

Comment on Pitt's Responses to Roodman & Morduch (2009)

David Roodman and Jonathan Morduch

December 2011

Non-technical Overview

A puzzle

Roodman and Morduch (RM, 2009) examine three microcredit impact studies based on a survey fielded in Bangladesh in the 1990s. The first and most influential of the studies is Pitt and Khandker (PK, 1998), published in the respected *Journal of Political Economy*. PK find a strong positive impact of microcredit borrowing on household spending, especially when the credit goes to women. The results are important because (1) they are derived from a carefully-designed econometric model that pays attention to potential sources of bias and (2) newer studies fail to find such strong impacts of microcredit borrowing. PK thus remains the chief evidence of the power of microcredit.

Some other studies of the same survey data from Bangladesh fail to corroborate PK. Using different methods, Morduch (1998) finds results that are difficult to square with those of PK. RM replicate Morduch (1998) but they fail to replicate PK. Specifically, RM find opposite signs on the key quantities of interest, those for the impact of female microcredit borrowing on household spending.

RM emphasize that they do not interpret their negative signs as showing that microcredit increases poverty. Rather, they argue that cause and effect cannot be inferred from the Bangladesh data. Thus RM do not question the plausibility of the PK results, only the statistical basis. Nevertheless, the sign mismatch on the relationship between microcredit and household spending remained a significant mystery.

A big step forward

Mark Pitt (2011a) solved the mystery. He finds two key ways in which RM diverge from PK's analytical methods. His work has led us revise our regressions in order to closely match the PK. We now understand how PK discerned a positive relationship between female microcredit borrowing and household spending.

But we still do not share PK's confidence about the conclusion that "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." One of our reasons relates directly to Pitt's comments. One issue he describes as a "logical error" on our part is in our view a surface manifestation of a deeper problem that RM pointed out, which undermines PK's claims to have estimated the *magnitudes* of the impacts of microcredit.

In addition, now that we can match PK much better, we have performed analyses that give us more insight into what, mathematically, is driving the PK results. These findings strengthen our conclusion about the lack of demonstrated causality from microcredit to poverty reduction. We report these in a separate paper.

Replication and transparency

The ongoing dialogue with Mark Pitt illustrates the value of replication and transparency in econometric research. The back and forth has shown that neither side has a monopoly on the truth. We have gained a better understanding of PK than we—or anyone else, going by the public record—had before.

This progression would have happened much faster if the *Journal of Political Economy's* current policy requiring authors to share data and code had been in place when the journal published PK. That would have given us

greater certainty about how the original analysis was carried out. In contrast, our decision to share all our data and code helped Pitt find an important error and improve our work.

Technical overview

In March 2011, Mark Pitt circulated a response (2011a) to RM. The two main points are that RM incorrectly use a censoring value of log 1,000, not log 1, for log microcredit borrowing by non-borrowers; and that a control for a household's eligibility for microcredit ("target status") is missing. Addressing both these issues indeed bring us to a close replication.

However, we note that:

- Though framed as "correct[ing] the substantial damage that [RM's] claims have caused to the reputation of microfinance," Pitt (2011a) does not address the main conclusion of RM, which is that the identification strategy is not credible.
- The choice of censoring value for log credit, the focus of one of Pitt's (2011a) main points, is arbitrary at the margin, yet directly affects the estimated impacts. This issue was described in RM and is elaborated upon here.
- All of the discrepancies that Pitt points out are to a degree traceable to our lack of access to a well-documented processed data set and, especially, the estimation code. Had *JPE's* current disclosure policy been in force in 1998, these problems would probably have been avoided and the cause of doing replicable science would have been served.

RM doubt PK's identification strategy in part because of is the poor results on overidentification-based instrument validity tests run on Two-Stage Least Squares (2SLS) regressions that parallel PK's LIML regressions. RM report such tests, as does a preliminary blog post from Roodman (2011). In response to the latter, Pitt wrote a second note (draft, 2011b) that characterizes the 2SLS-based instrument validity tests as "fundamentally flawed." But in contrast to Pitt (2011a), Pitt (2011b) hardly persuades us to make any changes to our analysis.

But we now realize that the 2SLS regressions have a weakness that Pitt (2011b) does not mention: weak instruments. As a result, we now view the 2SLS regressions, and Pitt's criticism thereof, as secondary. We rely more on direct analysis of the PK specification, which Pitt's corrections have made possible. We conclude that the quasi-experiment that is the basis for PK's claim to superiority, in addition to being impossible to detect, is not the source of identification. The hypothesis cannot be ruled out that borrowing is proxying for other nonlinear relationships between the controls and outcomes of interest.

Pitt did not give us an opportunity to review his first response (2011a) before making it public. Some preliminary, private back-and-forth might have caught secondary mistakes on both sides and economized on the time of readers. As it is, because of the public and emphatic tenor of Pitt's comments, we think it best to respond to them fully and publicly. This may spell tedium for the reader at times. For this reason, we have split our response. This paper comments directly on the criticisms in Pitt (2011a, 2011b). Separately, we explain the new insights about identification in PK.

Pitt (2011a)

Pitt's first response makes two major points and several minor ones.

On the handling of censored observations of log credit

One unusual feature of the PK estimator is that the equations in the first, instrumenting stage are Tobit. (In the second stage, outcomes are variously modeled as Tobit, probit, or uncensored linear.) The instrumented

variables are the logarithms of cumulative household microcredit borrowing over the previous 5–6 years, disaggregated by gender and lender, in Bangladeshi taka.

The appearance of these censored variables on the left of some equations and right of another creates a distinction whose importance we had not fully appreciated, between the *censoring threshold* and the *censoring value*—that is, between the threshold y_0 below which the latent linear index y^* is assumed to be censored and the value the observed y takes after censoring. For example, as Pitt (2011a) explains, the censoring threshold in PK is log 1,000 taka, the smallest observed level of log cumulative borrowing. Observations for non-borrowers take the value log 1 taka = 0. If log credit appeared only on the right of equations, then only the (observed) censoring value would matter; the data generating process for log credit would not be modeled, so the threshold would not be of interest. Conversely, if log credit appeared only on the left, as in a single-equation Tobit model, then the threshold would figure in the maximized likelihood but the censoring value would not. As Amemiya (1984) writes in his survey of Tobit models, “Note that the actual value of y when $y^* \leq y_0$ has no effect on the likelihood function. Therefore, the second line of [the line of the Tobit model, specifying the censoring] may be changed to the statement ‘if $y^* \leq y_0$ one merely observes that fact’.”

Pitt writes that there is “nothing unusual” about a Tobit model in which the censoring threshold and value differ. But it is unusual, at least in mattering for estimation. Pitt to the contrary, the PK model is not formally an example of Amemiya’s Standard Tobit model (equations (3) and (4) of Amemiya (1984)), which assumes the censoring threshold and value are same, at zero.¹ Certainly the PK model is an example in spirit, within Amemiya’s typology, but its formal exclusion from this authoritative and comprehensive survey speaks to its novelty. More to the point, the PK distinction between censoring threshold and value only matters in multistage models in which a preliminary stage is Tobit. Few researchers have fit such models. One reason is that it is not necessary for consistency to model instrumented, censored variables as censored. Standard linear treatment is consistent (Kelejian 1971) as well as robust to heteroskedasticity. Perhaps another reason such models are unusual is the historical lack of software (or familiarity therewith). Lee Lillard and Constantijn Panis’s stand-alone aML package can do the job, but appears not to be widely used among econometricians. Pitt wrote custom code in FORTRAN. Roodman wrote *cmp* for Stata in 2007, which brought such models to a broader community.

However rare, the distinction between censoring threshold and censoring value is sensible and valid. And Pitt is right that RM, in equating the two, are not faithful to PK.

The reason for this inaccuracy in the RM replication is not, as Pitt asserts, a software limitation.² *cmp* can handle the distinction, and can model censoring thresholds other than 0. The solution is to create two copies of each log credit variable, one for use on the left of the first stage, censored at log 1,000, the other for use on the right of the second stage, censored with log 1. (Storing log 1,000 in the left-side versions of the variables communicates the non-zero censoring threshold to *cmp*.) An example is posted on Roodman’s blog is at <http://j.mp/jweucd>.

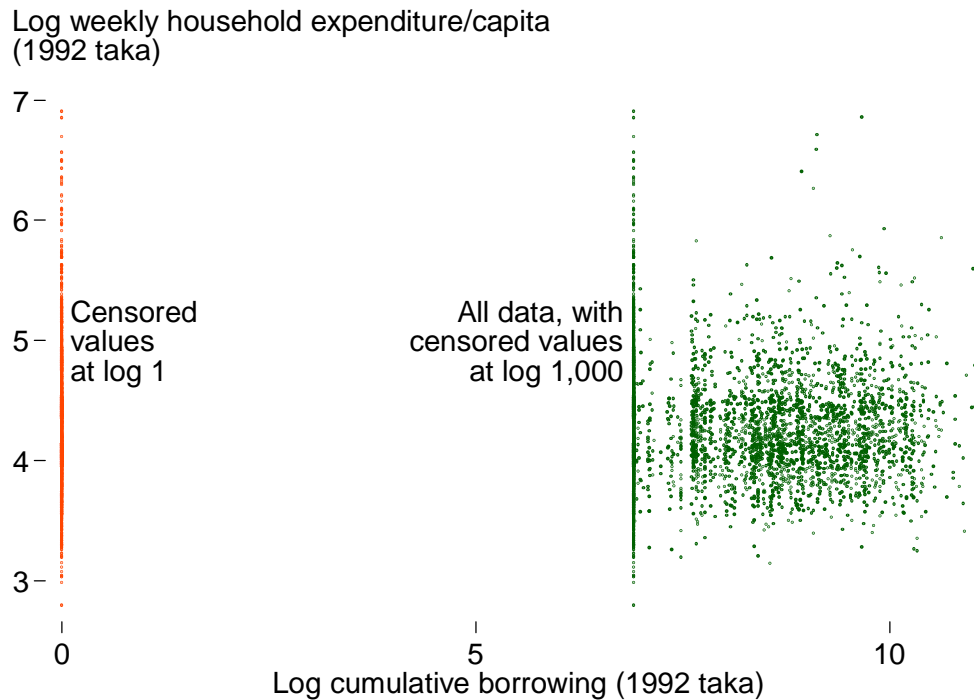
The real reason for inaccuracy, from our point of view, is a much deeper issue: PK’s choice of censoring value is arbitrary and undocumented. RM raise this issue but PK and Pitt (2011a) do not confront it. Not only is it a source of confusion in the replication. It also undermines PK’s interpretation of their results.

To see why, consider that PK could have censored with, instead of log 1, 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 the treatment variable, thus the lower the best-fit slope coefficient in a regression of household consumption per capita on the treatment. Figure 3 from RM, copied here as Figure 1, illustrates.

¹ “It is what Amemiya (1984) calls a Standard Tobit Model (Type 1 Tobit) in his classic survey paper of Tobit models.”

² “*cmp* does not correctly estimate models with a non-zero censoring threshold.”

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



One can see how a best fit-line to the data would have lower slope when non-borrowers are censored with log 1 instead of log 1,000. The effect can be seen in columns 3 and 4 of Table 1, where moving the censoring value from log 1,000 to log 1 reduces the coefficients on female borrowing by factors of 3–5. Thus PK’s conclusion that “annual household consumption expenditure increases 18 taka for every 100 additional taka borrowed by women...compared with 11 taka for men” means less than it seems.

The problem is that there is no natural choice of censoring value after taking logs for a censored variable that can be zero. One cannot censor the log of the variable with the log of 0. Fundamentally, taking logs of a regressor implies the assumption that the unlogged variable can never be 0. A hypothetical *ceteris paribus* movement from a zero to a non-zero state in the unlogged variable is thus outside the implied conceptual framework, and can only be hitched to it through an ad hoc assumption.

PK, we now understand, censor with log 1 (which might not have *seemed* arbitrary, being 0). What perhaps none of us has fully appreciated till now is that embedded in PK’s choice is an important *ceteris paribus* assumption: *moving from no treatment (proxied by 1 taka) to minimal observed treatment (a single one-year, 1,000-taka (\$25) loan taken and repaid sometime during the last five years) increases household consumption per capita by the same proportion as moving from 1,000 to 1,000,000 taka of treatment (about \$25,000).*³

That is a strong and debatable assumption.

It is undocumented too. The PK mathematical appendix presents a model in levels rather than logs, thus effectively avoiding the issue at hand. There, credit is censored at 1,000 with 0. What the 0 maps to in logs was only stated in Pitt (2011a).

³ The observed maximum for cumulative female borrowing is 64,850 taka.

For us, another barrier to understanding was an apparent impasse in our communication with Pitt. Our last query to him (e-mail from Roodman to Pitt, March 5, 2008) effectively raised the censoring issue, along with several others. Pitt's substantive reply, five days later, was a single sentence, which did not respond to the censoring question.

If the dialogue had continued, we might have avoided the impression that credit was censored at log 1,000. That last message to Pitt stated that in the data set he had recently sent (unlike in the one posted with Pitt (2011a)), the credit variables that appeared on the left side of the instrumenting equations were divided by 1,000 before being logged and censored at log 1, which was equivalent to censoring them at log 1,000. Pitt (2011a) denies ever censoring at log 1,000: "The credit variables that I sent are $\log(\text{credit}) - \log(1000)$ if $\text{credit} > 0$, and zero if $\text{credit} = 0$. In no case did a [censored] credit variable have the value $\log(1000)$." But what Pitt says PK didn't do and what Pitt says PK did do are mathematically identical. Dividing by 1,000, logging, and censoring with log 1 is the same as logging, subtracting log 1,000, and censoring with log 1. We see now from Pitt's (2011a) description that he never *conceived* of PK as censoring with log 1,000. But the data we received were open to that *interpretation*. This confusion illustrates the challenge of replicating a study with incomplete evidence on how it was carried out.

Sensing the structural issue, lacking understanding of PK's choices, perceiving precedent, and seeking minimal arbitrariness, RM use the censoring *threshold* as the censoring *value*: log 1,000. The structural assumption implied thereby is that moving from non-borrowing status (now represented by 1,000 taka) to minimal borrowing status (also 1,000 taka) has exactly zero impact. That assumption too, like PK's counterpart, is debatable.

But it seems less *unrealistic* than PK's. Taking a representative female credit coefficient of 0.04 from PK, their results imply that taking a single one-year loan of 1,000 taka some time during the previous five years raised household consumption by $e^{0.04(\log 1,000 - \log 1)} - 1 = 32\%$. Against an average-consumption benchmark of 84 taka/per week/per person for a typical household of five, that works out to an increase of some 7,000 taka per/year/household 0–5 years after taking the loan.

Yet the PK results turn on qualitatively on which structurally implied assumption about the relative impacts of commencing microcredit and increasing it is closer to truth. Compare columns 4 and 5 of Table 1: censoring with log 1—assuming massive relative impacts from starting microcredit—produces positive coefficients.; censoring with log 1,000—assuming no impacts from starting—destroys the results.

Thus Pitt's characterization of RM's choice—RM's "fatal econometric error"—misses this point: "**Roodman and Morduch arbitrarily assign 1000 units of treatment to the control group who were untreated.** [emphasis in original]." One could charge with equal validity that "Pitt and Khandker arbitrarily assign 1 unit of treatment to the control group." What must be examined is the structural assumption implied by a choice of censoring value. It is far from obvious that RM's choice is worse.

A wrong control variable

Pitt (2011a) points out two errors in RM's control set. RM wrongly include *censored*, a dummy for whether a household participated in a credit program but did not actually borrow.⁴ And they wrongly exclude *nontar*, a dummy for whether a household was too "well-off" to be a target for microcredit (whether or not a microcredit program operated in the village). The first of these charges has little significance for the results. (See column 3 of Table 1.) But the second is important. When this variable is restored *and* when log credit is censored

⁴ This was a poor choice of variable name on our part since the dummy is 1 for households that had joined a credit program but not actually borrowed, while it is 0 for households that were eligible, had a program in their village, and yet did not participate. From the point of view of the Tobit modeling, "borrowing" in both of these groups is censored, but the dummy is 1 only for the first.

with log 1 instead of log 1,000, the coefficients on female credit flip to positive (columns 4–6) and a close replication is at last achieved.

Pitt (2011a) describes RM’s substitution of *crcensored* for *nontar* as “inexplicable.” But of course it had a cause and so is explicable. Despite the impression created in Pitt (2011a), the data set he sent us in 2008 is not the one he posted in 2011. The 2008 copy lacked some variables, notably (and Pitt’s assertion to the contrary), the dummies for availability of microcredit by gender, which are required to define the samples of the first-stage equations.⁵ The 2008 data set also lacks informative variable names and labels, and includes no documentation of how variables were used in analysis. Its first 15 columns are labeled *xw1–xw15*. These are followed by *xm1–xm15* and *xb1–xb25*. (See j.mp/iYcgeg.) Some detective work confirmed that these groups are the right-side (**X**) variables for the female credit, male credit, and household consumption equations respectively. Pitt to the contrary, the 25 seeming regressors for the latter include *crcensored* (as *xb25*).⁶ Meanwhile the group of 25 excludes that crucial control missing from RM, *nontar*. *nontar* does however appear elsewhere in the file.

Further strengthening the natural hypothesis that these 25 represented the actual regressor set, they line up perfectly with variable lists in several regression results tables in the 1996 working paper version of PK (Pitt and Khandker 1996). (The PK paper in *JPE* does not list complete control sets in regression results, presumably to save space.) Pitt distances the working paper from the *JPE* publication: “That paper had a different title, presented different estimates, and used a different sub-sample of the data than PK.” Yet the two versions report identical sample sizes, identical means and standard deviations for independent and dependent variables, and identical results for all four household consumption specifications. For example, the *JPE* article’s Table 2 reports OLS household consumption results that perfectly match those in the working paper’s Tables A3 and A4. And the latter list *crcensored* (“participated but did not take credit”) as a control. Thus Pitt’s statement that *crcensored* “is not used in PK” appears incorrect, at least as regards the OLS regressions.

We agree, however, that it is best to drop this uninstrumented dummy because of potential endogeneity.

The foregoing also helps explain—but does not excuse—our omission of the crucial *nontar* control. Pitt is right that it does appear in text and in the list of independent variables (Table A1 in *JPE*, Table 3.1 in the working paper). But we had reason not to take that list as gospel since it was missing *crcensored*, a variable evidently used in the regressions. Meanwhile, *nontar* was missing from all control lists in the closely related working paper. (We appreciate now that it could not have been included in those lists because, as it happens, they are only for regressions restricted to target households, obviating any need for the *nontar* dummy.)

Again, our confusion highlight the value of transparency in econometrics. We never had access to the original code, which Pitt said is lost, and only obtained a data set that was cryptically labeled, undocumented, and misleading (unintentionally, we assume). Reconstructing a complex econometric study from fragmentary evidence about how it was done is an act of science in itself. Hypotheses are generated, then tested against the evidence. Sometimes different bits of evidence appear to contradict each other. No piece of evidence is unimpeachable because all are produced by fallible human beings. The hypothesis we chose about the control set is wrong but to some degree understandable. Our lack of full access to the PK code and data made our error more likely. Our complete sharing of our own data and code helped Pitt find it. If *JPE*’s policy requiring authors to post data and code has been in force in 1998, the story would have been different.

Other points

Pitt notes “two other less important issues affect the replication estimates.”

⁵ “They did not use the variable in the *PK estimation dataset* that I sent them.”

⁶ “I did not include this variable in the *PK estimation dataset* sent to Roodman.”

The first is that, “for a relatively small number of households, their separation of the sample into households with choice to borrow and households without choice to borrow, by gender, is wrong. They did not use the variable in the PK estimation dataset that I sent them.”

As mentioned, the 2008 data set does not actually contain the relevant dummies. This issue was also raised in the 2008 e-mail to Pitt that received a short reply. At any rate, Pitt does not substantiate the assertion that the differences between PK and RM in defining these variables are errors in RM.⁷ We attribute them to different choices in data cleaning. In a few cases, contradictory data forced somewhat arbitrary judgment calls. How does one interpret data showing female borrowing in villages without female credit groups? Zero out the borrowings? Assign them to husbands? Accept that the woman had the choice of microcredit after all?

The second minor point is that

RM use all three rounds of data in the estimation of the ‘first-round’ demand for credit equation, while PK use only one round of data...The `cmp` command does not allow there to be only one round of data for credit demand while at the same time having all three rounds for consumption. Khandker and I did so because we think of credit as a stock variable rather than a flow variable, and there is not a great deal of variation in the lifetime stock of credit between rounds that are only a few months apart.

As far as we know, this choice was not previously documented. But it is reasonable, even wise since it expunges the most recent borrowings, which are most likely to be endogenous. `cmp` can in fact handle this situation. Before running it, one creates copies of the regressor variables that stay constant at their first-round values through all three survey rounds.

It seems to us that for internal consistency, if the credit variables in the first stage are fixed in round 1 then so should those in the second stage. Otherwise the endogenous variables are not actually the ones instrumented. So in our run (penultimate column of Table 1), we fix all first-stage variables and all credit variables in the second stage to their round-1 values. Overall this seems to improve the replication a bit, at least for male borrowing.

⁷ The 2008 data set includes dummies *yf* and *ym* for whether females or males in a household actually borrowed and a dummy *choice* that is not differentiated by gender. See j.mp/iYcgeg.

Table 1. Incorporation of Pitt (2011a) changes Roodman and Morduch (2009) specification for household consumption per capita

	Changes to Roodman & Morduch (2009) specification, all using						PK original
	Pitt (2011a) data						
Drop <i>censored</i> ?	No	Yes	Yes	Yes	Yes	Yes	
Censor with log 1?	No	No	Yes	No	Yes	Yes	
Add <i>nontar</i> ?	No	No	No	Yes	Yes	Yes	
1st-stage & credit vars fixed?	No	No	No	No	No	Yes	
Log cumulative female borrowing, BRAC	-0.1230 (0.036)	-0.1253 (0.037)	-0.0287 (0.008)	-0.0846 (0.056)	0.0399 (4.318)***	0.0389 (3.987)***	0.0394 (4.237)***
Log cumulative female borrowing, BRDB	-0.1579 (0.046)	-0.1572 (0.047)	-0.0327 (0.009)	-0.1089 (0.068)	0.0407 (3.884)***	0.0407 (3.643)***	0.0402 (3.813)***
Log cumulative female borrowing, Grameen	-0.0998 (0.026)	-0.0980 (0.027)	-0.0299 (0.008)	-0.0632 (0.042)	0.0435 (4.320)***	0.0425 (4.032)***	0.0432 (4.249)***
Log cumulative male borrowing, BRAC	-0.0157 (0.046)	-0.0162 (0.047)	-0.0020 (0.013)	0.0270 (0.053)	0.0132 (0.730)	0.0156 (0.911)	0.0192 (1.593)
Log cumulative male borrowing, BRDB	-0.0013 (0.040)	-0.0003 (0.041)	-0.0012 (0.013)	0.0450 (0.050)	0.0168 (0.875)	0.0182 (1.024)	0.0233 (1.936)*
Log cumulative male borrowing, Grameen	-0.0157 (0.033)	-0.0145 (0.033)	-0.0035 (0.013)	0.0202 (0.040)	0.0110 (0.589)	0.0132 (0.755)	0.0179 (1.431)
Observations	5,218	5,218	5,218	5,218	5,218	5,218	5,218

Absolute z statistics clustered by village in parenthesis. *significant at 10%. ***significant at 1%.

Pitt (2011b)

In a preliminary blog response to Pitt, Roodman updated Table 4 of RM, which reports statistical tests of PK's identifying assumptions. The tests are based on 2SLS regressions that are constructed to parallel PK's preferred LIML household consumption regression, following Pitt (1999). Pitt (2011b) challenges these tests—not very convincingly, in our view.

But the tests are now secondary to our analysis. With Pitt's help, we have replicated PK much more accurately, and gained insights thereby, most of which we reserve for a separate paper. Meanwhile, we have come to appreciate that the 2SLS regressions probably have weak instruments, which invalidates the distributional assumptions underlying the estimator and the overidentification test. So our replies to Pitt (2011b) are brief.

The Hansen overidentification tests in RM

A version of RM Table 4 using Pitt's (2011a) data set is below, as Table 2. The bottom of the first column shows the Hansen overidentification test result for the 2SLS regression that parallels PK's preferred LIML specification. In this regression, the instruments are interactions between the dummies for female and male credit choice and all the controls, including village dummies. (See RM for an explanation.) The table shows that the regression fails the test with a p value of 0.018.

Dropped from this version of the RM table are Sargan tests, which we now see, agreeing with Pitt (2011b), are not theoretically valid.

Why were the Sargan tests in RM? RM note that PK's modeling assumes no heteroskedasticity. Since PK cluster errors by household they also assume error correlations only within households, not across them. They thus assume that errors are i.i.d. within each survey round. This seemed to open the way for a more high-powered overidentification test: performing the Sargan test on regressions run on one survey round's data at a time. The Sargan test is valid for spherical errors and more powerful than the Hansen test in that context. However, as Pitt (2011b) points out, and as we realized around the same time, the Sargan test is still not appropriate in this context because the data are weighted. Weighting observations effectively introduces heteroskedasticity.

In order to investigate which instruments might be invalid, RM modify the base 2SLS regression by including in the second stage a large subset of the instruments, those that are interactions between credit choice and village dummies. The newly included terms have joint explanatory power, as shown by the F test; and including them reduces the failure on the Hansen test. This suggests that these interaction terms are not excludable instruments. More concretely, this suggests that village "fixed effects" are in fact not fixed, but differ systematically for target and non-target households.

However, there are hints that most of the instruments in these regressions are weak, which, as mentioned, undermines the distributional assumption of the Hansen test. The regression in the first column of Table 2 actually fails an *underidentification* test.⁸ Aggregating the credit variables by gender (across lender) eliminates that problem, but still produces a Cragg-Donald F statistic of 3.543. Stock-Yogo critical values are not available for interpreting this statistic because the instrument count is so high. But the statistic itself is low, and paring back the instrument set produces regressions for which the critical values are available, and which indicate instrument weakness (results not shown).

⁸ The test is that of Kleibergen and Paap (2006), reported by `ivreg2` for Stata (Baum, Schaffer, and Stillman 2007).

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

Log cumulative female borrowings, BRAC	0.010 (0.610)	-0.036 (1.555)
Log cumulative female borrowings, BRDB	-0.025 (0.812)	-0.074 (0.997)
Log cumulative female borrowings, Grameen	0.009 (0.683)	0.038 (0.938)
Log cumulative male borrowings, BRAC	0.031 (1.291)	0.018 (0.336)
Log cumulative male borrowings, BRDB	0.002 (0.067)	0.049 (1.067)
Log cumulative male borrowings, Grameen	-0.002 (0.087)	-0.146 (1.417)
Interaction terms using village dummies (F test p value)		0.000
Hansen test of joint validity of instruments	0.018	0.157
Observations	5,218	5,218

Analogously with the PK LIML fixed effects regression, all regressions instrument with interactions of male and female credit choice dummies with household characteristics, survey round dummies, and village dummies. The second set includes the interactions with village dummies as controls. Absolute t statistics clustered by household in parenthesis.

Criticisms of the Sargan and Hansen tests

Pitt (2011b) does not mention weak instruments. It makes other points which we think are either incorrect or minor.

The first criticism is one already mentioned, that observation weights invalidate the Sargan tests. But with the Hansen test available to corroborate the Sargan, this is not important.

The other criticisms have even less bite. Mostly they question the first stage of the 2SLS regressions for lack of realism: the first stage doesn't model the deterministic zero-ness of microcredit for households without access to it; it doesn't constrain the instruments for female borrowing to have zero coefficients in the equation for male borrowing, and vice versa. But these equations are reduced forms and don't need complete realism for consistency.

To answer Pitt's demand for proof, the key assumptions necessary for the consistency of 2SLS (as a special case of linear GMM) are that: the structural equation must be linear; its error term must be orthogonal to the instruments (included and excluded); and the instruments must satisfy a rank condition (Hayashi (2000), 198–203, 217–18).⁹ Pitt does not question any of these assumptions for the case at hand. Nor do we think he could without questioning the identification assumptions in PK. The weakness in our argument is of course that the instruments appear weak; but that does not appear to be Pitt's point.

⁹ There are also technical conditions about the evolution of the data set as sample size goes to infinity.

The last methodological criticism is that the heteroskedasticity induced by observation weighting makes 2SLS inefficient. Pitt suggests “efficient GMM” instead, and reports that if one performs 100 GMM iterations, the p value on the Hansen test rises to 0.211. This Hansen test is, as far as we know, as valid as our two-step version, but not very reassuring. Taken at face value, it says that if the PK conclusions about the impact of microcredit on poverty are based on valid assumptions, then there is only a 1 in 5 chance that the Hansen statistic would be as large as it is.

To conclude, Pitt introduces certain interaction terms into the PK regression’s 2nd stage, as we do in the second column of Table 2—except that where we introduce actual instruments (interactions between credit choice and village dummies), Pitt introduces novel terms (interactions between landholdings and the core control set). The lack of significance for the latter does not have implications as clear as our finding that excluded instruments are significant when included.

Conclusions

Sift through the back and forth and a few important points emerge:

1. There are errors on both sides, many of which make little difference for the interpretation of the data.
2. As Pitt (2011a) says, two replication discrepancies mattered: RM censored with log 1,000 instead of log 1 and left out a control, the dummy for household eligibility for microcredit.
3. The issue of the censoring value for log credit makes a back-door entry to a problem pointed out in RM: the censoring value, and thus the apparent marginal impact of microcredit, is arbitrary.
4. The discrepancies in item 2 probably would not have happened if the *Journal of Political Economy*’s current data and code sharing policy had been in effect 1998. PK’s choices relating to censoring were particularly opaque. The definition of the control set was clear enough that we should have figured it out; yet here too the incompleteness of the evidence on what PK do made matters confusing. We welcome the *JPE* policy change and hope it will be widely emulated.

Although slow and awkward, the back and forth has improved our understanding of the work of Pitt and Khandker. For the first time, someone other than the original authors can replicate and examine the key regressions. Separately, we describe what we have learned from such an examination.

References

- Amemiya, Takeshi. 1984. Tobit Models: A Survey. *Journal of Econometrics* 24: 3–61.
- Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2007. Enhanced Routines for Instrumental Variables/Generalized Method of Moments Estimation and Testing. *Stata Journal* 7(4): 465–506.
- Kelejian, Harry H. 1971. Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables. *Journal of the American Statistical Association* 66(334): 373–74.
- Kleibergen, F., and R. Paap. 2006. Generalized Reduced Rank Tests Using the Singular-Value Decomposition. *Journal of Econometrics* 127: 97–126.
- Kelejian, Harry H. 1971. Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables. *Journal of the American Statistical Association* 66(334): 373–74.
- Morduch, Jonathan. 1998. Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh. Harvard University, Department of Economics. j.mp/kCd2na.
- 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/dLNItJ.
- Pitt, Mark M. 2011a. Response to Roodman and Morduch’s “The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence.” j.mp/j4x2xV.
- Pitt, Mark M. 2011b. Overidentification Tests and Causality: A Second Response to Roodman and Morduch. Draft. j.mp/l9O9mP.
- Pitt, Mark M., and Shahidur R. Khandker. 1996. Household and Intrahousehold Impact of the Grameen Bank and Similar Targeted Credit Programs in Bangladesh. World Bank Discussion Papers Number 320. go.worldbank.org/DTE18U2P30.
- 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.
- Roodman, David. 2011. “Response to Pitt’s Response to Roodman and Morduch’s Replication of..., etc.” David Roodman’s Microfinance Open Book Blog, March 31. j.mp/gwgo0g.
- Roodman, David, and Jonathan Morduch. 2009. The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence. Working Paper 174. Washington, DC: Center for Global Development.