidentification test						
H_0: system is underidentified (*p* value)	0.307	0.000	0.361	0.000	0.000	0.000
Kleibergen-Paap weak identification test						
F statistic	1.577	5.606	2.116	11.084	2.987	10.123
Observations	5,218	5,218	2,423	5,218	5,218	5,218

Absolute *t* statistics clustered by household in parenthesis. *significant at 10%.

Bimodality

Because of the lack of weak instruments theory for the PK estimator, the implications of these findings are ambiguous. However, we have discovered one symptom in the PK regressions known to be associated in 2SLS and linear LIML with weak instrumentation: bimodality. We discovered this after taking the advice of Chiburis, Das, and Lokshin (2011) to bootstrap standard errors when fitting a simpler but related IV model, the two-stage probit. After finding that bootstrapping widened the confidence intervals (not reported here), we investigated the distribution of the bootstrap estimates. It turns out that when applied to PK's data, the PK likelihood has (at least) two local maxima.

As shown in Table 7, columns 2 and 3, the chief difference between the two modes is that the alternate one puts mildly negative coefficients on female credit and reverses the sign on the estimated the correlation between $\epsilon_o$and ϵ_f ("ρ female"). In other words, the variation in credit explained by the instruments and the variation not explained both experience sign flips in relation to the outcome. The alternate mode has a lower log likelihood, $-6{,}548$ instead of $-6{,}541$. But this difference probably does not deserve much weight because it is slight and, as we will show, based on an incorrect likelihood function, as ϵ_o is not normally distributed. To illustrate the situation we graph the likelihood as a function of $(1-\lambda)\hat{\beta}_- + \lambda\hat{\beta}_+$, where $\hat{\beta}_-$ and $\hat{\beta}_+$ are the two discovered modes and λ ranges between –1 and +2. While all 255 parameters vary in this cross-section of the likelihood, the biggest changes are in the female microcredit coefficients, which move in near lockstep. So we label the λ axis with the coefficient on female borrowings from the Grameen Bank. (See Figure 4.)[31] This contour is perhaps better viewed not as two peaks but as single, wide one that straddles 0 and implies that the coefficient is estimated with great imprecision. The linear LIML estimates in Table 6, with their large standard errors, seem closer to the truth.

The bimodal likelihood leads to a bimodal estimator. The mechanism is intuitive: small changes in the data can perturb the relative heights of the two peaks or even cause one to disappear. To demonstrate this effect, we bootstrap the estimator's distribution with 1,000 samples from the PK data, drawing with replacement. Since PK reweight observations within villages, we draw at the village level: whole villages are included or not. For each sample, we maximize the likelihood starting from the standard initial fit chosen by the *cmp* program.[32] We then "flip" the resulting solution by negating the credit coefficients and the ρ parameters and restarting the maximization from this modified parameter

[31] The picture is nearly identical for all three lenders.
[32] *cmp* builds the initial fit by estimating each equation separately. The initial values for the cross-equation correlations are generated from the residuals of these estimates.

vector. Usually this leads to another mode. When it does, we retain the mode with the higher likelihood. Figure 5 shows the distribution of this estimator as a histogram and as an Epanechnikov kernel density plot, again labeling with the coefficient on female Grameen Bank credit. 65% of the distribution of this coefficient is below zero. That the positive mode is assigned slightly higher likelihood in the full PK sample, and that it is the one PK found, are in a sense flukes.

Previous work has demonstrated the potential for multimodality and instability in models akin to PK's. Even the simplest multi-equation ML model—two-equation, uncensored SUR—can produce two or three maxima (Drton and Richardson 2004). That example is relevant since PK's first stage is a two-equation SUR system (complicated by Tobit censoring).

Now, in accordance with the consistency of ML, the potential for multimodality in SUR disappears asymptotically *if* the model is correct—in particular, if the modeling error is normal. But the PK structural error is not. As shown toward the bottom of Table 7, the skew and kurtosis of the residuals in the second-stage equation deviate from those of the normal distribution (skew of 0 and kurtosis of 3). The differences are significant at levels below 10^{-10} according to the normality test of D'Agostino, Belanger, and D'Agostino Jr. (1990).[33] This may explain why the bimodality persists despite the seemingly large sample of 5,218 observations.

The non-normality appears to interact with the instrument weakness to generate the bimodality. In 2SLS and classical LIML with Gaussian errors, weak instrumentation exaggerates the tendency toward bimodality (Phillips 2006). Intuitively speaking, the weakness makes the estimator especially sensitive to the outliers that account for the skew and fat tails.

[33] The deviation from normality might point to additional specification problems that could bias results, much the way that heteroscedasticity can bias Tobit estimates. In light of the bigger problems we have found, we leave this one aside.

Having uncovered this instability, we backtrack to earlier results to see if they suffer from it too. The "probitized" variant of the PK specification also turns out to harbor an alternative, negative mode, as do the variants with instruments introduced into the second stage as well as the *de jure* specification restricted to villages where at most one gender could borrow. (See Table 7.) The flipping technique finds no instability in the specification restricted to both-gender villages and the remaining ones based on *de jure* eligibility. However, lack of one symptom of instrument weakness does not guarantee freedom from the underlying affliction. The estimators may still be poorly sized, falsely rejecting the null of no impact.

To pinpoint the deviations from normality that are contributing to instability, we plot a histogram of the outcome, log per-capita household consumption. The variable has a long right tail. (See Figure 6.) So we repeatedly estimate our replication regression, first for the full sample, then excluding the highest-consumption observation, then the highest two, etc., up to the highest 49. We use the flipping procedure to search for second modes. Figure 7 plots the discovered modes along with conventionally computed 95% confidence intervals, once more labeling with the female Grameen impact coefficient. The upper rightmost dot represents our replication of the full-sample headline PK specification. Scanning to the left, we see that the bimodality—and the statistical distance from zero—disappears after dropping the 16 highest-consumption observations, which are associated with 14 households.[34] Purchases of land, home improvement, servants' wages, and dowry account for most of the high spending in these observations. One of the 14 had female microcredit borrowing and two had male. (See Table 8.)

[34] Since the concern about normality technically pertains to the residuals in the PK headline regression, not the outcome variable itself, it is arguably more correct to perform this exercise with respect to the former. Doing so produces a quite similar graph.

We should note, however, that we have found no bimodality in the superior *de jure* specification, and that its strong positive results are robust to the removal of as many as 200 highest-consumption observations (results not shown). But as shown in Table 5, even this specification is not robust to tests that vary the sample by gender.

In sum, PK's main results emerge most strongly where the ability to identify them is weakest. The attempt to differentiate impacts by gender is creating a statistical mirage. The previously unremarked instability may also explain why the arguably modest specification discrepancies in the first edition of this paper reverse the signs of the coefficients on microcredit to women.

One important lesson from the weak instruments literature is that when instruments are weak, IV methods can perform worse than non-IV ones. The best response to the apparent problem of weakness in PK may therefore be to abandon IV altogether for OLS, regressing directly on some version of the instrument set. Intent-to-treat would become a regressor rather than an instrument. Morduch (1998), for example, regresses on three credit availability dummies, one each for BRAC, BRDB, and the Grameen Bank. His linear regressions, like ours, erase the strong positive coefficients on microcredit for women.

Table 7. Replication variants with two discovered modes

Explanatory variables	PK	Replication		Model credit as probit		Including instruments in 2nd stage: Choice dummies		Including instruments in 2nd stage: All instruments		*De jure*, excluding villages with male & female groups	
Log cumulative female borrowing, BRAC	0.0394	0.0389	−0.0191	0.3731	−0.2315	0.0422	−0.0244	0.0346	−0.0145	0.0369	−0.0284
	(4.237)***	(3.987)***	(1.287)	(4.649)***	(2.008)**	(5.019)***	(2.485)**	(3.914)***	(0.817)	(3.329)***	(1.734)*
Log cumulative female borrowing, BRDB	0.0402	0.0407	−0.0219	0.3832	−0.2578	0.0449	−0.0274	0.0401	−0.0125	0.0559	−0.0228
	(3.813)***	(3.643)***	(1.334)	(4.273)***	(2.050)**	(4.615)***	(2.446)**	(3.888)***	(0.643)	(4.137)***	(1.132)
Log cumulative female borrowing, Grameen	0.0432	0.0425	−0.0183	0.3874	−0.1949	0.0459	−0.0252	0.0404	−0.0133	0.0460	−0.0211
	(4.249)***	(4.032)***	(1.200)	(4.844)***	(1.759)*	(5.045)***	(2.504)**	(4.329)***	(0.693)	(4.883)***	(1.292)
Log cumulative male borrowing, BRAC	0.0192	0.0156	0.0221	0.1316	0.2272	0.0272	0.0299	0.0101	0.0114	−0.0198	−0.0195
	(1.593)	(0.911)	(1.460)	(0.721)	(1.786)*	(2.272)**	(2.648)***	(0.669)	(0.623)	(0.784)	(0.768)
Log cumulative male borrowing, BRDB	0.0233	0.0182	0.0232	0.1584	0.2267	0.0300	0.0331	0.0139	0.0170	−0.0045	−0.0049
	(1.936)*	(1.024)	(1.463)	(0.833)	(1.682)*	(2.499)**	(2.947)***	(0.957)	(0.941)	(0.342)	(0.410)
Log cumulative male borrowing, Grameen	0.0179	0.0132	0.0214	0.1229	0.2144	0.0253	0.0304	0.0110	0.0144	−0.0062	−0.0078
	(1.431)	(0.755)	(1.385)	(0.686)	(1.695)*	(2.106)**	(2.678)***	(0.775)	(0.814)	(0.373)	(0.483)
ρ female	−0.4809	−0.4739	0.3160	−0.5533	0.4277	−0.4983	0.3814	−0.4371	0.2344	−0.5364	0.3388
	(4.657)***	(4.340)***	(1.716)*	(5.414)***	(2.575)**	(5.701)***	(3.483)***	(4.406)***	(1.003)	(5.793)***	(1.734)*
ρ male	−0.2060	−0.1314	−0.2397	−0.1593	−0.2932	−0.2643	−0.3394	−0.0794	−0.1387	0.1479	0.1727
	(1.432)	(0.607)	(1.274)	(0.561)	(1.518)	(1.999)**	(2.798)***	(0.468)	(0.633)	(1.130)	(1.528)
2nd-stage errors											
Skew		0.64	0.63	0.62	0.59	0.60	0.58	0.55	0.58	0.55	0.60
Kurtosis		4.78	4.90	4.72	4.73	4.62	4.72	4.48	4.69	4.38	4.44
Instruments included linearly (*F* test *p* value)						0.041	0.029	0.000	0.000		
Log likelihood	−6,634	−6,541	−6,548	−4,725	−4,732	−6,530	−6,536	−6,322	−6,329	−2,726	−2,734
Observations	5,218	5,218	5,218	5,218	5,218	5,218	5,218	5,218	5,218	2,795	2,795

Absolute *z* statistics clustered by village in parenthesis. All skew and kurtosis figures reject the hypothesis of normal distribution (skew = 0, kurtosis = 3) at *p* values below 10^{-10}. *significant at 10%. **significant at 5%. ***significant at 1%.

Figure 4. A cross-section of the PK likelihood on PK data, with two local maxima marked

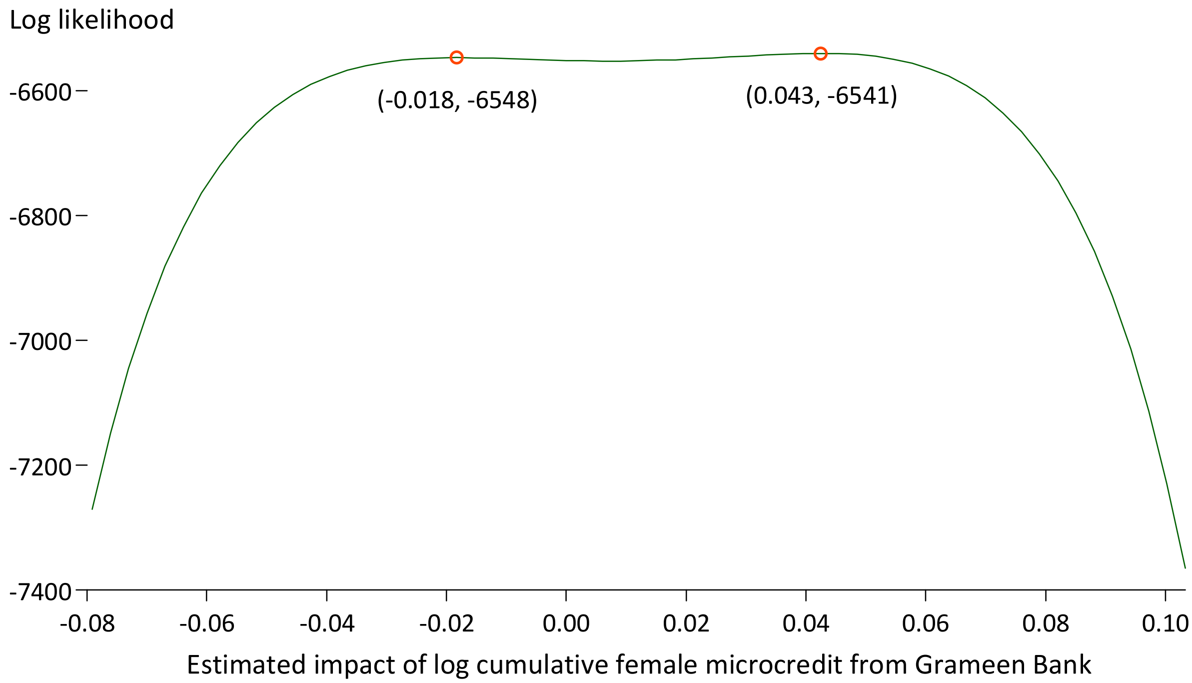


Figure 5. Bootstrapped distribution of the PK estimator on PK data, 1,000 replications

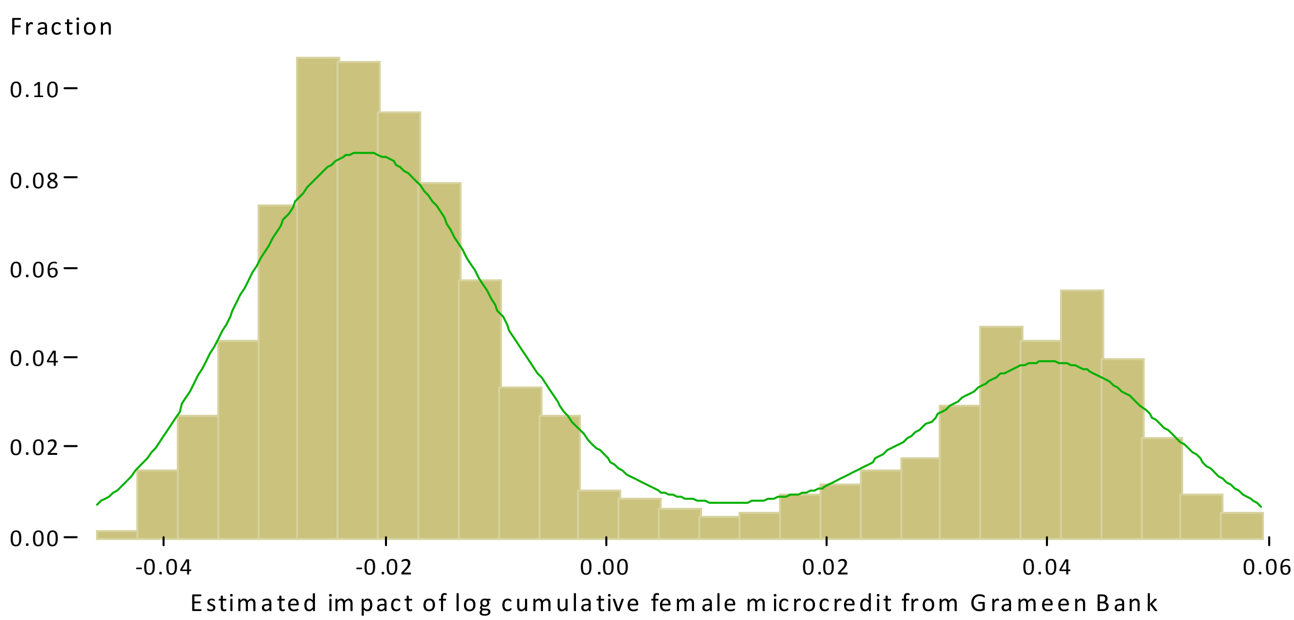


Figure 6. Weighted distribution of log per-capita household consumption, Pitt (2011) data

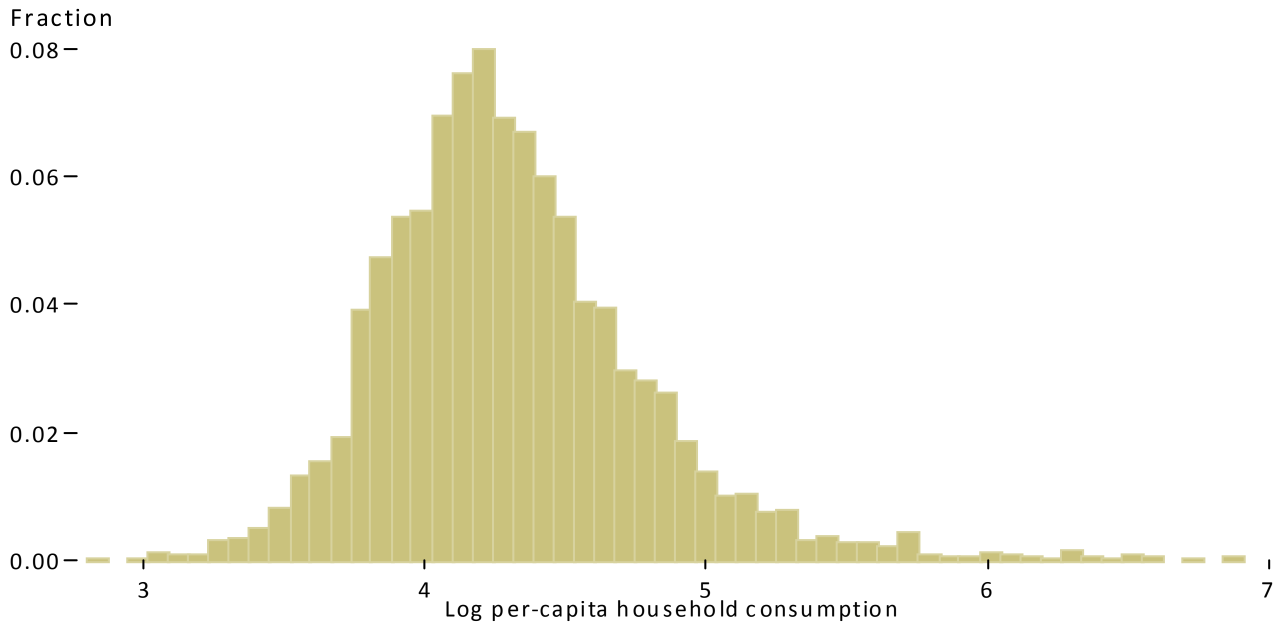


Figure 7. Modes of PK estimator on Pitt (2011) data when highest-consumption observations are excluded from sample

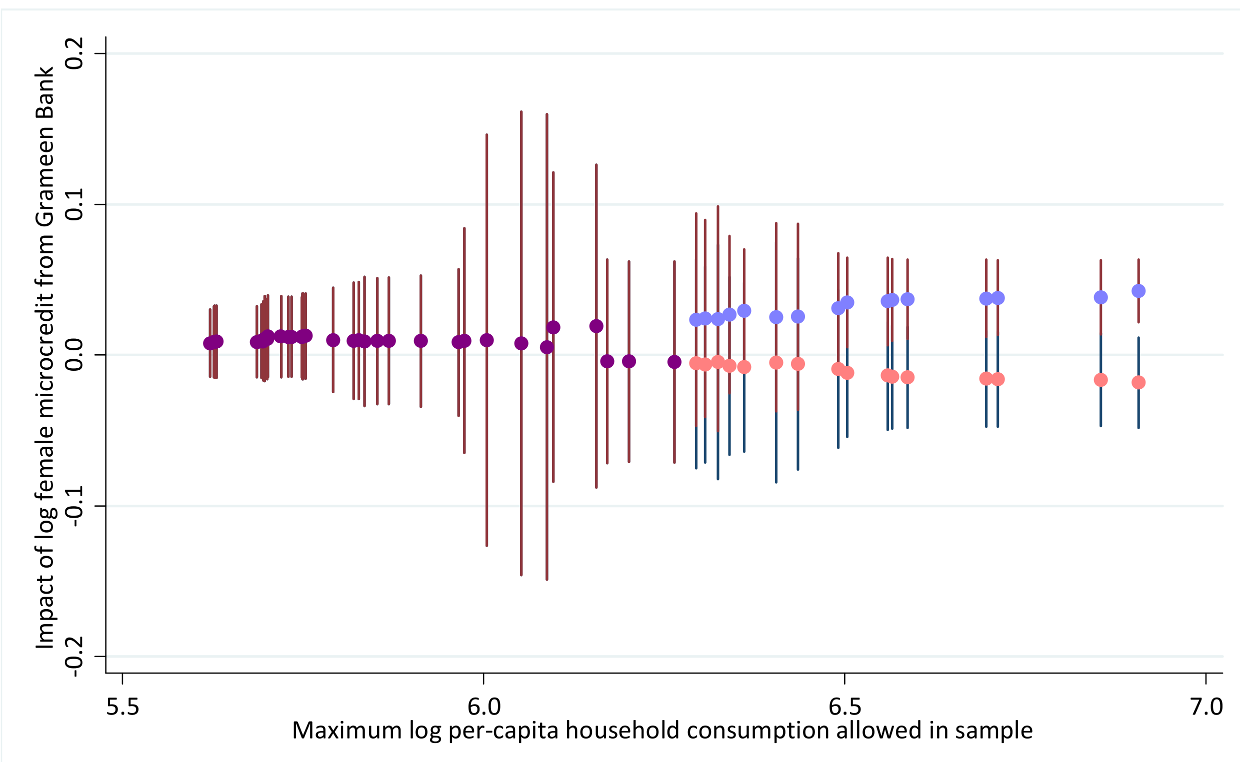


Table 8. Major expenditures of 16 highest-consumption observations

Household ID	Seasonal survey round	Division	District	Cumulative borrowing: Female	Cumulative borrowing: Male	Major expenditures: Purpose	Major expenditures: Amount	Total spending	Household members	Log consumption/ capita/week[1]
281035	2	Khulna	Jhenaidah	0	0	Home improvent	14,000	43,317	5	6.29
231025	3	Dhaka	Narayanganj	0	0	Home improvent	50,000	61,409	7	6.31
32014	1	Sylhet	Habiganj	0	0	Land/property purchase Ceremony	10,000 15,000	35,718	4	6.32
83015	3	Dhaka	Sherpur	0	0	Land/property purchase	53,000	64,827	7	6.36
52121	1	Dhaka	Manikganj	7,539	0	Land/property purchase	15,000	19,356	2	6.41
231025	2	Dhaka	Narayanganj	0	0	Home improvent Land/property purchase	40,000 10,000	69,830	7	6.44
133035	1	Dhaka	Mymensingh	0	0	Servant's wage	37,000	42,209	4	6.49
133035	2	Dhaka	Mymensingh	0	0	Land/property purchase	35,000	42,745	4	6.50
82015	2	Dhaka	Sherpur	0	0	Land/property purchase Dowry	27,000 10,000	45,173	4	6.56
72025	1	Rajshahi	Rangpur	0	9,044	Home improvent	50,000	68,225	6	6.57
121042	1	Dhaka	Kishoreganj	0	0	Servant's wage	9,000	11,613	1	6.59
282064	2	Khulna	Jhenaidah	0	0	Land/property purchase	30,000	38,843	3	6.70
132042	1	Dhaka	Mymensingh	0	9,103	Servant's wage	23,000	26,308	2	6.71
231015	3	Dhaka	Narayanganj	0	0	Home improvent	130,000	151,820	10	6.86
113035	3	Khulna	Bagerhat	0	0	Land/property purchase	53,000	63,978	4	6.91

All money figures in 1992 taka. $1≈40 taka. [1]Assuming 16 weeks/season.

4. Replication of additional results

Our final step is to replicate PK's results for the five outcomes other than per-capita household consumption. We do this for completeness and to put the earlier results in corroborating perspective.

The five additional outcomes are not in either of the data sets provided by Mark Pitt, so we construct them from the underlying survey data. Table 9 benchmarks our variables against the means and standard deviations reported in PK. The match is extremely good for male labor supply and boys' and girls' school enrollment. It is poorer for female labor supply. But here we have reason to doubt PK's numbers. Pitt and Khandker (2002, Table 1) reports the same means alongside seasonal subaverages with which they are mathematically incompatible.[35] Finally, the biggest discrepancies are in female-owned non-land assets. As shown, we obtain a much better match if we add land to nonland assets.

We perform LIML fixed-effect estimates of the impact of microcredit on these outcomes, similar to the household consumption specification studied above.[36] None of the results match PK exactly, but all are similar in signs and significance. The worst replication, as in the previous table, is for female non-land assets, though even here, the results are not statistically different. Adding land to nonland assets, which helped in the previous table, does not generate a good match.

We check all regressions for a second mode in the same manner as before. We find one only for male labor supply. The two discovered modes clash on the sign of the impact of male microcredit borrowing. (See Table 10.)

---

[35] For example, as also shown in our Table 9, Pitt and Khandker (2002) report that female labor supply averages 43.934 hours/month in nonprogram villages over all three seasonal survey rounds—and 29.121 in *aman*, 29.728 in *boro*, and 31.290 in *aus*. These three numbers cannot average 43.934.

[36] One subtle difference is that all of these outcomes are modeled as censored, so for observations in which the outcome, female credit, and male credit are all censored, the likelihood involves cumulative normal distributions of dimension 3. The *cmp* program computes these with the GHK (Geweke-Hajivassiliou-Keane) simulation-based algorithm. We use 100 Halton draws per observation along with antithetics.

PK observe that only for household consumption and male labor supply do the ρ parameters in the IV estimates differ statistically from 0, indicating that credit is in fact endogenous to those outcomes. For the other four outcomes, PK prefer uninstrumented fixed-effects regressions, which tend to produce stronger evidence of (positive) impact. However, this argument is based on the assumption that the instruments are valid, which we have strongly questioned.

Given our doubts about the instrumentation strategy, we think the most important thing to observe about the six PK LIML fixed-effects estimates is that they are of two sorts. One kind (female non-land assets, female labor supply, girls' and boys' school enrollment) features insignificant coefficients on credit, insignificant ρ parameters, and no apparent multimodality. The other (household consumption and male labor supply) features strong coefficients, significant ρ parameters, and bimodality that produces starkly contradictory impact coefficients for one sex. This is evident from a quick scan of Table 10: the bimodality surfaces in the same place as most of the asterisks. This strengthens our suspicion that bimodality is the proximate cause of the seemingly significant results in PK's household consumption regression. Perhaps it also not a coincidence that the instability occurs with the two least-bounded outcomes, the ones with the most scope for extreme deviations from Gaussianity. Log household consumption is an unbounded variable. Male labor supply is bounded from below but assumes its bounding value of 0 in only 8.6% of cases on a weighted basis.

Table 9. Weighted means and standard deviations of non-consumption outcomes, as reported in PK and in new data set

	Program villages									
	Participants		Nonparticipants		Total		Nonprogram villages		All villages	
	PK	New	PK	New	PK	New	PK	New	PK	New
Female nonland assets,	7,399.23	2,366.09	4,716.42	1,724.55	5,608.03	1,937.76	1,801.84	831.84	4,970.67	1,752.57
first survey round (taka)	(293.02)	(6,693.24)	(19,901.04)	(5,033.62)	(23,509.09)	(5,645.45)	(6,287.49)	(2,207.09)	(21,649.42)	(5,245.48)
	N = 899	N = 899	N = 542	N = 542	N = 1,441	N = 1,441	N = 292	N = 292	N = 1,733	N = 1,733
Female assets,		7,512.51		4,793.83		5,697.37		1,975.24		5,074.08
first survey round (taka)[1]		(31,572.90)		(19,922.00)		(24,443.40)		(6,428.01)		(22,498.90)
		N = 899		N = 542		N = 1,441		N = 292		N = 1,733
Women's labor supply,	40.328	40.389	37.68	32.467	38.905	35.087	43.934	31.269	39.54	34.467
all survey rounds	(70.478)	(70.558)	(71.325)	(64.297)	(70.934)	(66.529)	(74.681)	(60.214)	(71.432)	(65.556)
(hours/month, aged 16–59)	N = 3,420	N = 3,420	N = 2,108	N = 2,108	N = 5,528	N = 5,528	N = 1,074	N = 1,074	N = 6,602	N = 6,602
Men's labor supply,	202.758	202.749	185.858	185.758	191.310	191.239	180.940	180.528	189.477	189.346
all survey rounds	(10.527)	(100.820)	(104.723)	(104.904)	(103.678)	(103.897)	(98.805)	(99.405)	(102.902)	(103.191)
(hours/month, aged 16–59)	N = 3,534	N = 3,534	N = 2,254	N = 2,254	N = 5,788	N = 5,788	N = 1,126	N = 1,126	N = 6,914	N = 6,914
School enrollment of girls	0.535	0.535	0.528	0.527	0.531	0.530	0.552	0.552	0.534	0.534
aged 5–17, first survey	(0.499)	(0.499)	(0.500)	(0.500)	(0.499)	(0.499)	(0.498)	(0.498)	(0.499)	(0.499)
round (yes = 1)	N = 802	N = 802	N = 434	N = 434	N = 1,236	N = 1,236	N = 225	N = 225	N = 1,461	N = 1,461
School enrollment of boys	0.566	0.566	0.555	0.556	0.558	0.559	0.550	0.553	0.559	0.558
aged 5–17, first survey	(0.496)	(0.496)	(0.498)	(0.497)	(0.497)	(0.497)	(0.497)	(0.498)	(0.497)	(0.497)
round (yes = 1)	N = 856	N = 856	N = 468	N = 468	N = 1,324	N = 1,324	N = 265	N = 267	N = 1,589	N = 1,591

[1]Aggregates for this variable are displayed to show their similarity to PK's reported aggregates for non-land assets.

Table 10. LIML fixed-effects estimates of impact of microcredit on outcomes other than consumption, PK and new

	Log female non-land assets		Log female assets	Log female hours worked per month		Log male hours worked per month			School enrollment of girls, 5–17		School enrollment of boys, 5–17	
	PK	New	New	PK	New	PK	New		PK	New	PK	New
Log cumulative female	0.0318	0.1057	0.0833	–0.0117	–0.0017	–0.1813	–0.2165	–0.2276	–0.0203	–0.0566	0.0394	–0.005
borrowing, BRAC	(0.356)	(1.487)	(0.965)	(0.128)	(0.017)	(5.884)***	(8.254)***	(9.409)***	(0.552)	(1.196)	(0.917)	(0.082)
Log cumulative female	0.1257	0.1565	0.0988	–0.0139	–0.0260	–0.2308	–0.2654	–0.2723	–0.0099	–0.0540	0.1210	0.074
borrowing, BRDB	(1.043)	(1.716)*	(1.146)	(0.139)	(0.228)	(7.066)***	(9.290)***	(10.499)***	(0.220)	(0.985)	(2.573)**	(1.153)
Log cumulative female	0.1131	0.1682	0.1436	0.0152	0.0250	–0.2189	–0.2124	–0.2183	0.0128	–0.0300	0.1025	0.063
borrowing, Grameen	(1.317)	(2.348)**	(1.665)*	(0.162)	(0.275)	(6.734)***	(9.126)***	(10.103)***	(0.334)	(0.717)	(2.364)**	(1.324)
Log cumulative male	0.1005	0.0137	–0.0209	–0.0448	–0.1052	–0.1369	–0.1634	0.0185	0.0495	0.0020	–0.0040	–0.000
borrowing, BRAC	(0.468)	(0.130)	(0.243)	(0.520)	(1.085)	(2.155)**	(3.917)***	(0.508)	(1.152)	(0.027)	(0.107)	(0.008)
Log cumulative male	0.0334	–0.0940	–0.1476	–0.0144	–0.0857	–0.1440	–0.1713	0.0153	0.0321	–0.0161	0.0361	0.052
borrowing, BRDB	(0.141)	(0.880)	(1.711)*	(0.181)	(0.916)	(2.129)**	(4.018)***	(0.407)	(0.665)	(0.239)	(0.934)	(0.962)
Log cumulative male	–0.0457	–0.1039	–0.1798	–0.0570	–0.1044	–0.1592	–0.1584	–0.0076	0.0582	–0.0004	0.0736	0.095
borrowing, Grameen	(0.200)	(0.953)	(2.085)**	(0.677)	(1.241)	(2.524)**	(4.354)***	(0.230)	(1.298)	(0.006)	(1.688)*	(2.033)**
ρ female	0.1136	0.0719	0.0913	0.1255	0.1192	0.6564	0.6942	0.7358	0.1648	0.3038	0.2192	0.0175
	(1.325)	(0.658)	(0.757)	(1.062)	(0.975)	(7.461)***	(10.615)***	(14.315)***	(1.029)	(1.568)	(1.054)	(0.072)
ρ male	–0.0148	0.1527	0.2255	0.0560	0.1298	0.4929	0.5091	–0.0264	–0.1360	0.0918	–0.0284	–0.1650
	(0.053)	(0.883)	(1.124)	(0.592)	(1.228)	(2.512)**	(4.070)***	(0.257)	(0.720)	(0.303)	(0.177)	(0.767)
Observations	1,757	1,757	1,757	6,602	6,602	6,914	6,914	6,914	2,885	1,461	2,940	1,591
Log pseudolikelihood		–4,048	–4,195		–17,552		–20,865	–20,875		–2,446		–2,737

Regressions for first two columns run on household-level data and the rest on individual-level.Absolute *z* statistics clustered by household in parenthesis. *significant at 10%. **significant at 5%. ***significant at 1%.

5. Conclusion

Pitt and Khandker (1998) is in many ways a brilliant study of an important question. But its econometric sophistication backfires. The study is hard to understand and its complexities hide several major problems. These include:

- an imputation for the log of the treatment variable when it is zero that is undocumented, influential, and arbitrary at the margin;
- the absence of a discontinuity that is asserted as central to identification;
- a reclassification of formally ineligible but borrowing households as eligible, which presumably introduces endogeneity into the asserted quasi-experiment;
- evidence of a linear relationship between the instruments and the error;
- evidence of instrument weakness, especially when microcredit borrowings are disaggregated by sex;
- disappearance of the results when villages where both genders could borrow are excluded;
- bimodality driven by 16 outliers from 14 households.

Although the last finding is most striking, it is not the whole story. One reason is that the preferred *de jure* estimator does not appear to be bimodal and is robust to removing many high-consumption observations. More deeply, observing sensitivity to outliers does not explain it. What can explain it is instrument weakness caused by disaggregating credit by gender. Even the *de jure* results are fragile to the sample-varying tests on this theme. Thus we see the dependence on the econometrically problematic gender split as closer to the heart of the matter.

As we observed in the first version of this paper, Pitt and Khandker (1998) reinforced some broad ideas about microcredit: that it reduces poverty, and that it does so especially when given to women. In our view, nothing in the present paper contradicts those ideas, for the elementary reason that absence of evidence—lack of identification—is not evidence of absence. But the paper should reduce *confidence* in those ideas to the extent they were pegged on PK. Our critical conclusions about PK, along with our challenge to Khandker's (2005) follow-up (in the first version of this paper) and the muted results of the randomized trials of microcredit (Banerjee et al. 2009; Karlan and Zinman 2011; Crépon et al. 2011), combine to mean that 35 years into the microfinance movement, evidence in favor of the proposition that microcredit reduces poverty is extraordinarily scarce (Armendáriz and Morduch 2010, chapter 9; Odell 2010; Duvendack et al. 2011; Roodman 2012, chapter 6).

One broad question raised by our analysis is about the value of non-randomized studies. Our prior is that exclusive reliance on one type of study is not optimal. But for non-randomized studies to contribute to the measurement of causation in social systems where endogeneity is pervasive, the quality of the natural experiments must be high, and demonstrated.

Our replication also raises questions about quality control in the production of economics. Of course, hindsight is 20/20. So we point up this issue not to engage in retrospective perfectionism but to draw lessons for the practice of economics today. What was and is reasonable to expect is that authors, reviewers, and journal editors take steps to prevent methodological complexity from obscuring fundamental issues of identification. Assumptions should be checked to the extent they can be, the example at hand being the asserted discontinuity. Dependence on secondary assumptions, such as that required in PK for the identification of impacts by gender, should be tested. Where possible, complex estimators should be checked by simpler ones.

A more radical strategy for quality control is transparency: sharing data and code at the working paper stage. The more freely researchers circulate their data and code, the easier it is for others to subject that work to the scrutiny needed for science to proceed. The stakes are particularly high for research that influences policy (McCullough and McKitrick 2009). The *Journal of Political Economy*, which published PK, now requires such disclosure. Our own transparency allowed Pitt to find the errors in our replication.[37] Had *JPE*'s policy been in force when it published PK in 1998, the debate over the study might have been resolved long ago.

[37] Data and code for both versions of this paper are posted at http://j.mp/gpXmI1.

References

Anderson, T.W, and Herman Rubin. 1950. The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics* 21(4): 570–82.

Appelbaum, Richard. 2008. The Man Who Is Creating a World without Poverty. *Santa Barbara Independent*. January 10.

Armendáriz de Aghion, Beatriz, and Jonathan Morduch. 2010. *The Economics of Microfinance*. 2nd ed. Cambridge, MA: The MIT Press.

Banjerjee Abhijit V., and Esther Duflo. 2009. The Experimental Approach to Development Economics. *Annual Review of Economics* 1:151–78

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

Baum, Christopher, Mark E. Schaffer, and Steven Stillman. 2007. Enhanced Routines for Instrumental Variables/GMM Estimation and Testing. *Stata Journal* 7(4): 465–506.

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

Chiburis, Richard C., Jishnu Das, and Michael Lokshin. 2011. A Practical Comparison of the Bivariate Probit and Linear IV Estimators. Policy Research Working Paper 5601. World Bank.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté. 2011. Impact of Microcredit in Rural Areas of Morocco: Evidence from a Randomized Evaluation. Massachusetts Institute of Technology.

D'Agostino, R. B., A. J. Belanger, and R. B. D'Agostino Jr. 1990. A suggestion for using powerful and informative tests of normality. American Statistician 44: 316–321.

Deaton, Angus. 2009. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* 48(2): 424–55.

Drton, Mathias, and Thomas S. Richardson 2004. Multimodality of the Likelihood in the Bivariate Seemingly Unrelated Regressions Model. *Biometrika* 91(2): 383–92.

Duvendack, Maren, Richard Palmer-Jones, James G Copestake, Lee Hooper, Yoon Loke, Nitya Rao. 2011. *What Is the Evidence of the Impact of Microfinance on the Well-being of Poor People?* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

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

Karlan, Dean, and Jonathan Zinman. 2011. Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation. *Science* 332(6035): 1278–84.

Kleibergen, F., and R. Paap. 2006. Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics* 127(1): 97–126.

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

McCullough, B.D., and Ross McKitrick. 2009. *Check the Numbers: The Case for Due Diligence in Policy Formation.* Fraser Institute.

De Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008. "Returns to Capital in Microenterprises: Evidence from a Field Experiment." *Quarterly Journal of Economics* 123(4): 1329–72.

McKenzie, David, and Christopher Woodruff. 2008. "Experimental Evidence on Returns to Capital and Access to Finance in Mexico." *World Bank Economic Review* 22 (3): 457–82.

Morduch, Jonathan. 1998. Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh. New York University. Department of Economics. Available at j.mp/bC3Tge.

Odell, Kathleen. 2010. *Measuring the Impact of Microfinance: Taking Another Look*. Washington, DC: Grameen Foundation.

Phillips, Peter C.B. 2006. A Remark on Bimodality and Weak Instrumentation in Structural Equation Estimation. Paper 1171. Cowles Foundation for Research in Economics. Yale University.

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. j.mp/dLNltJ.

Pitt, Mark M. 2011. Response to Roodman and Morduch's "The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence." Brown University. j.mp/j4x2xV.

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.

Roodman, David. 2009b. Estimating fully observed recursive mixed-process models with cmp. *Stata Journal* 11(2): 159–206.

Roodman, David. 2012. *Due Diligence: An Impertinent Inquiry into Microfinance*. Brookings Institution.

Roodman, David, and Jonathan Morduch. 2011. Comment on Pitt's Responses to Roodman & Morduch (2009). Center for Global Development.

Staiger, Douglas, and James H. Stock. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica* 65(3): 557–86.

White, Halbert. 1982. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* 50(1): 1–25.

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

Yunus, Muhammad. 1999. The Grameen Bank. *Scientific American*. November.

Yunus, Muhammad. 2007. Q&A with Muhammad Yunus. Interview on "NOW." PBS. j.mp/oEeNni.

Yunus, Muhammad. 2008. Credit for the Poor. *Harvard International Review.* January.

Yunus, Muhammad, and Fazle Abed. 2004. Responses to New York Times Editorial Regarding New US Law on Poverty Measurement Tools. j.mp/o5xjq1.