WPS6273

Policy Research Working Paper 6273

Replicating Replication

Due Diligence in Roodman and Morduch's Replication of Pitt and Khandker (1998)

Mark M. Pitt
Shahidur R. Khandker

The World Bank
Development Research Group
Agriculture and Rural Development Team
November 2012

Policy Research Working Paper 6273

Abstract

"The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence," by David Roodman and Jonathan Morduch (2011) is the most recent of a sequence of papers and postings that seeks to refute the findings of the Pitt and Khandker (1998) article "The Impact of Group-Based Credit on Poor Households in Bangladesh: Does the Gender of Participants Matter?" that microcredit for women had significant, favorable effects on poverty reduction. In this paper the authors show that these latest Roodman and Morduch claims are based on seriously flawed econometric methods and theory and a lack of due diligence in formulating models and interpreting output from packaged software. On the basis of Roodman and Morduch's preferred two-stage least squares regression, an alternative calculation of the standard errors would lead one to conclude that the problem with Pitt and Khandker is that they underestimate the positive and statistically significant effect of women's credit on household consumption. As in their previous efforts, the methods of Roodman and Morduch are shown to bias the findings in the direction of rejecting the results of Pitt and Khandker. We also further examine two aspects of our instrumental variable approach that have been attacked by Roodman and Morduch. The first is the validity of the exclusion restrictions underlying the use of interactions between program choice and the set of exogenous variables (including the village fixed effects) as instruments. The second is the application of the "one-half acre" program eligibility rule. The authors show that identification does not require both of these, and present new results dropping each assumption in turn. The results originally reported in the Pitt and Khandker paper hold up extremely well in this new analysis.

This paper is a product of the Agriculture and Rural Development Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org.
The author may be contacted at skhandker@worldbank.org.

Replicating Replication: Due Diligence in Roodman and Morduch's Replication of Pitt and Khandker (1998)

Mark M. Pitt and Shahidur R. Khandker[1]

JEL code: C01; C21; C26; ARD.

[1] Brown University and World Bank. We thank Martin Ravallion, Will Martin, and Tiemen Woutersen for helpful comments and Kelley Smith for editorial suggestions. This paper and the code and data are available from http://www.brown.edu/research/projects/pitt. Views expressed in this paper are entirely those of authors and do not reflect in any way views of the World Bank or its affiliated organizations.

Replicating Replication: Due Diligence in Roodman and Morduch's Replication of Pitt and Khandker (1998)

Mark M. Pitt and Shahidur R. Khandker

1. Introduction

"The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence," by David Roodman and Jonathan Morduch (2011), hereafter referred to as RM, is the most recent of a sequence of papers and postings that seeks to refute the findings of Pitt and Khandker (PK) 1998, which found that microcredit for women had significant poverty-reducing impacts.

In the most recent iteration of their papers and web postings, RM claim that they have found three substantive and related weaknesses in PK: weak instruments, non-normality, and bimodality of the log-likelihood function. RM state that bimodality is "the most striking" finding, and implies that the PK estimates are inconsistent. They claim that this can be explained by "instrument weakness caused by disaggregating credit by gender" and that "non-normality appears to interact with the instrument weakness to generate the bimodality." They present test statistics to back up their claim of instrument weakness and non-normality, and suggest that instrument weakness and non-normality cause bimodality. They further suggest that the parameters estimated by PK lack validity based upon these three claims; additionally, using their preferred two-stage least squares (2SLS) linear LIML estimator, they fail to find the positive and statistically significant female credit effects found by PK, which they attribute to instrument weakness and non-normality. Like the earlier critiques by RM, we believe that these claims are based on flawed evidence.

This paper demonstrates that the tests of instrument weakness and non-normality presented by RM are invalid. As we will show, they have chosen testing procedures that necessarily tend to invalidly accept the hypothesis of weak instruments and reject the hypothesis of normal errors. Hidden from view in the paper are methods and results that reflect a serious lack of diligence in responding to and reporting errors that are plainly and repeatedly reported in the Stata output generated by the commands contained in the authors' Stata "do" files.

In the weak instrument case, RM "pad" the matrix of instruments they use for the test so that it is over 58 times larger than that actually used by PK. We show that this augmented matrix of instruments biases standard 2SLS coefficients in the direction of the least squares bias; that is, they underestimate women's credit effects. We provide simulation results that demonstrate that the PK model is very robust to variation in the weakness of the instruments, with instrument weakness measured using the methods of RM. The PK maximum likelihood approach recovers

the true parameters even when the instruments are quite "weak" and when the RM 2SLS approach demonstrates substantial bias. In addition, the linear LIML method that RM use generates estimates that are even less credible than standard two-stage least squares.

As for the test of non-normality, the skewness-kurtosis test for normality is the only parameter or test statistic that RM present in their paper that has not been adjusted for choice-based sampling. This failure to adjust for choice-based sampling results in over-rejection of normality when it is in fact true. No mention of this omission in the testing procedure is made in the paper.

RM do not investigate the econometric properties of the PK maximum likelihood when making their claims about bimodality and its causes and implications. We show below that, contrary to their implied claims, bimodality is a "standard" feature of this estimator even with strong instruments and normal errors, and provide evidence from a simulation exercise that the parameters identified at the global maximum mode are those of the true data generating process. As all agree that the PK results are those of the global maximum, there is no reason to suspect that PK have not recovered the true parameters.

RM also make incorrect claims concerning econometric theory. First, they wrongly believe that the bimodality of the PK log-likelihood implies that even when the model is properly specified (when errors are truly normal) the estimator is inconsistent, and that any claim otherwise is a violation of the theorem of the consistency of maximum likelihood. Second, RM maintain that the consistency of the maximum likelihood IV model requires normality, although it does not. Third, RM apparently believe that the recent literature on "weak instruments" applies to the PK model. There is no reason to think that it does, and indeed we present evidence that it does not.

There are other serious issues in the RM paper that we detail below. The lesson to be learned from this and the previous effort of RM (2009) is that their econometric methods are faulty and that their mistakes bias the findings in the direction of rejecting the results of PK. In the next section of this paper we briefly review the recent history of the RM replication attempts and our response to them. Section 3 addresses the supposed weak instrument problem and demonstrates that the RM two-stage setup used to generate tests of instrument weakness is wrong. Section 4 examines the faulty test of normality carried out by RM and shows that the PK model is robust to deviations of normality, just as econometric theory suggests. Section 5 examines the issue of bimodality, establishes that it is a "standard" feature of the PK model even with strong instruments, normal errors, and fairly large samples, reviews and cites the theorem on the consistency of maximum likelihood with respect to multi-modal likelihood functions, and demonstrates that the PK log-likelihood is not "flat" as RM suggest when appropriately graphed in the metric of probability. Section 6 addresses the issue of RM's lack of care in estimating their two-stage linear LIML models for which most t-values are exactly 0.01 in absolute value. Section 7 further examines two aspects of our instrumental variable approach that have been

attacked by RM. The first is the validity of the exclusion restrictions underlying the use of interactions between program choice and the set of exogenous variables (including the village fixed effects) as instruments. The second is the application of the "one-half acre" program eligibility rule. We will show that identification does not require both of these, and present new results dropping each assumption in turn. The results originally reported in the PK paper hold up extremely well in this new analysis. Section 8 summarizes our results.

2. Replication repeatedly redone

Although Roodman and Morduch's views on the validity of PK have never been published in a peer-reviewed journal, they have received broad exposure in media such as Roodman's "Microfinance Open Book Blog," tweets, Roodman's (2011) book *Due Diligence: An Impertinent Inquiry into Microfinance*, an extensive worldwide book tour, at conferences of development practitioners, and in testimony to the US Congress (posted on YouTube). The book and Roodman's subsequent book tour specifically target Pitt and Khandker, and conclude that "on current evidence, the best estimate of the average impact of microcredit on poverty is zero." His book recommends that microfinance institutions and donors "eschew any drive to extend credit to the poorest." This is a very strong statement that appears to arise in large part from Roodman's experience in replicating and seeking to refute PK.

Roodman and Morduch used their original replication paper (2009) to support the claim that the PK evidence is faulty, that the development community did not perform "due diligence" in evaluating the effectiveness of microfinance, and that microfinance has no discernible impact on the lives of the poor. Their latest revised version (2011) of that paper is just the most recent iteration in a continuing saga that dates back to Morduch's 1998 attempt at replication prior to the official publication of PK.

David Roodman's critical reviews on scholarly research cannot be taken lightly. His claim of a very high ranking among the world's economists is a prominent feature of his biographies. His book (Roodman 2011) bio says that "in 2011, Roodman ranked in the top 10 in the Research Papers in Economics (RePEc) list of young economists in the world." This is indeed remarkable given that his book bio also notes that "he has never taken a course in economics or statistics." [2] Furthermore, he is a member of the four-person Advisory Board of the *Replication Programme Advisory Group* of the International Initiative for Impact Evaluation (3ie).

[2] We note that the RePEc ranking is based in large part on file downloads, which include journal articles posted on RePEc, working papers, and software contributions. The vast majority of Roodman's RePEc downloads are software add-ons for *Stata* (xtabond2, in particular) and papers explaining how to use those add-ons. To illustrate how writing some computer code can affect the ranking, note that in the most recent ranking "Top 5% of Authors" of all ages based on downloads (May 2012), Roodman ranks 10th in the world, above most living Nobel Laureates including Sims, Prescott, Engle, Becker, Akerlof and Diamond. (http://ideas.repec.org/top/top.person.adownloads.html#pro120). As another example, we note that number 9 on this list is a Swiss professor of sociology whose major download is a piece of Stata code for formatting regression output for printing.

Although this paper addresses the critiques of Roodman and Morduch (2011), it is useful to review the recent history of Roodman and Morduch's replications. In short:

1. **RM (2009) found statistically significant and negative female credit effects in their 2009 version of the paper, the opposite of the PK findings**. This finding came about because of what can only be called a typographical error on their part. They included the variable named "*crcensored*" when they should have used "*nontar*" in a single line of their Stata code. That is it. This error was not the result of being confused by some complicated bit of mathematics, econometric modeling and programming, or by us not sharing computer code. The error that turned the signs of coefficients around was simply the lack of the word "*nontar*" and inclusion of the word "*crcensored*" in one line of their code.[3]

2. **RM claim a lack of sharing made their attempts at replication difficult and resulted in their errors.** Although there were certain details of the PK model that are not found in the PK paper, these details were all inconsequential in the RM replication. In reference to "sharing" it should be noted that RM had the PK data for at least 11 ½ years prior to their posting of the RM (2009), with Morduch being sent the data by us at least 9 months prior to the PK publication date of October 1998.[4] Morduch had apparently constructed all of the variables required for estimation for at least that long. Morduch, in the final version of his paper dated June 1998, lists all of the PK variables in the Appendix (most relevantly Appendix Table 1[5]) to his paper along with means and standard deviations. RM (2009) confirm that these variables are appropriately constructed, noting that "Our replication data set matches Morduch's original quite well, not surprisingly. " So both the raw data and the actual variables used in estimation have been in Morduch's possession since early 1998.[6]

3. **RM claim in their original replication that the signs of the credit effects do not matter.[7] They claim that the real issue is a lack of causality based upon test**

[3] The correct variable *nontar* is in the list of included exogenous variables in the 2009 RM paper, in PK (1998), in Morduch (1998), and in a 2008 email from Roodman asking Pitt for assistance. The incorrect variable *crcensored* is in none of these. Consequently, the results that RM presented were inconsistent with what their paper and emails to us said that they did.

[4] Roodman wrote the complete code for estimating the PK model by maximum likelihood prior to any contact between us, and has never suggested that there were any errors in his *cmp* code that affected its ability to estimate PK.

[5] This table is missing from the version available from Morduch's site at NYU, although it is included in identically dated versions available elsewhere.

[6] Roodman and Morduch did not make public the code that they used to convert the raw data into the variables used in estimation until December 2011 even though issues about the data and variables began almost three years previously. Their code, which was publically distributed two and one-half years after their paper was distributed, is in the form of a binary file written in Microsoft SQL format.

[7] Although they claim that "this is not about our inability to match the sign," the difference in sign is front and center in their only academic publication (of which we are aware) that discusses their replication, the chapter titled "Access

statistics that suggest that the PK instruments are invalid. This is the same claim made in the current version of their paper, and refuted below. However, in the 2009 version and in blog updates to that version, RM used different methods to argue this point. As in the current version of the paper, the methods used in the earlier version are faulty. The particular errors of those previous attempts to disprove causality in PK include invalid Sargan test statistics and wrong Hansen J-tests critical values, both as a consequence of their neglect of the implications of sample weights. Furthermore, the incorrect two-stage setup used by RM means that their Sargan and Hansen J-tests would be wrong even if there were no need to use sample weights. This error is discussed at length in this paper in the context of Kleibergen-Paap test statistics (Kleibergen and Paap, 2006), the test statistics adopted in this most recent iteration of RM, but it applies as well to Sargan and Hansen J tests.

3. Weak Instruments

The essential nature of the model as formulated in PK is quite simple, although we are unaware of any previous models of this sort in the literature. Some individuals are randomly offered treatment.[8] Those offered treatment choose a level of treatment c (credit), which can be zero. Other individuals are randomly not offered treatment, and their level of treatment is deterministically zero (c = ε = 0, where ε is an error term). The upper portion of the vector $\binom{c}{0}$ contains the level of treatment chosen by those offered treatment, and the bottom portion contains the non-stochastic zeros of those not offered treatment.[9] Analogously, the upper portion of the matrix $\binom{X}{0}$ contains a set of K exogenous variables (which may include dummy variables for place) that affect the level of treatment c among those with choice. The bottom portion of this matrix is zero, as is the bottom portion of the vector of errors $\binom{\varepsilon}{0}$, consistent with the deterministic nature of c=ε=0 of those without choice.

$$(1)\quad \binom{c}{0} = \binom{X}{0}\pi_m + \binom{\varepsilon}{0}$$

---

to Finance" authored by Morduch and Dean Karlan in the widely respected *Handbook of Development Economics* (Karlan and Morduch, 2009):

> "Roodman and Morduch (2009) attempt to find closure to the issue by returning to the data and rebuilding the analysis from scratch. They are unable to replicate results from Pitt and Khandker (1998) or Khandker (2005). In fact, their estimates carry the opposite sign. Rather than concluding that microcredit harms borrowers, however, they unearth a raft of identification issues which are not solved with panel data. Their revised analysis casts doubt on all of the findings from the related set of papers, including Morduch (1998)'s finding on consumption smoothing. "

[8] More generally, we require only that the offer of treatment be random conditional on a set of exogenous covariates that might include location fixed-effects, the assumption actually made in PK.

[9] Thus, elements of c can be stochastically or deterministically zero depending on if individuals are offered treatment.

An outcome of interest y, such as total household expenditure, is observed for both those offered treatment choice and those not offered treatment choice.

$$(2) \quad \begin{pmatrix} y \\ y \end{pmatrix} = \begin{pmatrix} x \\ x \end{pmatrix}\beta_x + \begin{pmatrix} c \\ 0 \end{pmatrix}\delta + \begin{pmatrix} v \\ v \end{pmatrix},$$

where $\begin{pmatrix} v \\ v \end{pmatrix}$ is a vector of errors. To be clear, all of the elements of X in the first term of the left hand side of equation (1) are identical to the elements of X in the upper sub-matrix of the first term on the left hand side of (2). To keep the notation simple, and because we need only to distinguish between observations that are deterministically zero and observations that are not, the lower sub-matrix of the first term on the left hand side of (2) is also called X because it contains the same variables, but the values of the variables (elements) in the lower sub-matrix correspond to a different set of individuals -- those individuals without credit program choice.

We are interested in estimating the treatment effect parameter δ. Least squares estimation of (2) yields biased estimates of the parameter of interest δ if the errors ε and ν covary. In PK, it is assumed (fairly innocuously) that the joint distribution of the errors is

$\begin{pmatrix} \varepsilon \\ v \end{pmatrix} \sim N[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 \end{pmatrix}]$, and that the covariance $\sigma_{\varepsilon v}$ may be nonzero.

The setup presented above is sufficient to provide a consistent instrumental variable estimate of $\hat{\delta}$. This model is identified even though there are no exclusion restrictions of the usual sort; that is, there is no set of exogenous variables Z that affects c but does not affect y conditional on c.[10] The model is identified by the deterministic lack of credit program choice for a subsample of the data. PK estimate the model by maximum likelihood. The lower sub-matrices in (1) are all zero and thus provide no information to help identify the first-stage parameters π_m, and hence are not in the portion of the likelihood identifying those parameters. The likelihood is presented in detail in an appendix to PK.

Roodman and Morduch Stata forgo ML in favor of the packaged 2SLS test statistics obtainable from Stata. The question is whether the model, given in its simplest form in (1) and (2), can actually be estimated by 2SLS packages without doing violence to the model's setup.[11] The requirements imposed by the packaged software (Stata) used by RM of particular concern are that (1) there must be a vector of identifying instruments Z with at least as many columns as there are endogenous variables, and (2) all the included exogenous variables in the second-stage equation must also appear in the first-stage.

---

[10] Note that adding exogenous variables $\begin{pmatrix} p \\ p \end{pmatrix}$ with associated parameter vector β_p to equation (2) does not aid identification.

[11] RM use the *ivreg2* package of *Stata* with the "liml" option that estimates linear 2SLS by limited information maximum likelihood. We will say more about this choice below; however, the "linear LIML" model estimated by *ivreg2* is essentially two-stage least squares and not the "LIML" estimated by iterating over a nonlinear log-likelihood until convergence as carried out in PK.

To see whether these requirements of the packaged software are consistent with the PK model, we re-write equation (1) to be consistent with the 2SLS format and where $c_0 = \begin{pmatrix} c \\ 0 \end{pmatrix}$, $Z = \begin{pmatrix} X \\ 0 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} X \\ X \end{pmatrix}$, and $e = \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix}$:

$$(3)\ c_0 = Z\pi_z + \mathbf{X}\pi_x + e$$

or

$$(3')\ \begin{pmatrix} c \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ 0 \end{pmatrix}\pi_m + \begin{pmatrix} X \\ X \end{pmatrix}\pi_x + \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix}$$

The first-stage regression equation (3) will yield $\hat{\pi}_x = 0.0$, which is sufficient to retain the property that C is deterministically zero ($\varepsilon=0$), the property that underlies identification, and $\pi_m=\pi_z$ will be consistently estimated. Two-stage least squares estimation carried out in this way is entirely consistent with PK, except that it includes the bottom sub-matrices containing all zero elements (recall that $\hat{\pi}_x = 0.0$). That is, it simply adds observations containing only zeros to the vector of dependent variable observations and to the matrix of independent variable observation. Adding all of these deterministically zero observations still recovers a consistent $\hat{\pi}_{m,}$ but not the correct parameter covariance matrix.

The model defined by equations (1) and (2) illustrates the source of identification in PK, but in practice the PK model extends this model by allowing for the two endogenous variables C_m and C_f (male and female credit), the effects of which vary for three different program providers for each gender. To start, consider only the gender disaggregation. As before, there is random assignment to credit choice. In the maximum likelihood approach of PK, the "first-stage" uses only the (informative) stochastic observations:

Male credit:

$$(4)\ \begin{pmatrix} c_{m0} \\ c_{mf} \end{pmatrix} = \begin{pmatrix} X \\ X \end{pmatrix}\pi_{mm} + \begin{pmatrix} \varepsilon_{m0} \\ \varepsilon_{mf} \end{pmatrix}$$

And, female credit:

$$(5)\ \begin{pmatrix} c_{fm} \\ c_{f0} \end{pmatrix} = \begin{pmatrix} X \\ X \end{pmatrix}\pi_{ff} + \begin{pmatrix} \varepsilon_{fm} \\ \varepsilon_{f0} \end{pmatrix}$$

Where c_{mo} refers to observations on the credit treatment choice of males where only males have choice, c_{mf} refers to observations on the credit treatment choice of males where both males and females have choice, c_{fm} refers to observations on the credit treatment choice of females where both females and males have choice, and c_{f0} refers to observations on the credit treatment choice of females where only females have choice. The second-stage with gender-specific credit is

$$(6)\quad \begin{pmatrix} y \\ y \\ y \\ y \end{pmatrix} = \begin{pmatrix} X \\ X \\ X \\ X \end{pmatrix} \beta_x + \begin{pmatrix} c_{m0} \\ c_{mf} \\ 0 \\ 0 \end{pmatrix} \delta_m + \begin{pmatrix} 0 \\ c_{fm} \\ c_{f0} \\ 0 \end{pmatrix} \delta_f + \begin{pmatrix} v \\ v \\ v \\ v \end{pmatrix}$$

In the packaged 2SLS estimation of the authors, the set-up for the first-stage equations is:

Male credit:

$$(7)\quad \begin{pmatrix} c_{m0} \\ c_{mf} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ X \\ 0 \\ 0 \end{pmatrix} \pi_{mm} + \begin{pmatrix} X \\ X \\ X \\ X \end{pmatrix} \pi_{mx} + \begin{pmatrix} 0 \\ X \\ X \\ 0 \end{pmatrix} \pi_{mf} + \begin{pmatrix} \varepsilon_{m,m0} \\ \varepsilon_{m,mf} \\ 0 \\ 0 \end{pmatrix}$$

$$(7')\ C_m = Z_m \pi_{mm} + \mathbf{X} \pi_{mx} + Z_f \pi_{mf} + \varepsilon_m$$

And, female credit:

$$(8)\quad \begin{pmatrix} 0 \\ c_{fm} \\ c_{f0} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ X \\ X \\ 0 \end{pmatrix} \pi_{ff} + \begin{pmatrix} X \\ X \\ X \\ X \end{pmatrix} \pi_{fx} + \begin{pmatrix} X \\ X \\ 0 \\ 0 \end{pmatrix} \pi_{fm} + \begin{pmatrix} 0 \\ \varepsilon_{f,mf} \\ \varepsilon_{f,f0} \\ 0 \end{pmatrix}$$

$$(8')\ C_f = Z_f \pi_{ff} + X \pi_{fx} + Z_m \pi_{fm} + \varepsilon_f$$

Note that $\pi_{mx} = \pi_{mf} = 0$ is required for deterministic zero C in equation (7), and $\pi_{fx} = \pi_{fm} = 0$ is required for deterministic zero C in equation (8). However, in general, least squares estimation will result in $\hat{\pi}_{mx} \neq 0$, $\hat{\pi}_{mf} \neq 0$ $\hat{\pi}_{fx} \neq 0$, and $\hat{\pi}_{fm} \neq 0$, as the three matrices of exogenous variables in each first-stage are not perfectly collinear. The consequence of this is that rather than having K parameters in the first-stage, there are 3K parameters. The additional instruments that have been added are necessarily "weak;" that is, they have individual t-statistics and group F-tests that are close to zero, if the data generating process set out in equations (1) and (2) is true. In particular, π_{mx} and π_{fx} , the gender-specific versions of parameter π_x of equation (3) – a parameter that would be estimated as identically zero in equation (3) – are now identified (that is, are nonzero) as a consequence of the artificial addition of the zero sub-matrices to the dependent and independent variable data matrices.

In the informative stochastic part of the sample, there is perfect collinearity between the first-two matrices on the right-hand side of equations (7) and (8). In equation (7), the matrix of

regressors $\begin{pmatrix} 0 \\ X \\ X \\ 0 \end{pmatrix}$ makes the effect of the exogenous variables on male credit depend on whether females have credit program choice, but then imposes that this interaction have an equal effect on the credit of males who have deterministically zero credit. Clearly, the values of the exogenous variables X cannot affect the level of credit participation when one has no choice to participate. Since the latter effect is ruled out by construction (the dependent variable is identically zero in the third sub-matrix from the top) and not by behavior, the estimated π_{mf} will be close to zero but not identically zero. The result of adding these two additional sets of variables to the first stage are twofold. First, the authors test for weak instruments and underidentification after arbitrarily adding necessarily weak instruments to the PK instrument set. Second, the zero predicted values of credit for those who deterministically have no choice to participate, a key identifying feature of the PK model, have been eliminated.[12] The identification tests presented by the authors are meaningless. Quite simply, RM identify parameters for the determinants of choice using "information" from the sub-sample that has no choice. The simple Stata ".do" file and regression output in Appendix A nicely demonstrate how the RM setup generates weak instruments by construction.

In practice, the problem of the authors arbitrarily adding a large number of weak instruments is even more serious than in the example above. In PK, not only are the two genders distinguished in the second-stage, but there are three different credit programs for each gender. As before, treatment choice by group is random, and the availability of the three programs is mutually exclusive and exhaustive. In PK, the three programs (Grameen Bank, BRAC, and BRDB-12) are allowed to differentially affect the outcome y, but the determinants of the level of treatment are assumed the same for each gender. Consequently, the first-stage equations in PK are still (4) and (5); however, the second-stage changes to

$$(9) \quad \begin{pmatrix} y \\ y \\ y \\ y \end{pmatrix} = \begin{pmatrix} x \\ x \\ X \\ X \end{pmatrix} \beta_x + \begin{pmatrix} c_{m0} \\ c_{mf} \\ 0 \\ 0 \end{pmatrix} G\delta_{mG} + \begin{pmatrix} 0 \\ c_{fm} \\ c_{f0} \\ 0 \end{pmatrix} G\delta_{fg} + \begin{pmatrix} v \\ v \\ v \\ v \end{pmatrix}$$

where the vector G, containing zeros and ones, picks out the appropriate credit provider (Grameen Bank, BRAC, or BRDB-12) from the three possibilities, and δ_{mG} and δ_{fg} are vectors of parameters of dimension 3 (g=1,2,3). In PK, this extension adds no additional endogenous variables to the second-stage equation and, as noted, the first-stage equations are exactly the same as the case in which we did not distinguish among credit programs by provider. Not so in

[12] Consequently, the residuals associated with deterministic credit are no longer zero. We leave zeros in the ε vectors to mark the observations that are deterministic even though in RM's econometric model they are no longer zero.

the RM paper, where there are 6 endogenous variables in the second-stage equation, and six first-stage equations. In RM, the first-stage equation for treatment of males by group 1 has the following form:

$$(10)\quad \begin{pmatrix} c_{m0,1} \\ 0 \\ 0 \\ \cdots \\ c_{mf,1} \\ 0 \\ 0 \\ \cdots \\ 0 \\ 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ X \\ X \\ \cdots \\ X \\ X \\ X \\ \cdots \\ 0 \\ 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix} \pi_{mm,1} + \begin{pmatrix} X \\ X \\ X \\ \cdots \\ X \\ X \\ X \\ \cdots \\ X \\ X \\ X \\ \cdots \\ X \end{pmatrix} \pi_{mx} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \cdots \\ X \\ X \\ X \\ \cdots \\ X \\ X \\ X \\ \cdots \\ 0 \end{pmatrix} \pi_{mf} + \begin{pmatrix} \varepsilon_{m0,1} \\ 0 \\ 0 \\ \cdots \\ \varepsilon_{mf,1} \\ 0 \\ 0 \\ \cdots \\ 0 \\ 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}$$

The vector of dependent variables is now partitioned into four parts corresponding to c_{m0}, c_{mf}, c_{f0}, and deterministic zero, reading from the top to the bottom. Each of the top three partitions is now disaggregated into the three groups g=1,2,3. For example, $c_{m0,1}$ is the treatment intensity of males when females are not offered treatment and when the treatment source is g=1. In the first-stage regression of the demand for male credit from group 1, the authors set male credit from other groups to zero, but make no change to the independent variables. Now the first matrix of independent variables, the Z_m set of instruments of equation (7'), is <u>not</u> zeroed out when men are treated by credit sources other than source 1. Consequently, the regression needs to estimate a $\hat{\pi}_{mm,1}$ that fits the data on male treatment from credit source 1, but also fits deterministic zero for the other male credit groups in the top-most partition. Forcing $\pi_{mm,1}$ to predict participation in the first group, $c_{m0,1}$ and deterministically zero credit in the other two groups clearly (where a $\hat{\pi}_{mm,1} = 0$ would fit best) makes no sense and necessarily results in a finding of "weak instruments."[13]

Compare the first-stage equation (10) estimated by RM with equation (4), the one actually estimated by PK and reproduced here:

[13] RM must be aware of the difference between their formulation and the PK model since they have more experience than anyone in estimating it by maximum likelihood, using the specialized software written by Roodman as well as by two-stage IV methods. Indeed, Pitt himself has repeatedly cautioned them about their two-stage least squares approach, beginning in his 1999 response to Morduch's 1998 critique, and has highlighted the issues associated with "padding" the matrix of instruments and the deterministic nature of credit program participation for those without program choice in the second of his papers (Pitt 2011b) responding to Roodman and Murdoch's 2009 draft of this paper. Indeed, in response to the 2SLS regressions in the first complete draft of Morduch's original replication paper dated February 10, 1998, six months <u>before</u> the PK paper was published in the *Journal of Political Economy*, Pitt wrote Morduch on February 17, 1998 saying "You should consider (empirically and otherwise) the implications of using hh's [households] with deterministically zero credit in a regression explaining credit." Here, over 14 years later, we still are making this same point in this paper.

$$(4)\ \begin{pmatrix} c_{m0} \\ c_{mf} \end{pmatrix} = \begin{pmatrix} X \\ X \end{pmatrix} \pi_{mm} + \begin{pmatrix} \varepsilon_{m0} \\ \varepsilon_{mf} \end{pmatrix}$$

PK estimate the male credit first-stage equation (4) using a matrix of independent variables that contains 54,534 elements; that is, 61 variables times 894 households with choice to join male credit programs. RM estimates the same first-stage equations using a matrix of independent variables that contains 58.6 times more elements than PK.[14] This wrongly expanded matrix consists of 204 independent variables times 5218 observations times 3 program types for a total of 3,193,416 elements. (The number of independent variables is only 204 in RM because Stata drops out 102 more independent variables as a result of perfect collinearity).

In the case of male participation in the Grameen Bank, the first-stage equation used by RM contains a matrix of independent variables for which fully 87.5 percent (931,872 out of 1,064,472) of the elements correspond to households in which males have (deterministically) no choice of participation in the Grameen Bank. In effect, RM's effort to examine the weakness (and other qualities) of the instruments used in PK uses a method that multiplies the size of the matrix of instrumental variables by a factor of 58 and, as a result, yields test statistics that have hardly any relevance to PK at all. It is no wonder that such methods allow RM to conclude that "The linear LIML estimates in Table 6, with their large standard errors, seem closer to the truth." (p. 32)

Issues arising from this artificial matrix remain even when RM aggregate credit effects across groups, resulting in only two first-stage equations, one for each gender. In this case the first-stage equation of the determinants of male credit program participation has 1,069,690 elements (205 independent variables times 5218 observations). This is 19.6 times the number of elements in the matrix of instrumental variables in PK.[15]

RM's claims fall short in another important way. There is no reason to believe that *any* of the literature on detecting weak instruments, and on the effect of weak instruments on parameter estimates, has any direct bearing on a PK-type model. The PK model is not a usual sort of instrumental variables model in which there is a set of identifying instruments associated with exclusion restrictions that meet the rank and order conditions for identification. Identification in PK clearly arises from having observations on the dependent variable y and the exogenous variables X for individuals that exogenously have no choice to obtain (credit)

[14] A relatively small part of this 58 times inflation is the result of RM replicating the data matrix for the first-stage into three rounds so that it fits into the single-command framework of the Stata *ivreg2* command.

[15] In an email David Roodman recently sent to Mark Pitt in response to an early undistributed version of this paper, Roodman maintains that the setup for two-stage estimation of PK is that given by equations (9) and (10) because it is the classical two-stage least squares setup. He argues that if test statistics show that the instruments in (10) are weak, then that means just what it says. In his email elucidation of this argument he fails to recognize the fact that credit is deterministically zero for a sizable part of the sample and that these deterministically zero "dependent variables" are the salient difference between the PK model and the classical two-stage least squares problem he is fixated on. Once again, if one tries to explain deterministic zero with a regression containing a set of independent variables, one gets nonsense.

treatment. Clearly, the number of observations without choice of treatment critically affects identification – at the limit, if there is only one such observation in the data, the model with both exogenous variables X and treatment c is not identified – and yet the authors proceed as if this were just another two-stage least squares problem. They seem to recognize there might be a problem, but then fail to properly address it:

> *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*. (p. 27)

If there is no statistical theory, then on what basis is there evidence?

Finally, one can easily simulate PK-like data sets and (i) estimate the model as PK do, by maximum likelihood, and (ii) estimate the model as RM do, using the wrongly augmented set of instruments and 2SLS.[16] The results of generating 200 such datasets are provided in Table 1.

Using the RM approach, the simulated data generating process has a positive bias for males and a negative bias for females[17]. Using PK's maximum likelihood method there is absolutely no evidence of bias arising from "weak instruments." The mean value for the F-statistic for the first-stage regressions given by equations (4) and (5) are 3.47 and 3.41, respectively. Using RM's 2SLS method, however, there is strong evidence of bias in the direction of the least squares bias. The male parameters are almost twice as large as the true values, and the female parameters are one-third too small, so that RM's method finds that male and female credit effects are about the same while the true female effect is three times as large as the true male credit effect. The Kleibergen-Papp rk rank test F statistic for weak instruments in the 2SLS has a mean value of 0.736, and the null hypothesis of "weak instruments" is never rejected in the 200 simulations. It is no surprise that the instruments are weak in RM's approach,

[16] RM actually estimate their model with linear LIML *not standard* two-stage least squares. This turns out to be another crucial error on their part that is discussed at length below. We use standard 2SLS here so as to highlight the effect of the weak instruments bias that arises as a consequence of RM's padded matrix of instrumental variables. The further degree of bias arising from the use of linear LIML is estimated with these same simulated datasets below and discussed at length.

[17] For the simulation, 40 variables X_i were generated from independent normal (0,1) random variables. The treatment chosen by females who are offered treatment is $c_f = 0.5\sum x + 12\mu + 8\varepsilon_f$, and the treatment chosen by males who are offered treatment is $c_m = 0.5\sum x - 12\mu + 8\varepsilon_m$, where μ , ε_f, and ε_m are independent normal (0,1) random variables. The outcome is $y = 0.75*c_f + 0.25*c_m + 0.5\sum x - 12\mu + 8\nu$, where ν is an independent normal (0,1) random variable. Female treatment was made available to 40 percent of the sample, male treatment to 40 percent, and no treatment to 40 percent. Half of the observations with female treatment also had male treatment, so when taking into account the overlap the percentages sum to 100 percent. Those treated were randomly assigned to one of three groups. Two hundred data sets with 5000 observations each were independently generated and the models estimated with maximum likelihood using Roodman's *cmp* estimation package, and the *ivreg2* (2SLS) package of *Stata*.

since they predict both behavior and deterministic zeros. All the 2SLS results of RM are the product of a flawed model that ultimately looks very little like the model of PK.[18]

4. Test of normality

Roodman and Morduch also claim that "non-normality appears to interact with the instrument weakness to generate bimodality" and that PK have an "incorrect likelihood function, as ε_o [the second-stage residual] is not normally distributed" and claim that "the PK structural error is not" normal based upon the skewness and kurtosis test for normality of D'Agostino, J. Belanger, and D'Agostino Jr. (1990) and implemented as part as the Stata command *sktest*. Just like their tests of instrument weakness, the normality tests that they carry out have errors that will necessarily lead to the over-rejection of the null hypothesis when it is true. That is, the authors have altered the PK model in such a way to make the instruments appear weak, and they have also calculated the test of normality in a way that makes the errors seem non-normal.

Their mistake is that they do not account for the sample weights in constructing the test statistic. This is odd, since they use sample weights in computing all other econometric estimates that they present in the paper, including in the table of means and standard deviations. So why is this not the case with the test of normality? The *sktest* in Stata, perhaps unique among the many Stata estimation commands that they use in the paper, does not allow for sample weights. Rather than go to the trouble of computing a sample weighted version of *sktest*, they proceed with computing unweighted test statistics, making no reference whatsoever in the text or table that sample weights were not used in only this peculiar case.

If the data sample arises from stratified random sampling (choice-based sampling in the PK case), then not correcting for the sample design will lead to over-rejection of tests of normality based upon skewness and kurtosis. Consider a sample r obtained by drawing N times from a normal distribution with zero mean and unit variance. Now sample (without replacement) a proportion $p<1$ of the values of r that are less than, say, 2.0, and "sample" a proportion $p=1$ all of the values of r having values greater than or equal to 2.0. Let q denote this stratified sample of r. An unweighted skewness/kurtosis test on q will necessarily over-reject the

[18] None of this means that the PK model cannot be estimated with two-stage least squares. Moreover, it should not be very difficult to accomplish. As shown above, however, one cannot expect to get it right by arbitrarily inserting a few million new data points into the procedure just so a two-stage model can be estimated with a single command from pre-packaged software. Estimating the first-stage equations (4) is just OLS. From those first-stage equations it is straightforward to predict credit for those with credit program choice, and make this prediction exactly zero for those without choice. The second-stage is just OLS substituting the predicted credit for the actual credit. All econometrics textbooks provide the formula for computing the standard errors for the case of predicted regressors, and these formulas can be appropriately adjusted for this case where some of the observations are exogenous in that they do not have predicted regressors. In every simulation do file that was appended to Pitt (1997), this two-step approach (using predicted credit for those with choice and exactly zero for those without choice) was the method used to illustrate PK. This was then followed in every instance with a check for the equivalence of 2SLS. The 2SLS was equivalent to this two-step method in every case where there was one gender and not equivalent in the case in which there were two genders, exactly as the simple algebra above predicts.

null hypothesis of normality even though the true process that generated the data is normal.[19] In a large sample such as PK's, in which sample proportions vary greatly across strata, it is to be expected that an inappropriate skewness/kurtosis test that did not apply sampling weights would reject the null hypothesis of normality.

To further demonstrate the inaccuracy of their approach, we have constructed a simulation of the PK model of the following sort. Thirty variables x1 through x30 were generated from independent normal (0,1) random variables. The treatment chosen by those who are offered treatment is $c = 0.40\sum x + 12\mu + 8\varepsilon$ where μ and ε are independent normal (0,1) random variables. The outcome $y = 1*c + 0.40\sum x + 12\mu + 8\nu$, where ν is an independent normal (0,1) random variable. Half of the sample of 30,000 was randomly assigned to treatment. There is only one gender and one program (credit) source. Eighty percent of the observations for which $c < 10$ were randomly dropped. One hundred data sets were generated and the model estimated with ML using Roodman's cmp estimation package with each data set, using sample weights. The mean value of the estimated parameters on c was 1.00 (true value = 1) with a standard deviation of 0.028. The predicted errors of the second stage were calculated and tested with the Stata *sktest* command, exactly as in the authors' paper. The median p-value of the skewness tests was 1.43e-06 and the median p-value for the combined skewness/kurtosis test was 8.49e-06. As the errors in this simulation exercise were normally distributed, these very low p-values rejecting the null hypothesis of normality are solely the result of not using sample weights in calculating the test statistics.

It is useful to consider whether non-normality of the errors should be a concern at all. In general, consistency of the maximum likelihood estimation of the multivariate regression model – also known as Seemingly Unrelated Regression (SUR) -- does not require that the errors actually be normal when that is the assumption made in constructing the log-likelihood (Davidson and MacKinnon 1993, p.315). In a simultaneous equations model estimated by LIML or FIML, such as PK, the same holds. The assumption that the errors terms are multivariate normal need not be true to insure that the parameter estimates are consistent and asymptotically normal (Davidson and MacKinnon, 1993, p. 641). Although it is well known that non-normality or heteroskedasticity yield biased estimates of Tobit models, the Tobit part of the PK model is the first-stage equations, and as is well known, consistent estimation of the first-stage equations is not necessary for consistency of the second-stage equations where the credit effects are estimated. Moreover, the RM tests of normality are only for the second-stage, which is absolutely continuous for the outcome of interest (household expenditure), and not for the Tobit

[19] The following Stata code demonstates this:

```
set obs 50000
gen x = rnormal()
sktest x
drop if uniform() < 0.9 & x < 2.0
sktest x
```

first-stages. Consequently, we can see no reason for the suggestion by RM that the actual distribution of the second-stage errors in PK has any important empirical ramification at all.

Even though RM's concern with bias arising from non-normality draws no support from econometric theory, we pursue the matter using simulation. This goes much further than RM's faulty skewness and kurtosis tests of the second-stage errors reported above. We allow for non-normality in both the first-stage and second-stage errors, and furthermore we specify two separate deviations from normality, one that generates skewness and one that generate excess kurtosis. The first-stage equations in every case are Tobits with a high proportion of censored values. Specifying a Tobit reveals how well the second-stage credit effects are estimated when the first-stage Tobit, which requires normality, is inconsistent.

The skew distribution was generated by summing a (non-negative) half-normal distribution and a normal distribution. Table 2 presents the results of generating 200 independent datasets with skew error distributions for both first-stage and second-stage equations. The median p-value for the tests for normality based on skewness is 3.49E-08, so skewness is quite significant. The RM 2SLS estimates perform poorly (although certainly as a consequence of the faulty construction of the RM instruments rather than because of skewness). The means of the maximum likelihood estimates of the credit effects using the PK log-likelihood are within 0.003 of the true value of 1.

The excess kurtosis was generated by adding a t(20) distribution to a normal. Table 3 presents the results of generating 200 independent datasets with error distributions characterized by excess kurtosis for both first-stage and second-stage equations. The median p-value for the tests for normality based on excess kurtosis is 7.64E-11, so excess kurtosis is quite significant. The RM 2SLS estimates perform poorly, again a likely result of the RM method. The means of the maximum likelihood estimates of the credit effects using the PK log-likelihood are within 0.005 of the true value of 1.

In short, econometric theory suggests that the PK maximum likelihood results are consistent when errors are non-normal, and the simulation exercises are in agreement.

5. Bimodal log-likelihood

RM claim to have discovered one symptom in the PK regressions "known to be associated in 2SLS and linear LIML with weak instrumentation: bimodality (p. 32)." Narrowing down the source of the problem, they conclude that it is explained by instrument weakness caused by disaggregating credit by gender (p. 35). The reasoning for this causal assertion is that, as described above, when they estimate their model by linear LIML they obtain Kleibergen-Papp rk rank test statistics suggesting weak instruments, although they fail to note that in calculating these test statistics they add scores of extraneous "instrumental variables" and thousands of deterministically zero observations to the first-stage matrix. In addition, as we show below, their use of linear LIML also results in an ill-conditioned parameter covariance matrix and outlandish

parameter estimates. Furthermore, they claim that "The non-normality appears to interact with the instrument weakness to generate the bimodality (p. 33)," although there is no demonstration in RM that this supposed interaction is in fact responsible for the bimodality. This assertion follows from their claim that "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. (p. 33)."[20] RM claim that the SUR label is appropriate one for the PK model, stating that "PK's first stage is a two-equation SUR system (complicated by Tobit censoring). (p. 33)" By that standard, any IV model with two or more first-stage equations is a SUR model and must, according to RM, have a globally concave log-likelihood function, asymptotically. The logic of RM is that the (i) PK model is a SUR model; (ii) SUR models have a globally concave log-likelihood as the sample size increases; (iii) the sample size in PK is large; so that (iv) something is wrong in PK since the log-likelihood is not concave; and (v) the source of the problem must be some interaction between instrument weakness and non-normality since there are test statistics that seem to suggest that both characteristics are true.

First of all, the assertion by RM in the paper and in Roodman's blog that the consistency of maximum likelihood requires that the log-likelihood function have a single mode is just wrong.[21] In theorem 2.5 of the *Handbook of Econometrics*, Newey and McFadden (1994, p. 2131) set out the conditions for the consistency of maximum likelihood, and having a single mode of the log-likelihood function is not one of the necessary conditions. Newey and McFadden discuss the example of maximum likelihood estimation of the parameters of the Cauchy distribution (in their example 1.1), which is consistently estimated by maximum likelihood even though it is multi-modal for small values of N, and note that in the multi-modal likelihood case in general "consistency results only apply to the global maximum (p.2117)." [22] As there does not seem to be any dispute that the PK parameter estimates correspond to the global maximum, there is no consistency issue that arises with the PK estimates if the likelihood were bimodal. Furthermore, the parameters associated with the local mode identified by RM are neither consistent nor asymptotically normal.

The log-likelihood of the PK model with the PK data is, in fact, bimodal. RM wrongly assert that the bimodality in the PK likelihood means that the PK results are not credible as a consequence of the twin evils of weak instruments and non-normality, and do so without examining the attributes of the PK likelihood.

It might well involve some effort to formally work out the properties of the PK likelihood. It is relatively simple, however, to simulate data sets that have normally distributed

[20] In contrast to this claim by RM, the consistency of SUR estimated by maximum likelihood does not require that the errors actually be normal when that is the assumption made in constructing the log-likelihood (Davidson and MacKinnon 1993, p.315).

[21] The claim on the Roodman blog is that "a foundational theoretical finding is that if the model being fit is correct, the probability of there being more than one mountain peak [mode] goes to 0 as the number of data points increases." (http://blogs.cgdev.org/open_book/2011/12/bimodality-in-the-wild-latest-on-pitt-khandker.php)

[22] On estimation of the multi-modal Cauchy distribution, see Ferguson (1978).

errors and check if (i) the log-likelihood is bimodal, and (ii) the frequency with which the maximum of the log-likelihood corresponds to the true value of the parameter of interest -- the credit effect, δ. The results of this simulation (Table 4) demonstrate that (i) bimodality was found in all 100 simulations, and is likely to be a "common" feature of PK-type models; (ii) all 100 simulations used data that were constructed with normal errors, so that non-normality is not an implication of the finding of bimodality, as RM suggest; (iii) the global maximum was tightly centered around the true value of δ, so that bimodality is innocuous in the simulations because the true estimate is recovered in every case; (iv) "weak" instruments are not necessary for bimodality, as RM suggest, as the instruments in the simulations were sufficiently "strong" to identify δ with tight bounds; and (v) the inferior mode was always in the direction of the least squares bias, that is, the mode had a smaller credit effect δ. In no case did the mode associated with the smaller value of δ have a higher likelihood. RM do not dispute that the PK estimates are the global maximum. Although the data-generating process of the PK data is, of course, unknown, this is suggestive evidence that the global maximum in PK, the one with a significant and positive credit effect, is the "true" maximum.

The simulated data-generating process has two genders and one program. Forty variables X_i were generated from independent normal (0,1) random variables. The treatment chosen by those who are offered treatment is $c_f = 0.5\sum x - 7\mu + 5\varepsilon_f$ for females, and $c_m = 0.5\sum x - 7\mu + 5\varepsilon_m$ for males, where μ , ε_f and ε_m, are independent normal (0,1) random variables. The outcome is $y = 1.0*c_f + 1.0*c_m + 0.5\sum x + 7\mu + 5\nu$, where ν is an independent normal (0,1) random variable. Half of the generated sample is female and half male. Treatment was made available to half of each gender-defined sub-sample. One hundred data sets with 5000 observations each were independently generated and the models estimated with maximum likelihood using Roodman's *cmp* estimation package to trace the converged likelihood over a range of the credit effect parameters $\delta = \delta_f = \delta_m$ starting at -0.3 and incrementing by 0.1 until $\delta_f = 1.5$. The gender breakdown is not really relevant in this exercise, since the point is to examine the shape of the log-likelihood surface with respect to δ, the parameter on credit.

In Table 4, as noted, the log-likelihood with respect to δ is bimodal in every single simulated dataset. The global maximum was tightly clustered around the true value of $\delta = 1$, and the local maximum (inferior mode) was clustered around $\delta = -0.1$.[23]

Figure 1 presents the log-likelihood surface for the 1st, 25th, 75th, and 100th simulated data set. The resemblance to the PK log-likelihood surface presented by RM (see also our Figure 2) is striking. One attribute of a bimodal log-likelihood is that selectively dropping observations can "cut off" the mode corresponding to the global maximum, resulting in a discrete (and negative, in this case) jump to the local mode. Indeed, it was not difficult to find data-generating

[23] When iterated to convergence, the mean value of the estimated δ was 0.998 with a standard deviation of 0.032, and the asymptotic t-ratios on δ were in excess of 30.0 in all simulated datasets. The data-generating process thus seems to result in a very well-identified model; that is, the instruments are not weak.

processes of the type outlined above (having, for example, normally distributed errors) for which the judicious dropping of less than 1 percent of the sample occasionally resulted in the local mode of the complete sample becoming the global maximum of the selected sample. These few observations may be deemed "outliers," but at least in the simulated case, dropping them from the sample generates incorrect inference about the underlying parameters of the true data generating process.

Bimodal log-likelihood surfaces were easy to obtain – one was found on the first try. Some limited experimentation found that generating datasets with 50,000 observations (instead of the 5,000 used in the simulations) from the same data-generating process still resulted in bimodal likelihoods. If the explanatory power of the instruments in the first-stage is made sufficiently large, bimodality seems to disappear. This occurred when the t-values for the credit parameter δ were pushed to exceed 42.0 simply by increasing the value of the parameters on the instruments in the first-stage so that average t-ratio for the 40 (independent) instruments in the first-stage was around 5.0.

What are the implications of all of this for the PK results, then? Are the instruments weak and the errors non-normal, and does it matter? The "foundational theoretical findings" of RM claiming a lack of consistency for maximum likelihood models with multiple modes are wrong; moreover, maximum likelihood of the broad class of IV models to which PK belongs is consistent in the presence of non-normal errors. Therefore, even if the errors are not normal (and we think it is not too hard to reject normality in large micro datasets such as this), this attribute is neither necessary nor sufficient to bias the results, a proposition supported by our simulation results. RM's "solution" to the presumed issue of non-normality, linear two-stage least squares by LIML (which they term *linear LIML*) regression, which they hold out as the "robust" estimate, is, as we will show, like their test statistics, an approach with a built-in bias to rejecting the PK results as a consequence of the artificial and unjustifiable inflation of the instrumental variable data matrix, making the instruments weak, along with the inclusion of observations with deterministically zero credit program participation in the first-stage.

How important is bimodality for the PK estimates? As we have shown, bimodality is not necessarily a perverse occurrence with the PK likelihood, and in the simulations, at least, the global maximum was always at the true value. What about RM's claim (p. 32) that the difference in the log-likelihoods between the modes "is slight," that the "difference probably does not deserve much weight" and "is better viewed not as two peaks but as a single, wide one that straddles 0 and implies that the coefficient is estimated with great imprecision"? It is difficult to reconcile the statement that the female credit coefficient is estimated with "great imprecision" with the asymptotic t-ratios that PK report (4.24, 4.25, and 3.81 for BRAC, Grameen Bank, and BRDB, respectively). For the PK estimates, the log-likelihood was concave around the global maximum and the t-values were computed as clustered (by household), Huber-White heteroskedasticity-robust standard errors. As the parameter standard error reflects the

slope and curvature of the log-likelihood around the maximum, how are such small standard errors consistent with the claim by RM that the log-likelihood is almost flat?

The top panel of Figure 2 presents the log-likelihood surface based upon the PK version of the dataset and restricts credit effects to vary by gender but not by group.[24] If the PK likelihood were really a true likelihood and not a pseudo-likelihood, about which more will be said below, a test of the hypothesis that any value of δ is equal to the maximum likelihood value of δ is given by twice the difference in their respective log-likelihoods and is distributed as χ^2 with one degree of freedom. It is straightforward to calculate the probability associated with this χ^2 statistic. This surface would then represent the probability that any parameter δ (on the horizontal axis) is equal to the value of δ at the global maximum (0.0437). The bottom panel of Figure 2 is the probability surface corresponding directly to the log–likelihood surface in the top panel. The local mode is barely visible in the bottom panel of Figure 2, as it is situated at very low levels of probability. Figure 3 is a close-up that focuses in on the area around the local mode so that one can read the probabilities associated with this mode off of the figure. The local mode corresponds to a probability less than 0.0002.

The complication is that the objective function that PK maximize is not a log-likelihood but a *pseudo*-log-likelihood, as a consequence of sampling weights, and the "two times the difference in the log-likelihood is χ^2 " rule does not apply to pseudo-log-likelihoods. The "true" probability surface can be below or above the one drawn in the figure. One simple way to see if the probability surface in the figure is too high or too low is to compare the probability of δ=0 from the clustered Huber-White parameter covariance matrix with the probability obtained from two times the differences in the pseudo-likelihoods. The p-value from the t-statistic is 0.0000013 (corresponding to a t-ratio of 4.83 or a $\chi2(1) = t^2 = 23.37$. The p-value from the pseudo-likelihood is 0.0000030, which is higher (by more than 2-fold) than the p-value from the t-ratio,[25] so that the probability surface pictured in Figure 3 and the bottom panel of Figure 2 is likely too <u>high</u> – that is, the probability value associated with the local mode is actually <u>smaller</u> than pictured in the figure and is likely less than 0.0001. The bottom line is that bimodality does not appear to be an issue in PK as (i) there is plenty of curvature in the estimated PK log-likelihood function, (ii) the log-likelihood surface with respect to the single parameter corresponding to the women's credit effect is not flat by the relevant probability metric, (iii) the probabilities associated with the local mode are extremely small, and (iv) simulation results suggest that the global mode corresponds to the parameters of the true data generating process.[26]

[24] Differences in credit effects by group are inconsequentially small in PK. To remain consistent with the estimates of Pitt (2011a) and (2011b), we use all three rounds of credit data in estimating the first-stage equation.

[25] This p-value corresponds to $\chi2(1) = 21.84$.

[26] RM report on a bootstrapping exercise in which they draw whole villages and then count the number of negative regression parameters. No implications can be drawn from such a procedure as it lacks any econometric justification, nor do they provide any.

6. Due diligence

Leaving aside their errors in logic – for example, not using choice-based sampling weights in calculating tests of normality – there are serious issues in the RM paper that arise from an apparent lack of attention to the methods and output that RM must have obtained. The former type of error – for example, the unweighted tests of normality – is not directly revealed by the computer output. When Roodman and Morduch computed the skewness/kurtosis test of normality with the Stata *sktest* command without using sampling weights, the software did not offer a warning that sampling weights should be used. However, things were very different when they estimated their "preferred" linear LIML estimates (their Table 6) of the effects of credit on a measure of household consumption. These estimates form the basis for their tests of instrument weakness and are used to support their contention that credit effects are zero. Re-running the estimates of RM's Table 6 using the data and code provided by Roodman and Morduch results in error messages and computer output that is the equivalent of sirens blaring and red lights flashing to proclaim that something is terribly wrong with the estimation. After each first-stage equation estimated as part of the their *ivreg2* 2SLS command with the LIML option, Stata reports:

> "*Warning: estimated covariance matrix of moment conditions not of full rank. model tests should be interpreted with caution*"

More worrisome than that is that 105 out of 110 slope coefficients in the second-stage output as reported by Stata have a reported t-ratio of exactly 0.01 in absolute value.[27] Regression output like this is an unambiguous and well-known signal something is very wrong. Nonetheless, RM overlook this, and report in their Table 6 the parameters, t-statistics,[28] and Kleibergen-Papp rk rank test statistics exactly as they are reported by Stata. Columns (2) and (3) of Table 5 provide all of the second-stage regression coefficients and t-statistics exactly as generated by Stata's *ivreg2* command – just copied and pasted into the table.[29] In addition, while a negative R-squared is possible with 2SLS, it is not conceivable that the (centered) R-squared is -381.4, so that the residual sum of squares is 382 times larger than the sum of squares, as the Stata output reports and which is presented at the bottom of Table 5.[30]

There is other evidence that RM's estimates of their Table 6 column (1) are suspect. The last two columns of our Table 5 provide standard two-stage least squares (as opposed to linear LIML) estimates of the same model. These are from the same data and exactly the same set of variables, the only difference being that RM used the '*liml*' option on i*vreg2*, and these standard

[27] Stata reports two digits after the decimal point when reporting t-ratios. The remaining five t-ratios are reported as 0.00, 0.02, 0.03, and 0.04, 0.25 in absolute value.

[28] RM report three digits after the decimal point so that five of them are 0.012 and one is 0.013.

[29] Ordinarily, one would not show all of the village fixed effects and other "nuisance" parameters in a table, but they are important here even though the names of the parameters are not. They are all presented in the table so that the reader can get the full visual impact of 105 coefficients with an absolute t-statistic of 0.01.

[30] Note as well that the value of the F-test of all of the parameters is F(110,1756)=1.85 in the linear LIML first-stage equation but F(110,1756)=14.88 for the exact same model estimated by standard 2SLS (our Table 5, bottom).

2SLS estimates leave out that option. Exactly what LIML means in two-stage least squares estimation is described in detail below. A comparison of the linear LIML estimates of columns (2) and (3) and the standard 2SLS estimates in columns (3) and (4) suggest that they have no relationship to each other at all. For example, the coefficient on the variable *nontarpk* in the 2SLS is 0.146 with a t-statistic of 3.39, but in the linear LIML it is -2.27 with (of course) a t-statistic of -0.01. RM estimated this model by standard 2SLS and posted it online as part of their response to Pitt (2011a), titled "Response to Pitt's Response to Roodman and Morduch's Replication of…, etc." (available at http://blogs.cgdev.org/open_book/2011/03/response-to-pitts-response-to-roodman-and-morduchs-replication-of-etc.php), and thus should be aware of this discrepancy in the t-statistics. Our Table 6 provides their 2SLS from that March 2011 response to Pitt as well as the linear LIML estimates that appear in this paper, which was distributed as a working paper in December 2011.

Consider as well the second column of parameter estimates in Table 6 of RM. Stata once again warns that the estimated parameter covariance matrix is not of full rank, and 105 out of 106 t-values in the second-stage are less than 0.5 in absolute value.[31] The Stata output of coefficient and t-statistics are copied and pasted into our Table 7. The (centered) R-squared of this regression is -7.26. RM once again ignore this evidence of serious issues with the estimates, and proceed to report parameters, t-ratios, and Kleibergen-Paap underidentification and weak identification tests.

The likely source of the clearly unidentified linear LIML models that RM present in their Table 6 is the extraordinarily artificial nature of the matrix of exogenous variables that they construct for the first-stage equations. To see why this matters, consider the nature of the linear LIML estimator. Leaving aside both sample weights and clustering so as to leave the notation uncluttered, the κ-class estimator of the vector of second-stage parameters $\mathbf{b} = [\boldsymbol{\beta} \;\; \delta]$ is

$$\mathbf{b} = \{\mathbf{X}'(\mathbf{I} - \kappa\mathbf{M_Z})^{-1}\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{I} - \kappa\mathbf{M_Z})^{-1}\mathbf{y}$$

where $\mathbf{Z}$ is the full set of exogenous regressors in the first-stage, $\mathbf{X}$ is the set of included exogenous regressors $\mathbf{X_1}$ plus the endogenous credit variables $\mathbf{C}$, $\mathbf{X} = [\mathbf{X_1} \; \mathbf{C}]$, in the second-stage, and $\mathbf{M_Z} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. The linear LIML estimator sets κ equal to the minimum eigenvalue of $(\mathbf{C}'\mathbf{M_Z}\mathbf{C})^{-1/2}\mathbf{C}'\mathbf{M_{X1}}\mathbf{C}'\mathbf{M_{X1}}\mathbf{C}(\mathbf{C}'\mathbf{M_Z}\mathbf{C})^{-1/2}$, where $\mathbf{M_{x1}} = \mathbf{I} - \mathbf{X_1}(\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'$. Exactly what this minimum eigenvalue represents in the RM formulation of the PK model is unclear because (i) the X and Z matrices have been by swelled by a factor of 58; (ii) most of the observations for $\mathbf{C}$ are deterministically zero but not treated as such; and (iii) 102 exogenous variables are dropped by Stata because of perfect collinearity. Which variables are kept by Stata and which are dropped is likely to be arbitrary, at least from point of view of statistical theory, so that dropping some columns from $\mathbf{X_1}$ rather than from $\mathbf{X_2}$ ($\mathbf{Z} = [\mathbf{X_1} \;\; \mathbf{X_2}]$), the matrix of excluded exogenous variables, may arbitrarily alter the mininum eigenvalue.

[31] The only exception to this invariance of the t-ratios is for a survey round dummy variable with a t-ratio of 6.2.

Even if the minimum eigenvalue used by RM in computing the value of κ lacks the statistical meaning to make it the linear LIML estimator, it can be argued that it does not matter much since the RM estimates can be interpreted as resulting from a *κ-class* estimator as in Nagar (1959) and others. Many methods of choosing κ have been presented in the literature. As the discussion above makes clear, linear LIML and other κ-class estimators are not the result of iterating on a log-likelihood function until convergence, but rather the result of an algebraic expression with an analytic solution, including the investigators' choice of κ. However, RM's assertion that "LIML is more robust to weak instruments" may not hold for choices of κ that are not LIML.

The LIML (κ-class) covariance matrix is computed as

$$\text{Var}\mathbf{(b)} = s^2\mathbf{\{X'(I - \kappa M_Z)^{-1}X\}^{-1}} ,$$

where s^2 is the estimated mean squared error, and differs from the standard two-stage least covariance matrix by the eigenvalue κ. However, the usual 2SLS covariance matrix,

$$\text{Var}\mathbf{(b)} = s^2\mathbf{\{X'(I - M_Z)^{-1}X\}^{-1}}$$

which sets κ=1, is still valid. In his popular econometrics textbook, Greene (2012, p. 567) notes that "A useful result is that the asymptotic variance of the two-stage least squares (2SLS) estimator, which is yet simpler to compute, is the same as that of the LIML estimator." Imbens (2010, p.17) adds that "Under the standard, fixed number of instrument asymptotics, the asymptotic variance for LIML is identical to that for TSLS, and so in principle we can use the same [2SLS variance] estimator." [32]

What happens if we compute t-ratios for the RM linear LIML model of RM's Table 6, column (2), in which there are separate gender effects but no disaggregation by credit group, using the perfectly valid 2SLS parameter covariance? Certainly, this is an appropriate exercise because of the problem of ill-conditioned covariance and data matrices, and the large and negative R-squareds. The statistically valid 2SLS t-ratios presented in our Table 7, when combined with Roodman and Morduch's own parameter estimates, suggest that there are very significant and positive female credit effects (t=10.63) and very small and insignificant male credit effects (t=0.05). On the basis of this 2SLS regression estimated by Roodman and Morduch, one might conclude that the problem with Pitt and Khandker (1998) is that they underestimate the positive effect of women's credit on household consumption. The point estimate of the female credit effect found by RM is vastly larger than that found by PK -- larger

[32] These two references are added to refute the contention of Roodman that the 2SLS covariance matrix (κ=1) is inappropriate in this case, a view he stated in an emailed response to an early draft of this paper. He argues that if the 2SLS covariance matrix was valid for linear LIML, which he (wrongly) doubts, then one should also able to use the OLS covariance matrix since that is also a κ-class estimator (with κ = 0). The error in his thinking is (1) the asymptotic variance for 2SLS is identical to that for linear LIML, and (2) for LIML, κ is always greater than 1 in finite samples, so that the κ=0 requirement for OLS is not feasible, and κ-class estimators are consistent whenever κ tends to 1 asymptotically at a rate faster than one over the square root of the sample size.

by a factor of 10. Of course, as we have shown above, the RM estimates are not interpretable, as they do not correspond either to the PK model or to any formal econometric model of program evaluation that RM or anyone else has ever set out, not to mention that the linear LIML estimation strategy RM adopt just makes the estimates even more strange.

Can one show that, in general, the linear LIML does not "work" in the PK model as RM have formulated it, or is it simply a reflection of some defect in the PK data? In our Table 8 below, we reproduce the simulation results of our Table 1 (above) but add the linear LIML estimates to the set of estimates reported. These are exactly the same simulation datasets described earlier. Recall that the errors are normally distributed and the instruments are not "weak" in these simulations. The linear LIML estimates of column (4) of our Table 8, the only new column in this table as compared to Table 1, have clearly bizarre point estimates (means of 200 simulations are presented) and huge standard deviations. It is no surprise that RM get outlandish parameter values and large standard errors in their Table 6 when they estimate the PK model with their linear LIML setup. Our Table 9 reports the estimates from one of the simulation datasets. The model estimates an effect of female credit from group 1 to be -244.388 when the true value is 0.75, and the male credit effect of group 3 to be 177.334 when the true value is 0.25. This should not be surprising given the large standard deviations reported in column (4) of our Table 9. Almost anything can result when an erroneous econometric model is applied to an artificially altered data matrix.

7. Sources of identification in Pitt and Khandker: Some new results

In this section, we further examine two aspects of our instrumental variable approach that have been attacked by RM. The first is the validity of the exclusion restriction underlying their use of interactions between program choice and the set of exogenous variables (including the village fixed effects) as instruments. The second is the application of the "one-half acre" program eligibility rule. We will show that identification does not require both of these, and present new results dropping each assumption in turn.

The PK model is identified even after dropping all of the exogenous variables including the village fixed effects from the first-stage equations given by equation (4). All that is required is a constant term. Consequently, the identifying instrument is whether or not a household had choice to join a female credit program and/or a male credit program. Estimating a model with only a constant term in the first-stage is of interest because it suggests the extent to which the set of exogenous variables X (and village fixed effects), which have been subject to pointed attack by RM, actually drive the results. Note that even if program placement is nonrandom - that is, even if the program availability is correlated with the village effects - dropping village fixed effects from the first-stage does not cause bias in the second-stage as long as the second-stage equation includes the village fixed effects, as it does in PK.

Column (1) of our Table 10 presents the baseline PK estimates as re-estimated with Roodman's *cmp* program and presented in Pitt (2011a), Table 1, column (3).[33] Column (2) of the table, labeled "No instruments,"[34] presents new estimates dropping all of the slope variables X and all of the village fixed effects from each first-stage equation (4) and (5), leaving only a constant term.[35] What is striking about a comparison of column (2) with column (1) is how little the parameter estimates have changed. Leaving out instruments that RM believe to be troublesome alters the points estimates hardly at all, improves the precision of the estimates, and leaves fully intact the PK conclusion that microcredit provided women positively affects household consumption while microcredit provided men has no statistically discernible effect.[36]

In column (2) of our Table 10 we eliminated exclusion restrictions by dropping all of the exogenous variables, X, plus the village fixed effects (leaving only a constant term) from the first-stage. The polar opposite method to eliminate exclusion restrictions is to retain all of these variables in the first-stage equations and just add them to the second-stage equation. To implement this method, we include interactions between the dummy variable indicating a non-target (ineligible) household with all of the slope variables X plus *thana* (sub-district) fixed effects in column (3) of Table 10, and, alternatively, we include these variables separately interacted with male and female choice in column (4) of Table 10.[37] Once again, these estimates, labeled "No exclusion restrictions" in Table 10,[38] demonstrate the robustness of the PK results to the exclusion restrictions that RM believe are not credible.[39]

To demonstrate that identification in the "no instruments" and "no exclusion restrictions" case does not depend on the Tobit nature of the first-stage equations, we re-estimate these models replacing the Tobit part of the pseudo-log-likelihood function in the first-stage with linear OLS. These estimates are presented in our Table 11. The first column of Table 11 just reproduces the "baseline" PK results of Table 10 and Pitt (2011a). Column (2) of the our Table 11 replaces Tobit with linear OLS with only a slight loss of precision as compared to the model

[33] To remain consistent with the estimates of Pitt (2011a) and (2011b), we continue to use all three rounds of credit data in estimating the first-stage equation.
[34] "No Instruments" means that the first-stage equations ((3) and (4)) have no slope terms.
[35] This looks very much like a Wald estimator as there is a single binary instrument corresponding to each endogenous variable. Because there are two correlated endogenous variables, the credit effect parameter is not given by the ratio of the difference in the means of the dependent variable over the difference in the means of the endogenous variable for each sub-sample as it is in the single endogenous variable case.
[36] The log-likelihood function in the "no instruments" case is bimodal.
[37] There are three villages in a *thana* (sub-district) in the sample frame. Using all villages interacted with both female and male choice yields very similar estimates.
[38] "No exclusion restrictions" means that the second-stage equation contains the full set of slope interactions.
[39] RM have done something similar in their Table 1 of RM (2011). They present two sets of estimates, labeled "*De facto*" and "*De jure*" each in which just the choice dummy variables and then all of the instruments are included, in turn, into the second stage. RM suggest that it is the nonlinearity of the Tobit instrumenting equations that permits model identification in this case. They are mistaken. The model is identified even if the first-stage instrumenting equations are linear OLS. More importantly, even RM report that women's credit effects are positive and significant under both the *de facto* and *de jure* rule when there are "no exclusion restrictions" in their Table 1, and that men's effects are insignificantly different from zero.

with Tobit first-stage equation in column (1). Analogously, column (3) of our Table 11 reproduces the "no instruments" estimates of Table 10 column (2) to enable comparison with the linear OLS estimates of column (4). Once again, the results support the original findings of a statistically significant and positive female credit effect, and a statistically insignificant male credit effect.

Just as we have been able to re-estimate the PK model after dropping all of the instrumental variables consisting of the interactions of choice with all of the exogenous variables (X) and fixed effects, we can also re-estimate the PK model without treating **any** households as having deterministically zero credit as a consequence of the "half-acre" eligibility requirement. That is, we can eliminate the eligibility rule completely and treat all households as endogenously at risk for choosing treatment from a credit program if one exists in their village. As long as the vector of exogenous variables X in equations (4) and (5) have any variables at all beyond a constant, we do not need ineligible (non-target) households in the sample to identify program effects. Identification comes off of variation in program availability by gender across villages.[40]

Contrary to the claims of RM, this eligibility rule and how it may have been enforced is not, in principle, required for estimating program effects.[41] However, if PK in their 1998 paper had treated some households as deterministically without choice when those households actually self-selected themselves into or out of credit programs, the PK estimates would be in error. This very issue was addressed at length in Pitt's 1999 response (Pitt, 1999) to Morduch's first replication, but seems to have been overlooked by RM in their discussion. It is worthwhile to reiterate that the eligibility rule set by the Grameen Bank is **not** that a household must own less than one-half acre of land of any type or quality to be eligible.[42] In that 1999 response paper, Pitt treats nonprogram households with somewhat more than ½ acre of land and as if they have

[40] In the simplest illustrative case, note that β_x in equation (9) can be estimated from the sample of villages without choice. One could then imagine replacing the dependent variable y in equation (9) with $\tilde{y} = y - X\hat{\beta}_x$, where $\hat{\beta}_x$ is the estimate of β_x obtained from villages without program choice. This transformation of y into $\tilde{y}$ makes all of the exogenous variables X (except village fixed effects) available as instruments in an instrumental variables regression of $\tilde{y}$ on female and male credit.

[41] They claim that "in light of the pervasive non-enforcement of the rule evident in Figure 2, the eligibility dummy as defined by PK, and thus the key instruments…should be presumed endogenous (RM (2011, p. 21)).

[42] According to Mahabub Hossain (1988), the most authoritative source of the time, the rule is that "a person from a household that owns less than 0.5 acre of cultivated land, or assets with a value equivalent to less than 1.0 acre of medium-quality land, is eligible to receive a loan" (Hossain 1988, p.25). Note that the rule about landownership applies (i) only to cultivated land, an important distinction among the land poor whose homestead (uncultivated) land may be a significant share of total landholding; and (ii) allows for eligibility when the quantity of cultivable land **exceed 0.5 acres** as long as the value of all assets is less than 1.0 acres of medium quality cultivable land. As Pitt (1999) makes clear in an econometric analysis of land values and program participation with these data, participating households owning more than ½ acre of total land at the time of the survey have dramatically lower land values even after conditioning on total land area and its square, participant status and thana fixed effects. Conditional on the other regressors, "mistargeted" households have unit land values of about one-half that of households which are not "mistargeted", suggesting that they likely fit under the actual *de jure* eligibility rule. Moreover, the other two microcredit programs examined (BRAC and BRDB) report that they followed the Grameen Bank land and asset eligibility rule in the years prior to our survey. The Grameen Bank, as the first microcredit program in Bangladesh, was the role model for subsequent programs initiated later in the 1980's.

choice. By raising the eligibility cutoff to land ownership greater than ½ acre, say 1 acre, all households, whether program participants or not, are treated as having the choice to participate. In this way, those who really do have choice but choose not to participate are now treated appropriately. On the other hand, some of the households that are not program participants with land below 1 acre are actually ineligible. Pitt (1999) claimed that this classification error does not alter the consistency of the estimates of program effects since treating a behavior as endogenous when it is in fact exogenous still yields consistent estimates. However, he could have made an even stronger claim – consistent estimates of credit program effects are possible even when every household, no matter their land or wealth status, is treated as having endogenous treatment choice.

Estimates of program effects in which every household in a treatment village is considered eligible are presented in column (5) of our Table 10.[43] A comparison with the "baseline" estimates of the first column of that table reveal that the qualitative results are essentially unaffected – female credit effects are statistically significant and positive and male credit effects are not statistically different from zero. The estimated female credit effects are about 25 percent larger when no households are considered ineligible and the precision of the estimates is greater. The last column of our Table 10 reproduces results reported in Pitt (1999) Table 4, last column. They are almost identical and likely suggests that the relatively few households with more than 2.0 hectares of land in the sample are not the primary source of parameter identification, but rather it is variation in program placement across villages that drives both these estimates and those in the previous column (column (5) of Table 10) in which there are no ineligible households. What is striking in this exercise is not only that the PK results hold up so well, but that the effect of including ineligible households in the sample is to slightly reduce the estimated credit effects, and the effect of dropping exclusion restrictions is to slightly improve the precision of the estimates.

Finally, we add one more robustness check to these new estimates. When sources of parameter identification used in PK are dropped, as they are in the new estimates described above, there may be enhanced concern that an insufficient set of exogenous covariates in the second-stage equation underlies the credit program effects that are estimated. This is particularly acute in the case of "no instruments" where the choice dummies are key components of identification. The issue of the validity of the instruments was a theme in Morduch's 1998 paper. In particular, Morduch is concerned with the interactions of land with all of the other exogenous regressors because there may be "systematic differences between the landless and landed in, say, the impact of age on income." In Pitt's 1999 response to Morduch, he addressed that concern by re-estimating the model with these interactions. Table 6 of Pitt (1999) presents estimates of the effects of program credit, by program and gender, on the log of household per capita expenditure, allowing for interactions between land ownership and all of the exogenous

[43] As every household is considered eligible, there is no longer a dummy variable indicating target/non-target status in the second-stage equation.

regressors and interactions between land ownership and all of the thana fixed effects. All 18 exogenous regressors and the 29 thana dummy variables are interacted with land and are included in the consumption equation. These interactions are the ones most likely to destroy identification if in fact it is the linearity of the consumption function that is driving the identification of credit effects. Based on these estimates, which allowed for interactions with landholding, Pitt (1999) concluded that the "bottom line" qualitative results of PK still hold -- there are positive and statistically significant effects of female credit program participation on household consumption, and much smaller and generally statistically insignificant effects of male credit program participation on household consumption, as in PK. The estimates first presented in Pitt (1999) were presented again in Pitt (2011b), only re-estimated with Roodman's *cmp* command. They are reproduced here in column (1) of our Table 12. The "headline results" still hold. Women's credit has statistically significant and positive effects on household consumption, although these effects are slightly attenuated with the addition of 47 additional control variables.

The question now is whether the addition of the 47 additional independent variables arising from interaction with landholdings will substantially affect the new results in which, in turn, all "instruments" are dropped and all households are considered eligible. The results from this more highly parameterized second-stage equation are presented in columns (3) and (4) of our Table 12. These estimates do differ very little from their counterpart in Table 10 (compare Table 12 column (3) with Table 10 column (2), and Table 12 column (4) with Table 10 column (5)). Once again, we find statistically significant and positive female credit effects and statistically insignificant male credit effects.

In brief, dropping each of the two major sources of identification in turn does nothing to weaken the effects reported in PK. Nor does combining these new estimates with OLS first-stage equation rather than Tobits, and heavily parameterizing the second-stage with interactions with landownership. The PK results are strikingly robust.

8. Conclusions

"The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence," by David Roodman and Jonathan Morduch, is the most recent of a sequence of papers and postings that seeks to refute the findings of Pitt and Khandker 1998. We find that RM have used testing procedures that necessarily tend to bias the findings in the direction of rejecting the results of PK. A key fault of RM is that it ignores the deterministic nature of choice for a large share of observations. Credit is deterministically zero for a sizable part of the sample and this characteristic is the salient difference between the PK model and the classical two-stage least squares problem on which they focus. RM also make incorrect claims about the causes and implications of bimodality and non-normal errors. Their preferred linear LIML estimates are vastly different from standard 2SLS estimates and are not identified as demonstrated by the 105 out of 110 coefficients reported by Stata to have t-ratios of exactly 0.01 in absolute value.

Nonetheless, they report and rely on the test statistics derived from this obviously faulty IV regression. Finally, we further examine two aspects of our instrumental variable approach that have been attacked by RM. These are the validity of the exclusion restrictions underlying the use of interactions as instruments and the importance of the "one-half acre" program eligibility rule. We find that identification does not require both of these. Indeed, the results originally reported in the PK paper hold up extremely well in this new analysis.

The final question is the value of continuing this process of "replication" and rebuttal, a process that can be dated back to Morduch's (1998) replication paper distributed months before the PK paper was first published. In Section 2 of this paper, we briefly reviewed the recent history of this replication/rebuttal sequence. In this paper, we make the same points about the two-stage setup used by RM that we made in 2011 and in 1999, and we cannot imagine that there is anything to be gained by restating them yet again in the future. Upon discovering errors in their work, RM have always come up with new lines of attack but have never got their econometrics and the 2SLS setup right. Perhaps one reason for the continuation of this process consisting of RM's triumphal announcement of finding devastating problems with PK, followed by rebuttal by PK, is that RM do not believe in non-randomized studies, deeming them "unreliable" and claiming that their complexity hides a form of "obscurantism."[44]

We note that others have also attempted to replicate PK. Duvendack and Palmer-Jones (2012) have published a paper that is both a "replication" of PK and a replication of an earlier replication of PK by Chemin (2008) in the *Journal of Development Studies*. Pitt's reply paper (2012), published in the same issue, lists many serious errors in their code and misrepresentations of method in their paper, and concludes that the Duvendack and Palmer-Jones results provide no credible evidence on the validity of PK or Chemin or on the effectiveness of

[44] On this, Roodman has written:

> My main conclusion about non-randomized quantitative studies is that they are obscurantism, however unintentional, and what their complexity hides, ultimately, is a failure to prove the assumptions needed to demonstrate cause and effect. Warren Buffett's investment rule—don't buy what you don't understand—works in microfinance research too. RCTs are easy to understand, so you should buy them more than non-randomized studies. *(http://blogs.cgdev.org/open_book/2011/06/rcts-are-people-too.php)*

and

> The confusing math [referring to some equations describing the model of Pitt and Khandker], *nota bene*, is from a non-randomized study. The math for randomized control trials (RCTs) could hardly be simpler. You flip a coin a bunch of times to decide who is offered a financial service and who isn't. Then you come back later, and for each of the two groups—those offered and those not—you compute the average outcome (are you still with me?). Then you subtract one average from the other. Got that? Then you're done. Even my third grader can understand that math. Well, my third grader derived the formula for the area of a regular octagon…but his classmates can understand RCT math too. So let's not taint the method with obscurantism. *(http://blogs.cgdev.org/open_book/2011/06/rcts-are-people-too.php)*

microfinance. While we doubt that either side will change its mind, we believe that it is best for all concerned to move on to new research endeavors.

References

Chemin, Matthieu (2008) The Benefits and Costs of Microfinance: Evidence from Bangladesh. *Journal of Development Studies*, 44(4), pp. 463-484.

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.

Davidson and MacKinnon 1993. *Estimation and Inference in Econometrics*. Oxford University Press.

Duvendack, Maren, and Richard Palmer-Jones (2012)"High Noon for Microfinance Impact Evaluations: Re-investigating the Evidence from Bangladesh. *Journal of Development Studies* (forthcoming).

Ferguson, Thomas S. (1978) "Maximum Likelihood Estimates of the Parameters of the Cauchy Distribution for Samples of Size 3 and 4" , *Journal of the American Statistical Association*, 73(361): 211-213.

Greene, William. 2012, *Econometric Analyses*, (7th edition). Prentice Hall, NY.

Hossain, Mahabub. 1988. *Credit for Alleviation of Rural Poverty: The Grameen Bank in Bangladesh*. IFPRI Research Report 65. International Food Policy Institute: Washington D.C. (relevant chapter available at http://www.ifpri.org/sites/default/files/pubs/pubs/abstract/65/rr65ch04.pdf)

Imbens, Guido. 2010. *Weak Instruments and Many Instruments.* Lecture Notes 4, Miami. http://www.bus.miami.edu/_assets/files/events/miami_weak_iv

Karlan, Dean and Jonathan Morduch. 2009. "Access to Finance," Chapter 2 in *Handbook of Development Economics, Volume 5*, Dani Rodrik and Mark Rosenzweig, editors. Manuscript available at http://www.nyu.edu/projects/morduch/documents/articles/2009-06-HDE_AccesstoFinance.pdf

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

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

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

Nagar, A.L. 1959. "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations", *Econometrica* 27, 575 - 595.

Newey, Whitney K. and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Volume 4, ed. R. F. Engle and D. McFadden. Amsterdam: North Holland, 2111-2245.

Pitt, Mark M. 1999. Reply to Jonathan Morduch's "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh." Department of Economics. Brown University. Available at http://www.brown.edu/research/projects/pitt/.

Pitt, Mark M. 2011a. Response to Roodman and Morduch's "The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence"

Pitt, Mark M. 2011b. Overidentification Tests and Causality: A Second Response to Roodman and Morduch. Available at http://www.brown.edu/research/projects/pitt/.

Pitt, Mark M. 2012. Gunfight at the NOT OK Corral: Reply to "High Noon for Microfinance," *Journal of Development Studies* (forthcoming) (available at http://www.brown.edu/research/projects/pitt/)

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. *Due Diligence: An Impertinent Inquiry into Microfinance*. CGD Books.

Roodman, David and J. Morduch. 2009. The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence, Working Paper 174, Center for Global Development.

Roodman, David and J. Morduch. 2011. The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence (Revised 2011), Working Paper 174, Center for Global Development.

Figures

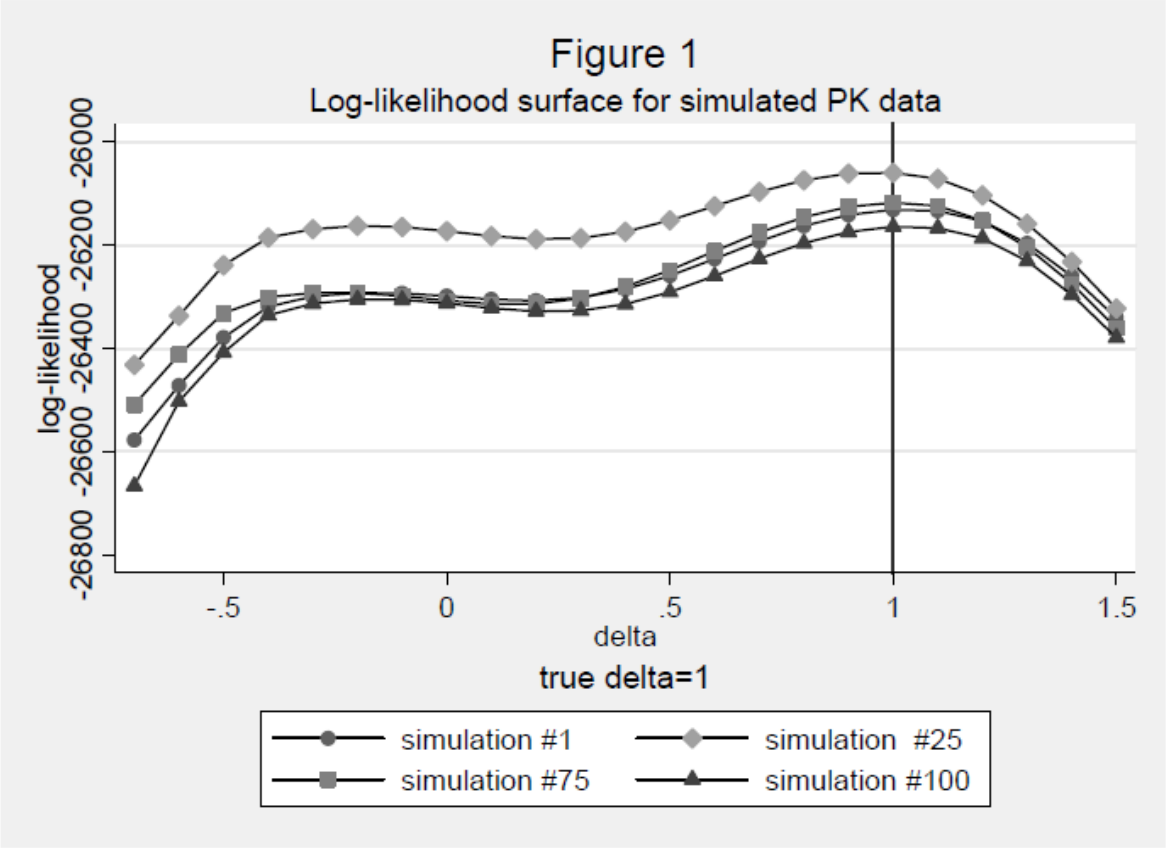
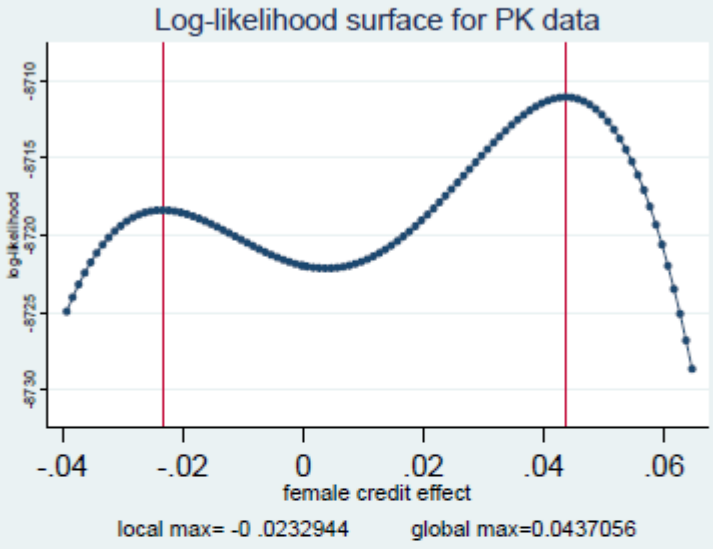
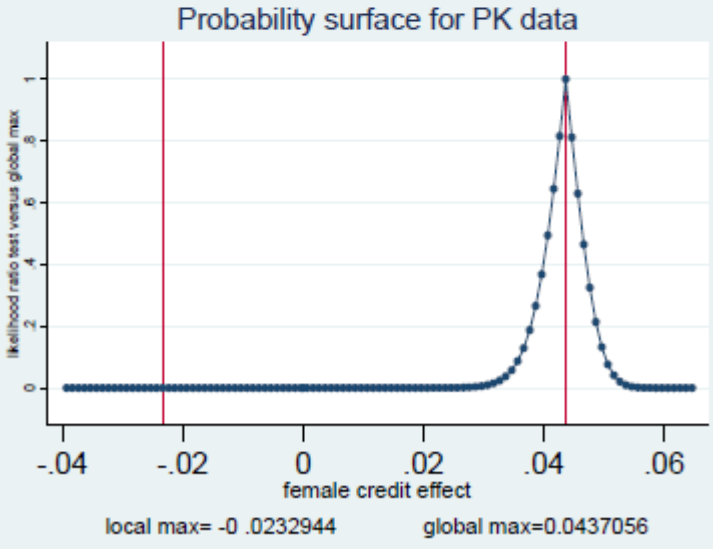


Figure 1
Log-likelihood surface for simulated PK data
log-likelihood
-26000
-26200
-26400
-26600
-26800
-.5
0
.5
1
1.5
delta
true delta=1
simulation #1
simulation #25
simulation #75
simulation #100

Figure 2


Log-likelihood surface for PK data
log-likelihood
-8710
-8715
-8720
-8725
-8730
-.04
-.02
0
.02
.04
.06
female credit effect
local max= -0 .0232944
global max=0.0437056


Probability surface for PK data
likelihood ratio test versus global max
0
.2
.4
.6
.8
1
-.04
-.02
0
.02
.04
.06
female credit effect
local max= -0 .0232944
global max=0.0437056

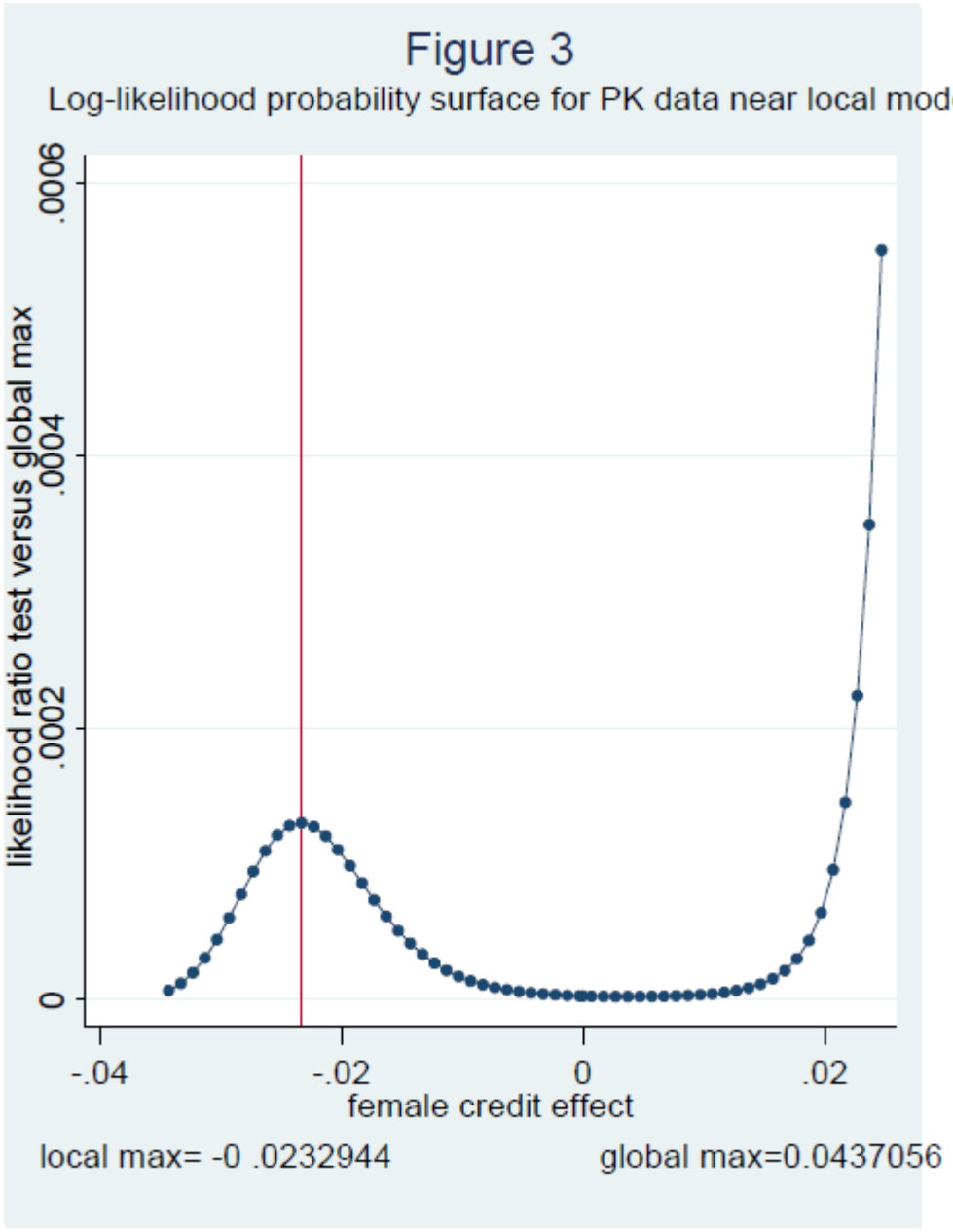


Figure 3
Log-likelihood probability surface for PK data near local mode
likelihood ratio test versus global max
0
.0002
.0004
.0006
-.04
-.02
0
.02
female credit effect
local max= -0 .0232944
global max=0.0437056

Tables

Table 1. Maximum likelihood and two-stage least squares estimates of program effects with simulated data (200 simulated datasets, 5000 observations each)

Program effect δ	True value	PK method (maximum likelihood)	RM instruments and method (standard 2SLS)
Male group 1	.250	.253 (.037)	.482 (.248)
Male group 2	.250	.253 (.036)	.504 (.214)
Male group 3	.250	.248 (.036)	.500 (.222)
Female group 1	.750	.747 (.038)	.526 (.233)
Female group 2	.750	.750 (.034)	.513 (.247)
Female group 3	.750	.748 (.035)	.531 (.235)

Note: Standard deviations in parenthesis

Table 2. Robustness of the PK estimator to Non-normal Errors: Skew Distributions
200 simulated datasets, true $\delta_m = \delta_f = 1.0$

	RM 2SLS		**PK method (ML)**	
	Coefficient	**Standard deviation**	**Coefficient**	**Standard deviation**
Male credit effect δ_m	1.282	.109	.997	.031
Female credit effect δ_f	1.282	.132	.998	.032
	Tests of normality			
Test type:	Median p-value			
Skewness	3.49e-08			
Kurtosis	.309			
Skewness and kurtosis	3.64e-07			

Note: Sum of normal and half-normal for all equations. First-stage specified as Tobit.

Table 3. Robustness of the PK estimator to Non-normal Errors: Distributions with Non-normal Kurtosis
200 simulated datasets, true $\delta_m = \delta_f = 1.0$

	RM 2SLS		**PK method (ML)**	
	Coefficient	**Standard deviation**	**Coefficient**	**Standard deviation**
Male credit effect δ_m	1.362	.163	.996	.034
Female credit effect δ_f	1.375	.148	.995	.041
	Tests of normality			
Test type:	Median p-value			
Skewness	.380			
Kurtosis	7.64e-11			
Skewness and kurtosis	2.34e-09			

Note: Sum of normal and t(20) distribution for all equations. First-stage specified as Tobit.

Table 4. Bimodality with simulated PK data

Panel A:

Global max of log-likelihood (100 simulated data sets)	
Value of δ_f	**Frequency of global max**
0.9	3
1.0	96
1.1	2
	Total=100 simulations

Panel B:

Local max of log-likelihood (100 simulated data sets)	
Value of δ_f	Frequency of local max
-0.2	5
-0.1	68
0	26
0.1	1
	Total=100 simulations

Table 5. Replication of RM's Table 6, column 1, labeled "Tests of underidentification and weak identification in linear LIML"
(dependent variable is measure of household consumption)

Names of variables in RM's Stata dataset	RM linear LIML (reported in RM Table 6, col 1)		2SLS of exactly the same model using RM's data and code	
	Coefficents	t-ratio	Coefficents	t-ratio
(1)	(2)	(3)	(4)	(5)
lfbraclvpk1 (fem. credit BRAC)	-3.815685	-0.01	.0103838	0.62
lfbrdblvpk1 (fem. credit BRDB)	-.6650787	-0.01	-.0249588	-0.82
lfgramlvpk1 (fem. Credit Grameen)	-.7829703	-0.01	.0085772	0.69
lmbraclvpk1 (male credit BRAC)	3.929513	0.01	.0307323	1.30
lmbrdblvpk1 (male credit BRDB)	-2.229001	-0.01	.0016568	0.07
lmgramlvpk1 (male credit Grameen)	-.8136244	-0.01	-.0021149	-0.09
scoheadpk	-.315985	-0.01	.0860843	2.17
afedhighpk	-.2452647	-0.01	.0177772	3.46
amedhighpk	.1776192	0.01	.0066512	0.92
afadultdpk	-2.481861	-0.01	.1641031	2.08
amadultdpk	-.8814257	-0.01	-.1186432	-1.22
sexheadpk	-2.478439	-0.01	-.0467679	-0.52
ageheadpk	.0129599	0.01	-.0020182	-2.42
llandbefpk	-.126666	-0.01	.0259332	3.41
edheadpk	-.1274544	-0.01	.0121871	1.58
spsislnddpk	-.0475071	-0.01	.0040901	0.45
spbrolnddpk	.1945562	0.01	.0220427	2.48
spparlnddpk	-.5885473	-0.01	-.0009438	-0.06
hdsislnddpk	-.1564894	-0.01	-.0047917	-0.50
hdbrolnddpk	-.0410882	-0.01	-.0067167	-0.76
hdparlnddpk	-.0685464	-0.01	.0294138	1.68
_Iwave_2	-.0195645	-0.04	-.0141153	-0.96
_Iwave_3	-.2421109	-0.25	-.2303045	-16.08
nontarpk	-2.370799	-0.01	.1463444	3.39
_Ivillag~_12	1.597659	0.01	-.0088011	-0.12
_Ivillag~_13	-1.622276	-0.01	-.0543815	-0.77
_Ivillag~_21	-3.618246	-0.01	.009135	0.11
_Ivillag~_22	-4.332557	-0.01	-.0692139	-0.73
_Ivillag~_23	-2.847171	-0.01	.0054571	0.06
_Ivillag~_31	2.947799	0.01	.3496697	3.76
_Ivillag~_32	.9148231	0.03	.5596841	4.75
_Ivillag~_33	-10.8526	-0.01	.2406412	2.85
_Ivillag~_41	-9.035545	-0.01	.0163955	0.16
_Ivillag~_42	.2209561	0.01	-.0301195	-0.24
_Ivillag~_43	1.762032	0.01	-.0421719	-0.48
_Ivillag~_51	-3.912438	-0.01	-.1332913	-1.48

_Ivillag~_52	-1.38246	-0.01	.0085667	0.10
_Ivillag~_53	1.100732	0.01	-.0421889	-0.48
_Ivillag~_61	-7.763293	-0.01	.1299685	1.30
_Ivillag~_62	3.10183	0.01	.4219448	4.44
_Ivillag~_63	.0094961	0.00	.3994444	3.45
_Ivillag~_71	1.431762	0.01	.1761968	2.06
_Ivillag~_72	-1.042307	-0.01	.1305833	1.17
_Ivillag~_73	-1.382047	-0.01	.1061143	1.32
_Ivillag~_81	-5.841731	-0.01	-.0265903	-0.30
_Ivillag~_82	-4.605215	-0.01	.1768697	1.78
_Ivillag~_83	-6.355661	-0.01	.2743733	2.49
_Ivillag~_91	-2.685624	-0.01	.3285115	2.64
_Ivillag~_92	-4.215806	-0.01	-.0056925	-0.04
_Ivillag~_93	-1.515515	-0.01	.0854598	0.56
_Ivillag~101	-4.504267	-0.01	.0231246	0.21
_Ivillag~102	-5.692452	-0.01	.0997992	0.83
_Ivillag~103	-5.413564	-0.01	.3622729	3.83
_Ivillag~111	-4.845666	-0.01	-.1400082	-1.25
_Ivillag~112	-3.446495	-0.01	.099935	0.84
_Ivillag~113	-4.372699	-0.01	.4040249	2.41
_Ivillag~121	-4.444865	-0.01	.1256401	1.03
_Ivillag~122	-4.247319	-0.01	.1008418	0.79
_Ivillag~123	-6.681669	-0.01	.1082682	1.03
_Ivillag~131	-2.823004	-0.01	.2367743	2.43
_Ivillag~132	-2.951351	-0.01	.3008999	2.24
_Ivillag~133	-4.333323	-0.01	.3330268	1.52
_Ivillag~141	-5.372782	-0.01	.0500823	0.58
_Ivillag~142	-4.960455	-0.01	.2856374	2.92
_Ivillag~143	-3.116554	-0.01	.0209507	0.21
_Ivillag~151	-5.820668	-0.01	.1168385	1.04
_Ivillag~152	1.167932	0.02	.2254976	1.73
_Ivillag~153	-1.583184	-0.01	.1439074	1.36
_Ivillag~161	-4.722555	-0.01	.1451311	0.97
_Ivillag~162	-3.377368	-0.01	.0571318	0.44
_Ivillag~163	-4.463402	-0.01	.1010068	0.63
_Ivillag~171	-3.816095	-0.01	.1235445	1.27
_Ivillag~172	-4.514936	-0.01	.2544887	2.53
_Ivillag~173	-2.267361	-0.01	.2602616	2.40
_Ivillag~181	-4.913823	-0.01	.2254458	2.27
_Ivillag~182	-6.007032	-0.01	-.1079303	-1.01
_Ivillag~183	-6.644546	-0.01	.1181285	1.18
_Ivillag~191	-3.348686	-0.01	.0641058	0.51
_Ivillag~192	-3.748686	-0.01	.0754787	0.80
_Ivillag~193	-3.627202	-0.01	-.0543931	-0.51
_Ivillag~201	-4.121989	-0.01	.1431313	1.19
_Ivillag~202	-5.350993	-0.01	-.0121169	-0.13
_Ivillag~203	-4.348962	-0.01	.0466108	0.50
_Ivillag~211	-5.425008	-0.01	.2397429	2.84
_Ivillag~212	-5.3651	-0.01	.0206076	0.21

_Ivillag~213	-5.516586	-0.01	.0020289	0.02
_Ivillag~221	-4.013762	-0.01	.0478596	0.42
_Ivillag~222	-4.344287	-0.01	.2236828	1.67
_Ivillag~223	-4.784044	-0.01	-.0477484	-0.42
_Ivillag~231	-5.68571	-0.01	.5123455	4.07
_Ivillag~232	-3.613266	-0.01	.3214447	3.23
_Ivillag~233	-3.835115	-0.01	.3726069	3.21
_Ivillag~241	-4.702653	-0.01	.1334711	1.31
_Ivillag~242	-2.336784	-0.01	.0158185	0.09
_Ivillag~243	-3.506021	-0.01	.2100095	1.70
_Ivillag~251	-6.645435	-0.01	.065295	0.63
_Ivillag~252	-7.184654	-0.01	.1985823	2.01
_Ivillag~253	-5.885879	-0.01	.3563095	2.84
_Ivillag~261	-6.272511	-0.01	.1707637	1.60
_Ivillag~262	-7.614288	-0.01	.154777	1.43
_Ivillag~263	-6.339185	-0.01	.2613025	2.88
_Ivillag~271	-7.180523	-0.01	.1111286	1.13
_Ivillag~272	-6.307533	-0.01	-.0047516	-0.04
_Ivillag~273	-6.458414	-0.01	.1555054	1.68
_Ivillag~281	-5.030298	-0.01	.6442677	2.95
_Ivillag~282	-5.369935	-0.01	.4205384	3.21
_Ivillag~283	-6.193199	-0.01	.2967033	2.40
_Ivillag~291	-6.70802	-0.01	-.1901881	-1.93
_Ivillag~292	-7.472168	-0.01	.0201885	0.20
_Ivillag~293	-6.687459	-0.01	-.0560393	-0.56
_cons	14.67148	0.02	4.129303	31.61
F(110, 1756)	1.85		14.88	
R^2 (Centered)	-381.4		0.3123	
Root mean squared error	9.03		0.38	

Table 6. Comparison of Roodman and Moduch's own 2SLS and linear-LIML 2SLS estimates (dependent variable is measure of household consumption)

Variable	2SLS	2SLS by linear LIML
	(1)	(2)
female credit BRAC	0.009 (0.53)	-3.816 (0.01)
male credit BRAC	0.034 (1.45)	3.930 (0.01)
Female credit BRDB	-0.024 (0.76)	-0.665 (0.01)
Male credit BRDB	0.001 (0.06)	-2.229 (0.01)
female Grameen	0.010 (0.75)	-0.783 (0.01)
male credit Grameen	-0.002 (0.07)	-0.814 (0.01)

Col 1: Posted by David Roodman in his blog as "Response to Pitt's Response to Roodman and Morduch's Replication of…, etc." http://blogs.cgdev.org/open_book/2011/03/response-to-pitts-response-to-roodman-and-morduchs-replication-of-etc.php

Col 2. RM (2012) Table 6 col (1)

Table 7. Replication of RM's Table 6, column 2, labeled "Tests of underidentification and weak identification in linear LIML"
(dependent variable is measure of household consumption)

Names of variables in RM's Stata dataset	**RM linear LIML (reported in RM Table 6)**		**Standard 2SLS t-ratios for the LIML estimates of column (2)**
	Coefficents	**t-ratios (κ= 1.08117)**	**t-ratios (κ= 1.0)**
(1)	(2)	(3)	(4)
lfproglvp~n1 (Female credit)	.4452663	0.10	10.63
lmproglvp~n1 (Male credit)	.0020398	0.00	0.05
scoheadpk	.2394207	0.17	2.11
afedhighpk	.0364407	0.20	2.64
amedhighpk	.0005662	0.01	0.03
afadultdpk	.6375085	0.14	3.68
amadultdpk	.5016348	0.08	1.57
sexheadpk	.7260898	0.09	2.62
ageheadpk	-.0109582	-0.12	-3.33
llandbefpk	-.0015224	-0.01	-0.06
edheadpk	.0195687	0.24	1.00
spsislnddpk	.0132932	0.15	0.46
spbrolnddpk	.0148581	0.18	0.54
spparlnddpk	.0494096	0.09	1.01
hdsislnddpk	-.0125403	-0.12	-0.43
hdbrolnddpk	-.0277971	-0.14	-1.01
hdparlnddpk	.0639722	0.16	1.09
_Iwave_2	-.0127039	-0.48	-0.85
_Iwave_3	-.2278736	-6.20	-15.25
nontarpk	1.025186	0.13	8.92
_Ivillag~_12	.2460201	0.10	0.47
_Ivillag~_13	.7495827	0.09	1.47
_Ivillag~_21	-.0568234	-0.06	-0.10
_Ivillag~_22	.6692253	0.09	1.35
_Ivillag~_23	.6307296	0.10	1.24
_Ivillag~_31	.7673286	0.19	1.47
_Ivillag~_32	1.184245	0.19	2.11
_Ivillag~_33	1.251508	0.12	2.58
_Ivillag~_41	.3353928	0.09	0.66
_Ivillag~_42	.7199609	0.09	1.54
_Ivillag~_43	.4409875	0.09	0.81
_Ivillag~_51	.3248061	0.07	0.63
_Ivillag~_52	.6897275	0.10	1.37
_Ivillag~_53	.676902	0.10	1.08
_Ivillag~_61	1.15305	0.11	2.16
_Ivillag~_62	.3071422	0.24	0.50

_Ivillag~_63	1.180982	0.15	2.17
_Ivillag~_71	.7479336	0.13	1.40
_Ivillag~_72	.9285035	0.12	1.86
_Ivillag~_73	.7465784	0.12	1.47
_Ivillag~_81	.4560349	0.09	0.89
_Ivillag~_82	.5741275	0.14	1.08
_Ivillag~_83	1.023173	0.13	2.02
_Ivillag~_91	1.197617	0.12	2.47
_Ivillag~_92	1.227133	0.09	2.63
_Ivillag~_93	.7645231	0.09	1.44
_Ivillag~101	1.454267	0.10	3.19
_Ivillag~102	1.416334	0.11	3.07
_Ivillag~103	1.26254	0.13	2.63
_Ivillag~111	1.420351	0.09	3.11
_Ivillag~112	.6056902	0.10	1.14
_Ivillag~113	1.823532	0.12	3.99
_Ivillag~121	1.580979	0.10	3.61
_Ivillag~122	1.444818	0.10	3.09
_Ivillag~123	1.182559	0.11	2.42
_Ivillag~131	1.618145	0.11	3.65
_Ivillag~132	1.763618	0.11	3.84
_Ivillag~133	1.424838	0.12	2.93
_Ivillag~141	1.205933	0.10	2.74
_Ivillag~142	1.774451	0.11	3.98
_Ivillag~143	1.258546	0.10	2.77
_Ivillag~151	1.02176	0.11	2.19
_Ivillag~152	.8304332	0.10	1.49
_Ivillag~153	.7252304	0.10	1.45
_Ivillag~161	-.0352342	-0.04	-0.06
_Ivillag~162	1.413944	0.10	3.20
_Ivillag~163	.4273315	0.10	0.80
_Ivillag~171	.790332	0.11	1.63
_Ivillag~172	1.319859	0.12	2.79
_Ivillag~173	-.7290758	-0.07	-1.18
_Ivillag~181	1.153316	0.11	2.22
_Ivillag~182	.5981195	0.08	1.24
_Ivillag~183	1.099592	0.11	2.17
_Ivillag~191	.3456389	0.10	0.65
_Ivillag~192	.6281852	0.10	1.22
_Ivillag~193	-.0005375	-0.00	-0.00
_Ivillag~201	1.208697	0.10	2.53
_Ivillag~202	.7595727	0.10	1.49
_Ivillag~203	.1743798	0.13	0.32
_Ivillag~211	1.430806	0.11	3.19
_Ivillag~212	.0781275	0.10	0.13
_Ivillag~213	.1989295	0.10	0.33
_Ivillag~221	.6965733	0.10	1.52
_Ivillag~222	.8745241	0.12	1.57
_Ivillag~223	.6915736	0.09	1.41

_Ivillag~231	1.232368	0.17	2.24
_Ivillag~232	.0659367	0.03	0.11
_Ivillag~233	.5127151	0.35	0.89
_Ivillag~241	.3895818	0.15	0.64
_Ivillag~242	1.125981	0.09	2.33
_Ivillag~243	1.318739	0.10	2.58
_Ivillag~251	1.421445	0.10	3.09
_Ivillag~252	1.689377	0.11	3.68
_Ivillag~253	1.650934	0.13	3.52
_Ivillag~261	1.439462	0.11	3.06
_Ivillag~262	1.803983	0.11	4.01
_Ivillag~263	1.596262	0.12	3.63
_Ivillag~271	1.599431	0.11	3.43
_Ivillag~272	1.273785	0.10	2.72
_Ivillag~273	1.504177	0.11	3.25
_Ivillag~281	1.800998	0.15	3.85
_Ivillag~282	1.65278	0.13	3.66
_Ivillag~283	1.642492	0.12	3.53
_Ivillag~291	1.042607	0.08	2.27
_Ivillag~292	1.510559	0.10	3.35
_Ivillag~293	1.279386	0.09	2.81
_cons	2.03682	0.10	3.76
F(106, 1756) = 7.49			
R^2 (Centered) = -7.2631			
Root mean squared error = 1.327			

Table 8 . The effect of Roodman and Morduch's use of linear LIML on estimates of program effects using simulated PK data (200 simulated datasets)

Program effect δ	True value	PK method	RM instruments	
		maximum likelihood	2SLS	Linear LIML (RM method)
	(1)	(2)	(3)	(4)
Male group 1	.250	.253 (.037)	.482 (.248)	1.146 (7.754)
Male group 2	.250	.253 (.036)	.504 (.214)	-.386 (6.693)
Male group 3	.250	.248 (.036)	.500 (.222)	.861 (15.100)
Female group 1	.750	.747 (.038)	.526 (.232)	.338 (18.614)
Female group 2	.750	.750 (.034)	.513 .246)	.469 (8.057)
Female group 3	.750	.748 (.035)	.531 (.235)	1.161 (17.022)

Note: Standard deviation in parenthesis

Table 9. The effect of Roodman and Morduch's use of linear LIML on estimates of program effects using simulated PK data (results for simulated dataset number 141 only)

Program effect δ	**True value**	**PK method**	**RM instruments**	
		maximum likelihood	**2SLS**	**Linear LIML (RM method)**
	(1)	(2)	(3)	(4)
Male group 1	.250	.276	.493	21.738
Male group 2	.250	.319	.651	-42.193
Male group 3	.250	.280	.451	177.334
Female group 1	.750	.742	.349	-244.388
Female group 2	.750	.750	.557	94.795
Female group 3	.750	.724	.503	218.990

Table 10. Pitt and Khandker estimates with and without instruments, exclusion restrictions, and eligibility requirements

Explanatory Variables	**Baseline Pitt (2011a)**	**No instruments**	**No exclusion restrictions (target interactions)**	**No exclusion restrictions (choice interactions)**	**No HHs ineligible**	**Ineligible only if land owned > 2.0 hectares (Pitt 1999)**
	(1)	(2)	(3)	(4)	(5)	(6)
Amount borrowed by female from BRAC	.0443 (4.78)	.0431 (5.10)	.0413 (4.83)	.0407 (4.83)	.0559 (8.67)	.0500 (8.077)
Amount borrowed by female from BRDB	.0458 (4.30)	.0408 (4.56)	.0459 (4.80)	.0468 (4.86)	.0595 (7.62)	.0512 (6.801)
Amount borrowed by female from GB	.0420 (4.80)	.0409 (5.18)	.0410 (5.17)	.0407 (5.27)	.0547 (8.87)	.0527 (7.805)
Amount borrowed by male from BRAC	.0093 (0.52)	-.0113 (-1.21)	.0123 (0.94)	.0045 (0.29)	.0014 (0.12)	.0066 (0.265)
Amount borrowed by male from BRDB	.0128 (0.70)	-.0097 (-1.15)	.0176 (1.41)	.0098 (0.66)	.0035 (0.31)	.0077 (0.303)
Amount borrowed by male from GB	.0072 (0.45)	-.0121 (-1.43)	.0113 (1.01)	.0055 (0.43)	.0003 (0.03)	.0006 (0.023)
No. of observations	5218	5218	5218	5218	5218	5218

Sources: Col (1): Pitt (2011a) Table 1, col (3); Col (5): Pitt (1999) Table 4, col (5);
All others are author's calculations.

Table 11. Pitt and Khandker estimates with and without instruments or exclusion restrictions: OLS compared to Tobit first-stages equations

	Baseline Pitt (2011a) (Tobit)	**Baseline**	**No instruments (Tobit)**	**No Instruments**
		OLS first-stages		**OLS first-stages**
Explanatory Variables	(1)	(2)	(3)	(4)
Amount borrowed by female from BRAC	.0443 (4.78)	.0434 (3.37)	.0431 (5.10)	.0383 (3.42)
Amount borrowed by female from BRDB	.0458 (4.30)	.0415 (3.13)	.0408 (4.56)	.0356 (3.13)
Amount borrowed by female from GB	.0420 (4.80)	.0430 (3.39)	.0409 (5.18)	.0387 (3.52)
Amount borrowed by male from BRAC	.0093 (0.52)	.0076 (0.27)	-.0113 (-1.21)	-.0181 (-1.46)
Amount borrowed by male from BRDB	.0128 (0.70)	.0113 (0.37)	-.0097 (-1.15)	-.0164 (-1.37)
Amount borrowed by male from GB	.0072 (0.45)	.0067 (0.24)	-.0121 (-1.43)	-.0197 (-1.62)
No. of observations	5218	5218	5218	5218

Sources: Col (1), Pitt (2011a), Table 1, col (3)
All others are author's calculations.

Table 12. Pitt and Khandker estimates with and without instruments, exclusion restrictions, and eligibility requirements: Landholding interactions included as second-stage covariates

Explanatory Variables	Baseline Pitt (2011a)	Full interactions with quantity of land owned		
		Baseline Pitt (2011b)	No instruments (target interactions)	No HHs ineligible
	(1)	(2)	(3)	(4)
Amount borrowed by female from BRAC	.0443 (4.78)	.0372 (4.15)	.0363 (4.26)	.0500 (8.02)
Amount borrowed by female from BRDB	.0458 (4.30)	.0388 (3.77)	.0348 (3.88)	.0538 (7.15)
Amount borrowed by female from GB	.0420 (4.80)	.0358 (4.19)	.0351 (4.37)	.0494 (8.26)
Amount borrowed by male from BRAC	.0093 (0.52)	.0108 (0.81)	-.0061 (-0.63)	.0004 (0.04)
Amount borrowed by male from BRDB	.0128 (0.70)	.0151 (1.14)	-.0038 (-0.42)	.0031 (0.31)
Amount borrowed by male from GB	.0072 (0.45)	.0073 (0.62)	-.0084 (-0.98)	-.0016 (-0.18)
No. of observations	5218	5218	5218	5218

Sources: Col (1): Pitt (2011a) Table 1, col (3); Col (2): Pitt(2011b) Table1, col(3);
All others are author's calculations.

Appendix A

This is a very simple example in Stata of generating 5000 observations for which the first 2000 observations have female choice, observations 1001 to 3000 have male choice, and last 2000 observations have no choice. There are only two exogenous regressors x1 and x2 that determine female credit cf and male credit cm. Following Roodman and Morduch, the instruments are the interactions of choice and the exogenous regressors. The code is:

```
set seed 8279104
set obs 5000
gen x1=4*uniform()
gen x2=4*uniform()
gen cf = x1 + x2 + 5*rnormal() if _n < 2001
replace cf=0 if cf==.
gen cm = x1 + x2 + 5*rnormal() if _n > 1000 & _n < 3001
replace cm=0 if cm==.
gen fchoice = _n < 2001
gen mchoice = _n > 1000 & _n < 3001
gen zfx1=x1*fchoice
gen zmx1=x1*mchoice
gen zfx2=x2*fchoice
gen zmx2=x2*mchoice
regress cf x* zf* zm*
testparm zm*
regress cm x* zf* zm*
testparm zf*
```

The regression output is:

```
. regress cf x* zf* zm*

------------------------------------------------------------------------------
        cf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
        x1 |   .0419391   .0499062     0.84   0.401    -.055899    .1397772
        x2 |   .0077155   .0502128     0.15   0.878   -.0907237    .1061546
      zfx1 |   .9917728   .0617552    16.06   0.000    .8707055     1.11284
      zfx2 |   1.000131   .0619533    16.14   0.000    .8786753    1.121587
      zmx1 |  -.0894075   .0615399    -1.45   0.146   -.2100528    .0312378
      zmx2 |   .0150237   .0619004     0.24   0.808   -.1063283    .1363757
     _cons |  -.0583808   .1191893    -0.49   0.624   -.2920442    .1752825
------------------------------------------------------------------------------
. testparm zm*: ( 1)zmx1 = 0  ( 2)zmx2 = 0, F(  2,  4993) = 1.88, Prob > F =  0.1533
. regress cm x* zf* zm*
        cm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
        x1 |   .0496884   .0486819     1.02   0.307    -.0457494    .1451263
        x2 |  -.0103255   .0489809    -0.21   0.833    -.1063497    .0856987
      zfx1 |  -.0899551   .0602402    -1.49   0.135    -.2080523    .0281421
      zfx2 |   .0959736   .0604334     1.59   0.112    -.0225023    .2144496
      zmx1 |   1.009026   .0600302    16.81   0.000     .8913402    1.126711
      zmx2 |    .933007   .0603818    15.45   0.000     .8146321    1.051382
     _cons |  -.0972115   .1162652    -0.84   0.403    -.3251424    .1307194
------------------------------------------------------------------------------
```

. testparm zf*: (1)zfx1 = 0 (2)zfx2 = 0, F(2, 4993) = 1.36, Prob > F = 0.2558

The *Stata* regression coefficients on x1 and x2 in both first-stage regressions correspond to π_{fx} and π_{mx} for the first and second regression, respectively, and the *Stata* regression coefficients on zmx1 and zmx2 in the first regression, and zfx1 and zfx2 in the second regression, correspond to π_{fm} and π_{mf} from equations (8) and (7), respectively. These are then the "weak instruments" added by the RM setup.