Re-Re-Reply to "The Impact of Microcredit on the Poor in Bangladesh:

Revisiting the Evidence"

Mark M. Pitt[1]

Abstract

"The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence," by David Roodman and Jonathan Morduch (2014) 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.   This response paper refutes the claims of Roodman and Morduch that were not addressed in the earlier World Bank working paper of Pitt and Khandker (2012). This response paper, like the Pitt and Khandker (2012) paper and others that preceeded it, shows that many of the Roodman and Morduch claims are based on a flawed econometric understanding and a lack of due diligence in formulating and interpreting statistical models.

JEL classification numbers: N01, C510

Keywords: microcredit, microfinance, replication, Bangladesh, Grameen Bank, program evaluation

[1] Brown University. I thank Shahidur Khandker and Tiemen Woutersen for comments and suggestions. A shortened version of this paper appears in the *Journal of Development Studies* as Pitt (2014).

3 February 2014

Re-Re-Reply to "The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence"

by Mark M. Pitt[1]

"The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence," by David Roodman and Jonathan Morduch (henceforth RM) is the most recent of a sequence of papers and web postings that seeks to refute the findings of the Pitt and Khandker (1998; henceforth PK) 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 household consumption and other outcomes. In this version of RM, the authors have backed off many of their prior claims and methods after earlier replies noted their faults (see Pitt (1999), Pitt (2011a), Pitt (2011b), and Pitt and Khandker (2012)). Nonetheless, important claims against PK remain in this new version of RM and are addressed below. Readers should refer to Pitt and Khandker (2012) for a discussion of other issues with RM, including a discussion of the bimodal likelihood.

1. The econometrics of RM and PK

The essential nature of the model as formulated in PK is quite simple. Some individuals have the choice of participating in a credit program (receiving treatment), and the choice to participate is exogenous conditional on a set of exogenous covariates that include village fixed-effects. Actual participation and the level of treatment (the amount borrowed) are endogenous. Using the RM notation but abstracting from gender-specific choice for brevity,[2] the log-likelihood L_{PK} that is maximized is the sum of the standard two-stage least squares likelihood (L_{2SLS}) and the standard ordinary least squares (regression) likelihood (L_{OLS}).

(1) $L_{PK} = L_{2SLS}(y_0\text{-}y_{fm}\delta\text{-}\mathbf{x}\beta_0,\ y_{fm}\text{-}\mathbf{x}\beta_{fm})$ for those with the choice to participate

$+ L_{OLS}(y_0\text{-}\mathbf{x}\beta_0)$ for those without the choice to participate

Note the following about the log-likelihood L_{PK}:

(A) There is only one set of exogenous variables, **x**. There are no exclusion restrictions and thus there is no separate set of instruments **z**.[3] The PK model is not the usual sort of instrumental variables model in

[1] Brown University. I thank Shahidur Khandker and Tiemen Woutersen for comments and suggestions. A shortened version of this paper appears in the *Journal of Development Studies* as Pitt (2014).

[2] In order to focus only on the two key differences between the PK and RM instrumental variable models, which are (i) RM's use of a set of identifying instruments **z** and, (ii) the inclusion of the non-choice sample in first-stage of two-stage least squares, we also abstract from more than one program type, sample weighting, clustered errors, and censored regression.

[3] Information on β_0 that comes from the exogenous OLS part (L_{OLS}) of the likelihood function $L_{(PK)}$ (1) allows δ to be separately identified from β_0 in the endogenous part (L_{2SLS}).

which there is a set of identifying instruments associated with exclusion restrictions that meet the rank and order conditions for identification. Identification in PK arises from having observations on the dependent variable y_0 and the exogenous variables $\mathbf{x}$ for individuals that exogenously have no choice to obtain (credit) treatment. It is the number of observations without choice of treatment that critically affects identification – at the limit, if there are too few such observations in the data, the effect of treatment (credit) δ is not identified.

(B) The two additive parts of L_{PK} are log-likelihoods that go back decades in the econometrics literature. While RM has repeatedly characterized these as complex, this is not so, either separately or added together. Although the assumption of normality is made in constructing the OLS log-likelihood and the 2SLS likelihood, (1) linear regression provides **exactly** the same parameter estimates as OLS by maximum likelihood (L_{OLS}) no matter what the distribution of the errors, and (2) the assumption that the error terms are multivariate normal also need not be true to insure that the parameter estimates from L_{2SLS} are consistent and asymptotically normal (Davidson and MacKinnon, 1993, p. 641). *Stata* very easily estimates both L_{2SLS} and L_{OLS} with a single simple command but it does not estimate the sum of these two log-likelihoods (except using Roodman's *cmp* command). That limitation in *Stata* does not make this model complex.

The RM likelihood is

(2) $L_{RM} = L_{2SLS}(y_0\text{-}y_{fm}\delta\text{-}\mathbf{x}\beta_0,\ y_{fm}\text{-}\mathbf{x}\beta_{fm}\text{-}\ \mathbf{z}\pi)$ for everyone in the sample including those without choice

where $\mathbf{z}$ is a set of identifying instruments having parameters π. The $\mathbf{z}$ instruments, as RM explain, are interactions of $\mathbf{x}$ with a dummy variable for having the choice to participate. Both PK and RM estimate their models by maximizing their respective likelihoods.

There are two differences between the PK log-likelihood (equation 1) and the RM log-likelihood (equation 2). First, for those without credit program choice, RM estimate the determinants of the outcome y_0 (household expenditure) as part of the 2SLS likelihood even though credit is deterministically and uniformly zero for every observation so that ordinary least squares is unbiased and fully efficient. It is not clear what it means to do 2SLS over a subsample in which the variable that is treated as endogenous only takes the value of deterministic zero. In contrast, PK treat the determinants of y_0 for those exogenously excluded from credit as exogenous (in the OLS part of likelihood (1)). Second, in seeming contradiction to point #1, RM add instruments $\mathbf{z}$ based upon the exogeneity of choice. Given that they have treated exogenous observations as endogenous, they must do this in order for the impact of credit to be identified. In contrast, PK do not need to add instruments $\mathbf{z}$ because they treat exogenous observations as exogenous. There is no *a priori* reason to believe that maximum likelihood estimates of credit impacts based upon RM's approach using the 2SLS likelihood for all observations is any more robust to deviations from normality, as claimed by RM (p. 19), than PK's estimates that use OLS for some of the observations. Furthermore, PK's log-likelihood is much simpler because it does not need the artifice of constructing instruments **z,** and because it uses OLS rather than 2SLS for the exogenous observations.

Consider the implications of RM's estimating the parameters β and π from a first-stage credit equation sample that includes observations lacking the choice to participate. Since credit is deterministically zero (that is, credit is zero with a sample and theoretical variance of zero) for those without choice, how can these observations provide any information to the estimation of the first-stage parameters?[4]

RM do not establish if their estimator (2), which treats deterministic observations as if they exercised a choice of whether or not to participate, is a consistent estimator of the PK model or if it is even a sensible alternative econometric model. Below it is explained why it is likely to be neither. In what follows, it is useful to keep in mind that it is the specification of the first-stage equation that is the key difference between RM and PK, and that fully efficient estimates of the first-stage can be obtained by linear regression in both cases. Consequently, the comments below focus on the first-stage and the "predicted" credit estimated from the first-stage in a model is which there is gender-specific treatment, and abstracts from issues of censored regression and sample weighting. Consequently, these comments are not specific to the PK data only but apply in general to any dataset having multiple endogenous treatment variables in which some observations have been exogenously denied the choice to be treated.

(1) The RM first-stage regression predicts non-zero credit with positive sample variance for those who have deterministically zero credit. This is akin to predicting a non-zero probability that an infant will have graduated university in a model of schooling achievement that includes all ages. In comparison, the PK first-stage "predicts" zero credit with zero variance, in accord with the definition of deterministic, for those who are deterministically without choice.

(2)The RM first-stage will, in general, estimate different marginal effects than the PK first-stage for all of the exogenous variables. That is, the marginal effect of an increase in, say, land, on predicted credit is different when estimated with the RM setup (equation 2) than the PK setup (equation 1). Consequently, RM generates different predicted values for credit than PK even for observations with choice. As there can be no better estimates than a first-stage regression using only observations with choice, the RM method adds a possibly biasing noise to the prediction.

(3) If additional nonchoice observations are added to the dataset, the first-stage estimates of RM will change as if these observations added new information even though observations with deterministically zero credit necessarily contain no information on the determinants of credit behavior. The PK first-stage, which is a linear regression using only observations with choice, is unaltered by the addition of any number of nonchoice observations as they do not enter into the first-stage estimation sample.

[4] Another issue that arises from RM's inclusion of nonchoice households in the first-stage sample is that having wealthier, landed households in the first-stage RM can lead to bias because "Even with village-level fixed effects, bias can emerge when programs base their placement decisions on the characteristics of sub-village groups. …The village-level fixed effects will only control for the average characteristics of the village sample (RM, fn. 7)." This critique by RM is not valid for the PK model, but is valid for the RM setup. PK do not include landed, ineligible households in the first-stage equation where the village fixed effects are estimated -- PK only estimate the village fixed effects with the functionally landless sub-sample of the villages with programs. RM do include wealthy, landed household in their first-stage.

(4) The first-stage equations of RM differ from those of PK not only because RM includes nonchoice observations but also because they add the "instruments" **z** – the interactions of choice with the exogenous variables **x**. In the first-stage equation for female credit, adding the interaction of female choice (c_f) and the variables in **x** to a sample containing only households for which females have choice (the PK first-stage sample) is completely innocuous. This is because c_f**x** = 1***x** when females have choice since c_f=1 and thus c_f**x** and **x** are perfectly collinear and one or the other will drop out of the regression. If there is only one endogenous variable (female credit), then RM's approach of including nonchoice observations in the first-stage works as **x** will drop out of the regression and the RM and PK first-stages coincide. (The RM first-stage in this case has all zeros for the dependent variable of the nonchoice observations and all zeros for c_f**x**. The RM t-ratios will all be wrong since the regression just adds observations with zeros for the dependent and independent variables.)

However, if there are two (or more) endogenous variables requiring first-stage equations, for example one for female credit and one for male credit, then RM and PK will no longer agree. In general, parameters for c_f**x**, c_m**x** and **x** are all separately identified in the RM setup. These extra parameters are identified because RM (i) fit the first-stage regression to observations with deterministically zero credit**,** and (ii) c_f**x** and c_m**x** are not perfectly collinear.[5] Thus, in this two endogenous variable case, the RM first-stage has three times the number of independent variables as PK, plus it has the non-choice observations. As Pitt and Khandker (2012) report, the RM first-stage for the case of three program types times two sexes yields a data matrix of independent variables that is 58 times larger than in PK, and for the model that does not distinguish by program type (there are three program types) the RM first-stage regression has a data matrix that is over 19 times larger than PK.[6]

In summary, the RM first-stage equations have a very different specification than the PK first-stage equations even when estimated as single equations by ordinary least squares linear regression; that is, leaving aside all of the "complexity" of maximum likelihood estimation that concerns RM. Both models have the same second-stage.[7] RM's model, which differs by including non-choice observations in its first-stage equation and as up to three times the number of independent variables than in PK, should not be

[5] It is sufficient that for some households only females have choice to participate and some households only men have choice to participate for all the interactions to be separately identified.

[6] In the case of male participation in the Grameen Bank, for example, the first-stage equation used by RM (used to estimate column (1) of Table 5) 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, one of the three sources of program credit.

[7] Since the second-stages have identical variables, one can judge which model seems to make more sense by examining the first-stage equations alone. Does it make more sense to estimate the determinants of a behavior from a sample from whom this behavior is observed (e.g. pregnancies of women of reproductive age) or from a sample that also includes those from whom the behavior is deterministically zero (e.g. include children or men in the pregnancy equation)?

used to test for "instrument weakness" of the PK model.[8] The RM results are not a robust variant of the PK results that can be used to test the assumptions of PK, but rather something almost completely different. In particular, adding "weak" (deterministically zero) observations and tripling the number of exogenous variables by including the nearly collinear sets of **x**, c_f**x** and c_m**x**, will result in a model with weak instruments and biased estimates of credit impact.

Simulations presented in PK (2012) suggest that, in the RM setup, linear LIML is particularly unstable compared to plain 2SLS, and this instability affects the estimates of credit effects (the δ's) that RM present. For example, examination of the complete computer output for the estimates presented by RM in col (1) of their Table 5 shows 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.[9] In addition, while a negative R-squared is possible with instrumental variables estimation, it is not conceivable that this regression's (centered) R-squared of -381.4 reflects credible estimates, as it means that the residual sum of squares is 382 times larger than the sum of squares.[10] A comparison of these linear LIML estimates with standard 2SLS estimates, both modeled with the RM setup, seemingly have no relationship to each other (PK (2012), Table 6). This is no doubt a reflection RM's inclusion of households without choice to participate in a credit program in the first-stage equation of the determinants of choice. As a further clue to the misspecification in RM, in PK (2012) I use the standard 2SLS formula to compute t-ratios for the credit effects parameters that RM estimate by linear LIML and report in column (2) of Table 5 of RM. This is of interest because asymptotically the parameter variance of the two-stage least squares (2SLS) estimator is the same as that of the linear LIML variance estimator. A comparison of the parameter covariance matrices seems to constitute a general misspecification test. The linear LIML t-ratio for the female borrowing parameter (0.445) is reported by RM to be t=0.10 in col (2) of Table 5, as compared to the 2SLS t-ratio of t=10.63 for the same parameter (0.445).[11] From the female credit effect estimated by Roodman and Morduch in this regression, one might wrongly conclude that the problem with Pitt and Khandker (1998) is that they underestimate the strongly statistically significant and positive effect of women's credit on household consumption, as well as the t-ratio. The point estimate of the female credit effect found by RM is much larger than that found by PK -- indeed by a factor of 10.[12]

[8] Weak instruments bias instrumental variable estimates in the direction of the OLS estimates. Thus, RM's claim that the instruments are weak in PK seemingly implies that PK have in fact under-estimated the positive effects of female credit on household consumption expenditure.

[9] 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. All of these results are reported in Table 5 of PK (2012).

[10] 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.

[11] This is in accord with the findings of Bekker (1994) that show that that the asymptotic variance of LIML changes if there are many instruments, in particular, it gets larger if there are irrelevant instruments.

[12] Large differences between standard 2SLS and linear LIML are a consequence of the questionable RM setup and not the PK dataset. PK (2012) demonstrate that linear LIML generates outlandish estimates with simulated data for the RM model. Moreover, it is not only the (endogenous) credit program

2. Instability and Outliers

The discussion of outliers in the RM paper provides the reader with only a selective view of how the distribution of the data affects the PK results, purportedly “destroying” the PK results. RM drop the 16 largest values for consumption and then re-estimate the model using PK’s method (equation 1). There are a variety of issues with how this trimming of the data is performed by RM. First, the practice of dropping outliers in general is "pre-analyzing" the data, a form of sample censoring imposed by the investigator, or in this case the replicator. Censoring on the basis of an outcome may produce inconsistency as in the endogenous selection literature (for example, MacDonald and Robinson 1985). Second, RM seem to have found the worst case scenario and do not fully investigate how further trimming or other trimming strategies affect the results. Third, RM ignore the estimation and testing protocol of PK, in which the exogeneity of credit is formally tested and then imposed if not rejected (both individually and jointly) for the sex-specific credit variables. These latter two points are the most important.

The procedure outlined in PK is to estimate the instrumental variables model, perform tests of exogeneity, and then, if the null hypothesis of exogeneity is not rejected, to impose it by estimating a (weighted) linear regression by ordinary least squares (PK, 975). Table 5 in PK (p. 987) summarizes the statistical significance of the results for all six of the outcomes studied, and imposes exogeneity for the four outcomes for which it is not rejected. Dropping the 16 outliers identified by RM and following through with the protocol of imposing exogeneity if the null hypothesis cannot be rejected results in a clear non-rejection of exogeneity (t=0.98 for female credit) for the household consumption variable. Imposing exogeneity for household consumption results in positive female and male credit effects that are jointly significant (p=0.03). Columns (1) and (2) of Table 1, provide the estimates when the deleted number of high consuming households are larger. When more of the high consumption observations are deleted, the estimated (positive) female credit effect rises and statistical precision increases (t=2.18 for 100 deleted observations and t=3.56 for 1000 deleted observations.)

As RM note, the concern about normality pertains to the residuals in the regression, not the outcome variable, so it is appropriate to focus on the former. More generally, the discussion about outliers in econometric practice concerns residuals; that is, how well a model fits at each data point, and not dependent variables. RM claim (footnote 20) that dropping observations on the basis of residuals produces a similar graph (in which the two modes collapse) to dropping observations on the basis of the outcome, so why have they not used residuals as they are the more appropriate measure? Moreover, why not trim symmetrically and check on the sensitivity of results to the extent of trimming?

Columns (3)-(6) of Table 1 in this reply provide estimates using both ordinary least squares (OLS) and IV-ML (the PK method) with differing levels of symmetric trimming on the basis of the residuals calculated from the full sample. Symmetric trimming of only 16 observations from each tail results in non-rejection of exogeneity and the OLS estimates (indicated on the basis of the exogeneity tests) have statistically

parameters estimated by RM’s linear LIML setup for which the linear LIML t-ratios and 2SLS t-ratios are vastly different. This difference extends to the exogenous slope variables as well (PK 2012).

significant and positive (but smaller) credit effects for both females and males. Trimming more observations symmetrically results in larger estimated female credit effects and higher t-ratios. Symmetrically trimming 100 observations from each tail nearly doubles the female credit effects (with t=4.88). Symmetrically trimming 1000 observations from each tail yields female credit effects that are six-fold larger than trimming 16 observations (and with t=28.27).

Symmetric trimming of 16, 100, 500 and 1000 observations and re-estimation by PK's IV-ML (reported in Table 1 columns (5) and (6)) results in the non-rejection of exogeneity except in the case of 1000, implying that the weighted OLS estimates of columns (3) and (4) are preferred for those other cases. Since the exogeneity of female credit is rejected with symmetric trimming of 1000 observations (p=0.03,) the IV-ML estimates should be preferred. For this case, the IV-ML estimates for both female and male credit effects are positive and statistically significant (t=9.95 and t=2.44, respectively), and the point estimates for female credit is more than triple that of male credit, and very close to the PK estimates as originally reported.[13]

Finally, the RM motivation for trimming is to reduce skew in the error that they claim leads to bimodality of the PK likelihood *and* estimator and biased estimates of credit effects.[14] I show in Section 4 (below) that non-normality can only be an issue for those observations with choice to participate as it is only those observations that are in the L_{2SLS} part of the likelihood in which there are "two stages." As equation (1) above makes clear, the determinants of the outcome (household consumption expenditure) for those without choice is simply estimated by ordinary least squares (OLS) in the L_{OLS} part of the likelihood function for which the distribution of the errors does not matter. In Section 4, I report that symmetrically trimming a total of 5 percent of observations from the tails of household consumption for those households with choice actually strengthens the PK results and leads to the non-rejection of normality for the subsample for which it might matter. If in addition the 16 high consumption "outliers" that most concern RM are also excluded (they are mostly non-choice observations), the parameter estimates of the credit effects (δ_f and δ_m) remain qualitatively unchanged.

In summary, the RM approach of deleting a single selected set of observations and then obtaining estimates that do not follow the protocol clearly described and implemented in PK cannot be seen to "destroy" the PK results. Rather, more defensible symmetric trimming methods, as well as asymmetric

[13] Consider top-coding (rather than deleting) high consumption observations and bottom-coding low consumption observations, one at a time from each tail and then re-estimating with ML, and repeat. The top-code sets the k largest values of consumption to that of the k-th largest, and analogously for bottom-coding. Doing this for 32 observations (that is, all 16 of the high consumption observations that RM delete), leaves the PK results almost unchanged. Continuing this top- and bottom-code procedure to larger numbers of observations results in uniformly positive female credit effects.

[14] RM have previously claimed that a bimodal likelihood violates the theory of maximum likelihood but in this version they now agree with PK (2012) that it is does not. Instead, they claim that "the apparent bimodality of the PK estimator, as distinct from the likelihood, is therefore a first-order concern." I do not know what the quoted statement means. Likelihoods can be bimodal, the bootstrap distribution can be bimodal, but maximum likelihood estimators cannot be. What is a bimodal maximum likelihood *estimator* and how is it distinct from the *likelihood*?

trimming of more observations, reinforce the PK results when estimates are obtained using the protocol clearly described and implemented in PK.

Having established that the source of identification in PK (1998) are observations without choice to join a credit program, now consider the interpretation of the sensitivity analysis of RM in Table 4, carried out with ML, in which they re-estimate the PK model over sub-sets of the data. Re-estimating the PK model with a sub-set of the data is not like breaking up a 2SLS wage equation sample into white and nonwhite sub-samples. Parameter identification arises from the deterministic zero credit observations (no choice to join) in the PK model (the second part of the log-likelihood given by equation (1)), so that dropping large shares of nonchoice observations, as RM do, eliminates a large share of the source of parameter identification. Essentially, RM demonstrate that when identification is weakened by the selection of a sub-sample, the model is less well identified. Why should that be surprising and what conclusion can be drawn? RM would like us to only draw the conclusion that the model is not well specified. In estimating the men's credit effect, for example, they throw out all of the observations from villages having a women's credit program. Doing so drops 63 percent of the (unweighted) observations without male choice, the source of parameter identification. This reduction in identifying restrictions arises not just as a consequence of a smaller total sample size and smaller nonchoice sample size, but also because two first-stage credit equations are no longer estimated jointly with the outcome (consumption) equation. Working with sub-samples may, in fact, be an enlightening exercise but only when the implications of dropping sources of identification is part of a broader discussion that informs the reader of the implications of doing so.[15]

3. "A missing discontinuity"

RM claim to "follow the advice of Imbens and Lemieux (2008) on preliminaries to discontinuity-based regression." The first thing Imbens and Lemieux (2008) suggest is to graph the average value of the outcome variable (household consumption expenditure) around the discontinuity in the forcing variables, and this is the plot that they are referring to when they state (Imbens and Lemieux, p.622) that "if the basic plot does not show any evidence of a discontinuity, there is relatively little chance that the more sophisticated analyses will lead to robust and credible estimates." Instead of the outcome graph, the RM paper includes a graph of the value of credit by gender against landholdings (e). Credit is not the outcome variable and landholdings is not the forcing variable. The forcing variables in PK are c_f and c_m, the interactions of the land threshold and program availability by gender in each village.

There are a number of issues with this simple exercise in addition to it not being the graph Imbens and Lemieux suggest be done first. First, contrary to the RM claim, the microfinance landholding rule generates a fuzzy Regression Discontinuity (RD). In the fuzzy RD design, the probability of receiving the treatment does not change from zero to one at the threshold. Instead, there is just a smaller jump in the *probability* of assignment to the treatment at the threshold (Imbens and Lemieux, p.619). Most households in our sample that are offered treatment (choice to participate in a microcredit program)

[15] It would be very useful if RM were to provide formal test statistics relating to the comparison of sub-samples rather than suggesting that the reader infer misspecification simply by eyeballing the table.

choose not to accept, so participation is not a deterministic function of eligibility. According to van der Klauw (2008, p.225) "this situation is analogous to having no-shows (treatment group members who do not receive treatment) and/or crossovers (control group members who do receive the treatment) in a randomized experiment."

Second, because this is a fuzzy RD setup, landholdings affect not only whether a household can join a credit program, but also whether it actually does and, conditional on joining, how much it borrows. Consequently, in a region around the land eligibility rule,[16] the local relationship of land to borrowing can be obscured by land's effect on whether a household participates and how much they borrow, not just if it can borrow. In addition, the other determinants of participation and borrowing in PK (a total of 80 variables that include education, age, household composition and the village fixed effects) are correlated with land.[17] By not controlling for these confounding covariates, this bivariate Lowess graph of borrowing versus land tells us little.[18] Moreover, RM offer no caveats and note none of these points in their discussion.

Third, we can test for the significance of the forcing variables c_f and c_m on borrowing conditional on landholding and the full set of confounding exogenous variables (including the village fixed effects) by standard regression. In order to determine if there is any discontinuity it is important to permit the effect of continuous landholding to be very flexible (that is, not overly smoothed) by including polynomials in land (Imbens and Lemieux 2008, p.624). Including a sixth degree polynomial in land yields significant and positive coefficients on the binary forcing variables female choice and male choice on female credit (t=3.29) and male credit (t=2.21), respectively, providing strong evidence of a discontinuity even when controlling for landholding in an exceptionally flexible way.

Fourth, RM choose a Lowess bandwidth of 0.8 and justify it by saying that it is the *Stata* default. Lowess in *Stata* provides locally weighted scatterplot smoothing for general use, not specifically as a method to test a fuzzy regression discontinuity design for discontinuities in a bivariate plot. No attempt is made by RM to use an optimal bandwidth or discuss the sensitivity of the results to this arbitrary choice. Imbens and Lemieux (2008) devote a great deal of attention to the choice of bandwidth and Imbens and Kalynaraman (2012) derive the optimal bandwidth choice for the RD estimator, yet this literature is ignored and the key parameter chosen by RM is left completely unjustified.

[16] The region is quite large in Figure 2 of RM as they specify a bandwidth of 0.8 which means that 80 percent of the observations are used for calculating smoothed values for each point in the data except for end points, where smaller, uncentered subsets are used.

[17] This is true even within program villages. The set of sixteen exogenous independent variables that exclude fixed effects are jointly significant in the determination of whether a household has female choice in villages with a female credit program (F(16,1256)=16.38, p=.0000) and analogously for males (F(16,1001=18.92, p=.0000). Thus, the exogenous covariates that determine both participation and outcomes (household expenditure) are correlated with the presence or absence of a program by gender.

[18] Note that PK estimate their model with controls for the confounding covariates over the full support of the data, not just "close" to the landholding eligibility point, which in that important sense differs from much of the regression discontinuity literature.

Fifth, the confidence interval bands in RM's Figure 2 are large around ½ acre simply because the local smoother has no data greater than or less than ½ acre in the two sub-samples defined by the ½ acre rule, and thus there are fewer observations used to smooth near the ½ acre limit. Just a glance at RM's Figure 2 confirms this. It is inconceivable that a graph presenting the Lowess smoothing results based on the entire sample would show a discontinuity in the variance around ½ acre. Thus, one cannot judge whether the differences around either side of the discontinuity are significantly different from each other simply by examining the confidence regions near ½ as RM direct the reader to do.

Sixth, illustrative of the issue discussed above, my Figure 1 below presents a smoothed graph of the outcome (consumption) on landholding created using Lowess regression done with Roodman's own code and choice of bandwidth. This is the graph that Imbens and Lemieux would suggest as the first diagnostic to be done if landholding were the only determinant of choice. The only change to the RM setup that I make is to separate the data at landholdings of 0.55 acres rather than 0.50 acres, an increase of only 200 square meters. As Pitt (1999) notes, the ½ acre limit refers to cultivable land and does not include the homestead (small plot on which the house is located). In that paper, I find that homestead land exceeds 0.05 acres on average in this sample.[19] RM have used cultivable plus homestead land. The result of this Lowess smoothing is a negative effect on household consumption at the discontinuity that is consistent with PK, and this discontinuity is large enough that the drop in consumption (as landholding increases beyond the discontinuity) seems to lie outside of the 95 percent confidence interval of consumption on the other side of the discontinuity. I do not wish to claim that a figure such as this necessarily demonstrates the correctness of PK. Rather it demonstrates the unreliable nature of the exercise performed by RM that does not deal with confounding covariates (including land itself) in a fuzzy RD design, uses smoothing based upon an arbitrary choice of bandwidth, and ignores the advice of Imbens and Lemieux that they claim to follow.

Finally, 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. As PK (2012) establishes,[20] consistent estimates of credit program effects are possible even when every household, no matter its land or wealth status, is treated as having endogenous treatment choice as long as it is in a village where any treatment is offered.[21] PK's (2012) estimate of program effects in which every household in treatment villages is considered eligible makes no qualitative change to the results – female credit effects are statistically significant and positive and male credit effects are not statistically different from zero.[22] Quantitatively, the estimated female credit effects are about 25 percent larger when no households are considered

[19] This should not be confused with the separate issue concerning the adjustment of land to "medium quality cultivable land." The quality issue is not addressed at all here (see Pitt 1999).

[20] First suggested in Pitt (1999).

[21] The non-choice households are restricted to those residing in villages with credit programs. The endogeneity of program placement is accounted for with village fixed effects.

[22] RM estimate the same model with essentially the same results (Table 3, col 4) but fail to note that a model in which all households in program villages have choice was previously estimated by PK (2012). If completely leaving out the one-half acre rule as a source of identification actually (in the words of RM) "*strengthens* the PK results for female microcredit" why do they continue to make it out as the linchpin to identification when it is not required at all?

ineligible and the precision of the estimates is greater. This finding effectively renders irrelevant this whole discussion of discovering a land discontinuity in the data.

4. Consistency, sources of identification, simulation and the "logarithm of zero"

RM's discussion of the PK method and its comparison to their linear LIML estimator raises more issues than the key one discussed in the previous section of this paper. In particular, RM state that linear LIML "is derived from a model that assumes normal errors in its formulation, but is consistent under substantial violations of that assumption" and that "the nonlinearities in the PK estimator turn out to make it less robust to such violations."

RM's claim about the consistency of their linear LIML model does not hold up under closer scrutiny under any distribution of the data. There are no exogenous slope (**x**) variables in any of the models that RM simulate. The only variables in the first-stage are intercepts, and as argued below, those intercepts should not be identifying instruments. Even with normal errors, the RM model is not consistent and the PK model is consistent if there are sufficient exogenous covariates and deterministic (non-choice) observations to identify the model. It is the existence of exogenous covariates that generates the plethora of interaction terms in the misspecified 2SLS setup of RM that make the RM instrument set weak and the linear LIML estimates invalid. As PK (1998) estimates a model with 254 free parameters, the RM simulation without any exogenous **x** slope variables is a seriously deficient model from which to make claims about the PK and RM approaches, particularly without making any mention of the crucial importance of these covariates.

To examine the claims of RM concerning the consistency of their approach and the inconsistency of PK's approach, Table 2 presents of estimates of credit effects (δ_f and δ_m) using Stata code similar to that used by RM (see Appendix 1 of RM) but now with 80 exogenous slope variables (covariates) rather than none.[23] The means of 100 replications estimated by RM's linear LIML method yield mean parameter estimates that are more than 60 percent larger than the true values of 1.00. The PK estimator applied to the same simulated data results in mean parameter estimates only 1 percent from the true value when the errors are normal or skewed and the first-stage is linear. In the case of a Tobit first-stage, the mean parameters are 1.00 when the errors are normal, and off only slightly when the data are skew in the way that RM specify. RM's very odd choice of simulating a model without any exogenous variables masks the inconsistency of their approach and the robustness of that of PK.

The general claim made by RM that having any skew in the second-stage error "violates" the PK model is clearly wrong on its face. As equation (1) above makes clear, the determinants of the outcome

[23] The 80 independent variables are generated as independent N(0,1) random variates. In the PK (1998) paper, the number of independent variables K is about 80. As the source of inconsistency in the RM approach is their treatment of household observations without choice, this simulation generates samples with a large share of non-choice households, and for which there are three program types, as in the actual Bangladesh case examined in PK (1998). This simulation thus mirrors the RM model used to estimate column 1 of RM's Table 5.

(household consumption expenditure) for those without choice is simply estimated by PK using ordinary least squares (OLS) in the L_{OLS} part of the likelihood function. As noted earlier, even though the assumption of normality is made in constructing the OLS likelihood, it is well known that OLS by linear regression provides exactly the same parameter estimates as OLS by maximum likelihood no matter what the distribution of the errors. Consequently, having skewed errors for those without credit program choice is of no consequence. Adding the level of skewness suggested by RM only to the nonchoice observations in the simulation model reported in Table 3 (nonchoice observations constitute about 80 percent of the sample), yields no evidence of bias whatsoever (last row of Table 2). Exogenous observations are treated as exogenous in PK, and treating nonchoice observations as if they have choice and are endogenous (as RM do) results in mistaken conclusions such as this.

RM's failure to understand that non-normal errors can only be an issue for observations with choice matters because RM test for skewness and excess kurtosis from predicted "second-stage" residuals for the full sample – choice and nonchoice – and apply that level of skewness to all of the residuals in their simulation exercise.[24] In addition, as noted above, they motivate dropping the high consumption observations from the PK Bangladesh data set, the 16 observations with the highest consumption in particular, that they call "the ones most responsible for the model-violating skew in the error," in an effort to reduce the "bias" from skewness. However, a (weighted) 86 percent of these 16 observations are non-choice observations and thus cannot be implicated in any bias arising from skewness. These high consumption (i.e., non-poor) observations are primarily non-choice because they are so well off they are ineligible for microcredit, and thus it is expected that their consumption would be at the right-tail of the distribution. It also suggests that the level of skewness and excess kurtosis is much lower in the sub-sample that matters, those with choice, than in the sub-sample most affected by the trimming strategy of RM – those without choice. Moreover, even if the errors in the choice sub-sample have skewness and excess kurtosis, we have shown that identification that arises from the inclusion of covariates in the model swamps any possibly faulty additional identification that may arise from specifying error distributions when the first-stage is Tobit, and which is demonstrated below with the actual Bangladesh data.[25]

If skewness, and non-normality in the error more generally, are not a source of bias when they occur in the non-choice (OLS) part of the likelihood, and since 86 percent of the 16 observations that RM drop are non-choice, then RM's trimming strategy clearly does little if anything to alleviate any bias that might arise from a non-normal error in the choice (2SLS) part of the likelihood. RM suggest that "model-violating" error skew is responsible for the supposed bias that they claim makes the PK results unreliable, but they have not trimmed the correct observations to test their claim. One crude test for

[24] PK (1998) estimate a single (second-stage) error variance for the choice observations and the non-choice observations. The test statistic for the null hypothesis that the variance is equal across sub-samples is $\chi^2(1)=0.71$, p=0.398. This test statistic was obtained by directly adding an extra variance term into PK's IV-ML estimation.

[25] In the PK model, when the first-stage is linear the only source of identification is covariates and model is more robust to error distribution (Table 2, row 2). The RM model is biased whatever the distribution of the errors.

any bias simply follows RM's procedure of dropping the 16 observations with the largest values for household consumption but only for those households with choice since it is only for these households that any non-normality in the error matters. Dropping the 16 highest consumption observations among those with choice has almost no effect on the credit parameters estimated with the PK IV-ML method. The female (δ_f) and male (δ_m) credit parameters effects with the trimmed sample are δ_f=0.043 (t=5.00) and δ_m=0.01 (t=0.48), respectively, as compared to δ_f=0.044 (t=4.83) and δ_m=0.01 (t=0.52), respectively for the complete sample. Using a more justifiable symmetric trimming strategy, trimming 2.5 percent of observations from each tail of the distribution of household consumption (80 observations per tail) for those with choice to participate in a credit program yields estimated female credit effects of δ_f=0.049 (t=8.87) and male credit effects of δ_m=-0.01 (t=-0.41) and the non-rejection of normality on the basis of tests of skewness (p=0.14) and excess kurtosis (p=0.24, joint test p= 0.17).[26] On closer inspection, the RM claim that a non-normal (skew) second-stage error is a significant source of bias in PK has no support in the data.

PK (2012, p. 24) highlight the additional bias arising from the use of linear LIML as compared to standard 2SLS in the RM setup even when the errors are normal. Simulated data sets with normal errors demonstrate that the linear LIML estimates, column (4) of Table 3 below (reproduced from Table 8 of PK (2012)), are clearly biased and imprecisely estimated (huge standard deviations) even in comparison to the bias of standard 2SLS. In addition, R-squareds are typically negative and large in absolute value when linear LIML is used, as is the case of RM's application of linear LIML to the PK data reported in RM's Table 5, column (1). RM's application of linear LIML is clearly not the "robust estimator" they they claim it to be.

The RM simulation exercise is lacking in another respect, the discussion of which sheds some light on the sources of identification and the "logarithm of zero" problem that RM highlight. In particular, RM state that the PK "simulations produce bimodality[27] only by deviating from the PK model and estimator in two major ways that weaken instruments. They simulate borrowings as averaging zero in the treatment group, so that average treatment is the same for borrowers and non-borrowers and credit choice is a perfectly weak instrument for treatment. And they *control* for credit choice rather than instrumenting with it. (RM footnote 22)." In the PK (2012) simulation, borrowings, which are treated as a continuous variable, do indeed have a mean of zero and credit choice is controlled for in the second stage and therefore is not an identifying instrument. RM fail to realize that by controlling for credit

[26] If in addition the 16 high consumption "outliers" that most concern RM are also dropped (they are mostly non-choice observations), the parameter estimates of the credit effects δ_f=0.39 (t=6.64) and δ_m=0.00 (t=0.01) remain qualitatively unchanged.

[27] The issue of bimodality is discussed at length in PK (2012). There PK show that bimodality is a "standard" feature of the PK likelihood 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. RM treat bimodality of the likelihood as if it were a diagnostic. Correctly specified models can have a bimodal likelihood (e.g. Phillips (1983) and Ferguson (1978)). It would be a major breakthrough in statistics and econometrics if a researcher finds a way to use bimodality as a diagnostic. However that is not likely given that correctly specified models can have a bimodal likelihood.

choice in the second stage, PK have made the estimates of program effects invariant to linear translation. Invariance to linear translation means that if any constant term is added or subtracted to credit (that is, to the value of the treatment of the treated) so that the credit treatment could have any mean value, the estimated program effects are unaffected.[28] Consequently, it is incorrect for RM to say that the PK simulation is deficient by choosing zero mean borrowing since the PK specification would have reported exactly the same results had the mean been one million or negative one million.

Why is invariance to linear translation important? In modeling the functional form for estimation, it is useful to convert the amount borrowed to logarithms so as to reduce the skew in the data, an issue of great concern to RM, and the method followed in PK (1998). In the PK simulations, consider the implications of the credit variables taking the form of logarithms of credit. In this case, there is then the problem of how to assign a value of the "treatment", call it log(α), to those who do not receive treatment (borrow) because they do not have choice. RM argue that this is done arbitrarily in PK who set α=1, so that log(1)=0. The problem is that a comparison of credit equal to 1 for the nonchoice observation with nonzero credit for those with choice is what helps identify the program effect, so this estimate may be sensitive to the choice of α. However, even if α=1000 were chosen instead of α=1, the estimates do not change in the PK simulation setup simply because PK controls for choice in the second stage equation; that is, it does not treat choice as an identifying instrument. If choice were treated as an identifying instrument, as RM suggest, and the mean of credit was translated via a logarithmic transformation as it is in PK, the RM linear LIML estimates would be biased. In essence, the choice dummy variables let the data decide the appropriate scale of (log) transformed treatment. Subtracting the log(1000) from the log of credit in taka, thereby rescaling credit to borrowings in thousands of taka, has no effect on PK simulations, but results in a huge bias in the RM simulations. In addition, the estimated credit coefficients δ are marginal treatment effects. In summary, (1) contrary to RM, program choice should not be an identifying instrument in the simulation, (2) a model with only choice as identifying variables, as in the RM simulation, is incorrect when the credit variable is interpreted as a logarithm, and (3) using a credit variable having a mean of zero is innocuous when the second stage equation controls for choice as the choice dummies perfectly adjust for the mean.

Although the PK (2012) simulation model is invariant to linear translation of the credit variable, the estimates presented in PK (1998) are not fully invariant for two reasons. Most importantly, the first stages are modeled as Tobit, which have zeroes even for those with choice. Second, PK (1998) controls for the two components of choice – eligibility (the "nontarget" variable) and presence of a credit program in the village (via village fixed effects) – but not the product of the two. Adding dummy variables for female and male choice to the PK second-stage does little to alter the estimates (not reported here), with t-ratios becoming larger. Nor does it matter much to rescale credit by dividing all credit by 10, which is equivalent to increasing the credit assigned to nonparticipators from log (α=1) to log(α=10) , a change in scale responsive to the untested view of RM that log(α=1) is "implausibly low." Doing so modestly increases the estimated female credit effect without changing the t-ratio even

[28] Since the model is also invariant to scale, invariance to linear translation makes it invariant to any linear transformation of credit.

though with α=10 the gap between the smallest credit treatment and no treatment falls by 90 percent and the variance of female and male credit falls by nearly half. [29]

5. The process of replication

In the first (2009) version of the RM paper, RM (2009) erroneously claimed statistically significant and negative female credit effects, the opposite of the PK findings, a finding that they advertised widely.[30] This finding came about because of what can only be called a typographical error on their part.[31] Before that error was discovered, they made many public pronouncements declaring that PK (1998) is faulty as a consequence of irreproducibility, instrument weakness, and non-normality; testified to a committee of the US Congress (and posted the video of that testimony on Youtube at http://www.youtube.com/watch?v=5Y10XbCLys8) on the supposed faults of PK; written a book for the popular market (Roodman 2011) with "Due Diligence" in the title that has as a central claim that PK and those who funded microfinance programs did not perform "due diligence" in their work. They have made very self-assured proclamations that PK is flawed, claiming (in David Roodman's Microfinance Open Book Blog: http://www.cgdev.org//open_book) that "…*academia has some explaining to do: first the most prestigious study says microcredit reduces poverty, then it is overturned [by RM] (posted March 1, 2010)…,*" that the RM (2009) paper is "*the academic equivalent not of a citation but an indictment... It is a long document packed with logic and evidence that the flaws are not merely possible but provable in academic court and important enough to generate wrong results (posted March 1, 2010) ,*" and that

[29] Increasing α to 20 (reducing the "gap" by 95 percent) has very little further impact on the results. RM also report that turning credit into a 0-1 variable, another way of obtaining linear scale invariance albeit with a loss of information, also "corroborates PK's results (RM, p. 12)." In the absence of full invariance to linear translation, which is not possible in this model, the choice of log(α) must be regarded as a functional form assumption similar to the choice of a logarithmic transformation. It is therefore reassuring that the PK results are relatively insensitive to the choice of α or to "probitizing" (dichotomizing) the treatment.

[30] They also widely advertised the results of Sargan tests that they claim refute causality in PK. They eventually agreed with me that these tests were flawed (see Pitt (2011a)), and then placed their emphasis instead on similarly flawed Hansen J tests.

[31] The correct variable 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 is in none of these, but is only buried in the Roodman code. Consequently, the typographical error by Roodman led everyone astray for some time because what they actually did was not what they say they did or what they thought they did. 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, as they clearly understood what the correct variables are or they would not have listed them correctly in the tables of their papers or in their email. 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 not publically distributed until two and one-half years after their paper was distributed, is in the form of a binary file written in Microsoft SQL format.

> [the] message that a lot of research published in prestigious journals is wrong *does* carry over to economics in general and microfinance in particular. Cases in point are the papers that Jonathan and I replicated (posted January 6, 2011).[32]

Finally, in a blog entry title titled "Taking the Con Out of Econometrics," Roodman writes:

> "… how could the economics profession have gone so wrong for so long? …the old research is fundamentally suspect and the new much better (though hardly perfect). The fancy math in what was once the leading study of microcredit's impacts is, though beautiful, typical of the old generation in its propensity to obscure rather than resolve the fundamental barriers to identifying cause and effect (posted March 28, 2010)."

Replicators have every right to make the results of their research known and to challenge academic conventions. But perhaps statements as definitive as these should wait until after the referee and review process has been completed (all of these comments by Roodman were written before I had even posted my first reply). Otherwise the replicators appear to be wedded to an outcome via their public pronouncements, and their objectivity as the process proceeds appears somewhat diminished. The serious issue of specification searches ("data mining") in empirical economics applies as well to mis-specification searches in replications (and apparent 'data selection' and 'model selection'). The work of replicators is important and it is their responsibility to take extra care in the quality of their methods and the claims they make (e.g. Pitt 2012).

5. Summary

This response paper refutes the claims of Roodman and Morduch that were not addressed in the earlier World Bank working paper of Pitt and Khandker (2012). This response paper, like the Pitt and Khandker (2012) paper and others that preceeded it, shows that many of the Roodman and Morduch claims are based on a flawed econometric understanding and a lack of due diligence in formulating and interpreting statistical models.

One claim highlighted by RM is that the PK results are not robust to deviations from normality of the second-stage errors, and that this non-normality is an important source of bias in PK. The determinants of the outcome (household consumption) for those without choice to participate in a credit program is simply estimated by PK using ordinary least squares and, as is well known, does not require normality. Roodman and Morduch, having improperly stated the distributional requirements of the errors, proceed

[32] Roodman's view that even publication in prestigious journals does not signal the credibility of a paper apparently does not carry over to the RM paper that is the subject of this response. Upon RM's acceptance in *the Journal of Development Studies,* Roodman posted in at least two different microcredit blog sites that "The acceptance is milestone for Jonathan and me, for it represents a ratification of our work, and is very long in coming."[my emphasis](found in http://www.financialaccess.org/blog/2013/10/milestone-great-debate-over-microcredit-impact-study and http://www.usfinancialdiaries.org/blog/2013/10/milestone-great-debate-over-microcredit-impact-study)

to drop high consumption observations as a *cure* even though they are predominately in the OLS part of the model and thus are not part of the error term that matters – the error associated with observations that have a choice to join a credit program. Furthermore, this paper demonstrates that symmetrically dropping 2.5 percent of the sample with choice from each tail of household consumption (i) actually strengthens the PK results, and (ii) cannot reject normality of the error for those with choice.

RM provide evidence from simulations that seemingly demonstrate the consistency of their approach as compared to that of Pitt and Khandker. Their simulation is for the unrealistic special case in which there are no exogenous variables at all in the model, only a constant. In Pitt and Khandker it is the exogenous variables that, coupled with non-choice observations, are ultimately the source of identification and since there about 80 exogenous independent variables in every equation PK estimate, this is a very peculiar specification by RM with which to examine the econometric properties of the PK results. This peculiar choice is better understand by knowing that it is the independent variables that are the source of bias in the Roodman and Morduch specification. It is the existence of exogenous covariates that generates the plethora of interaction terms in the misspecified 2SLS setup of RM that make the RM instrument set weak and the linear LIML estimates invalid. This was discussed at length in Pitt and Khandker (2012) and yet RM continue to ignore the issue. This issue is key in that using the asymptotically correct standard 2SLS formula to compute t-ratios for the credit effects parameters that RM estimate by linear LIML and report in column (2) of Table 5 of RM yields statistically significant and large positive effect of women's credit on household consumption.

Although RM follow an incorrect strategy to deal with non-normal second-stage errors, their dropping ("trimming") of high consumption observations does alter the IV-ML results of PK but does not follow the PK protocol. The RM approach of deleting a single selected set of observations and then obtaining estimates that do not follow the protocol clearly described and implemented in PK cannot be seen to "destroy" the PK results. Rather, more defensible symmetric trimming methods, as well as asymmetric trimming of more observations, reinforce the PK results.

The paper also refutes the results of Roodman and Morduch concerning the "logarithm of zero." PK convert the amount borrowed to logarithms so as to reduce the skew in the data. There is then the problem of how to assign a value of the "treatment" to those who do not receive treatment (borrow) because they do not have choice. RM fail to realize that the inclusion of a choice dummy variables lets the data decide the appropriate scale of (log) transformed treatment. I show that the qualitative results of PK are robust to altering the value of the "treatment" provided those who do not participate in a credit program.

This response paper also notes a number of errors in the RM's search for a discontinuity in PK regression discontinuity design. There are a number of issues besides that they do not actually follow the method of Imbens and Lemieux that they say they follow. First, contrary to the RM claim, the microfinance landholding rule generates a fuzzy Regression Discontinuity (RD). Land-holdings affect not only whether a household can join a credit program, but also whether it actually does and, conditional on joining, how much it borrows. Consequently, in a region around the land eligibility rule, the local relationship of land to borrowing can be obscured by land's effect on whether a household participates and how much they borrow, not just if it can borrow. I present a smoothed graph of the outcome (consumption) on landholding created using Lowess regression done with Roodman's own code and choice of bandwidth that contradicts the RM claim of no discontinuity. In addition, as PK (2012) establishes, consistent estimates of credit program effects are possible even when every household, no matter its land or wealth status, is treated as having endogenous treatment choice as long as it is in a village where any treatment is offered. Doing so leaves the PK results intact and effectively renders irrelevant this whole discussion of discovering a land discontinuity in the data.

The point of replication should not be to search in every conceivable direction for an empirical specification that changes ("destroys" in the words of RM) some of the results while disregarding other evidence in support of the original specification. The RM replications are so full of flaws from their very first paper in 2009 to their published version that, when coupled with their numerous and early public pronouncements declaring the presumed faults of PK and the "obscurantism" of any program evaluation that is not a randomized control trial, the credibility of the entire replication exercise is called into question.

References

Bekker, P., 1994. "Alternative Approximations to the Distribution of Instrumental Variables Estimators," Econometrica 62, 657-681.

Davidson, Russell and James MacKinnon 1993. Estimation and Inference in Econometrics. Oxford University Press.

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.

Imbens, Guido and Karthik Kalyanaraman 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," Review of Economic Studies (2012), 1–27.

Imbens, G.W., & Lemieux, T. 2008. "Regression discontinuity designs," Journal of Econometrics, 142, 615–635

MacDonald, Glenn and Chris Robinson 1985, "Cautionary Tails About Arbitrary Deletion of Observations," Journal of Labor Economics 3: 124-152

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

Phillips, P.C.B.. 1983. Exact small sample theory in the simultaneous equations model. In M.D. Intriligator and Z. Griliches (eds.), Handbook of Econometrics, Volume I, chapter 8, 449-516, North-Holland, Amsterdam.

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." Available at http://www.brown.edu/research/projects/pitt/.

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, Dec. 2012, 48:12, 1886-1891. Expanded version 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.

Pitt, M.M. and Khandker, S.R. (2012) Replicating Replication: Due Diligence in Roodman and Morduch's Replication of Pitt and Khandker (1998). Working Paper 6273. World Bank.

Pitt, Mark M., 2014. Response to 'The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence', Journal of Development Studies, 2014 (in press).

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. 2014. The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence, Journal of Development Studies.

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.

van der Klaauw, Wilbert. 2008. "Regression–Discontinuity Analysis: A Survey of Recent Developments in Economics," Labour 22 (2) 219–245.

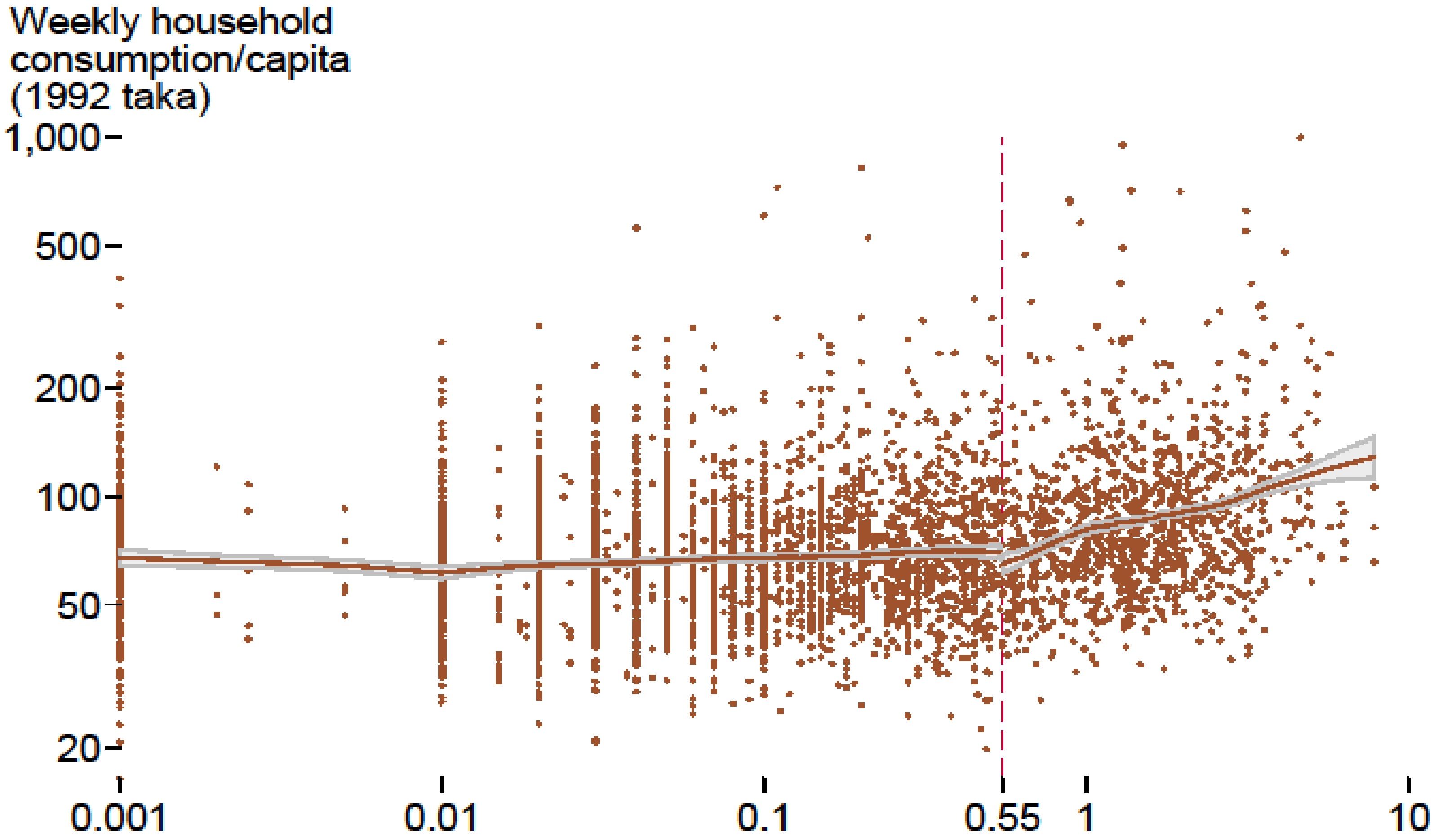

Figure 1: Household landholdings before borrowing (acres)

Table 1. A more complete picture of the PK results when observations are trimmed

(full sample = 5218, asymptotic t-ratios in parenthesis)

	Trim N largest consumption observations		Symmetrically trim N largest and N smallest residuals			
	Weighted OLS		Weighted OLS		Instrumental variables by nonlinear ML	
N (observations trimmed)	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)
16	0.0040 (1.91)	0.0054 (1.98)	**0.0051** **(2.51)**	**0.0056** **(2.07)**	-.0144 (-1.00)	.0133 (1.14)
100	0.0042 (2.18)	0.0043 (1.71)	**0.0089** **(4.88)**	**0.0038** **(1.60)**	.0193 (1.26)	.0115 (1.37)
500	0.0050 (2.87)	0.0041 (1.98)	**0.0207** **(14.52)**	**0.0060** **(3.42)**	.0292 (3.56)	.0063 (0.81)
1000	0.0055 (3.46)	0.0034 (1.78)	0.0306 (28.27)	0.0094 (7.24)	**.0386**[a] **(9.85)**	**.0141** **(2.44)**

[a] Exogeneity rejected (p=0.034) when N=1000 observations are trimmed from each tail. Consequently, the estimates in bold are the ones that correspond to the PK protocol.

Table 2. Simulated coefficients with and without skew in the second-stage error

	Normal error		Skewed error	
	δ_f	δ_m	δ_f	δ_m
RM Linear LIML	1.60 (0.050)	1.60 (0.053)	1.61 (0.054)	1.61 (0.052)
PK (linear first-stage)	1.00 (0.067)	1.01 (0.064)	1.00 (0.88)	0.99 (0.089)
PK (Tobit first-stage)	1.00 (0.051)	1.00 (0.051)	1.05 (0.081)	1.04 (0.083)
PK (Tobit first-stage, skew in non-choice errors only	n.a.	n.a.	0.99 (0.051)	0.99 (0.053)

Note: Means of 100 replications with standard deviations in parenthesis. True parameters are 1.00 in every case. Skew and normal second-stage errors are as specified in RM (2014). The δ_f and δ_m values for each simulated data set are themselves the means over the three program types.

Table 3 . 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. Source: PK (2012), Table 8.