July 21, 2012

Gunfight at the NOT OK Corral: Reply to "High Noon for Microfinance" by Duvendack and Palmer-Jones (Uncut version)

by

Mark M. Pitt
Brown University

ABSTRACT *The paper "High Noon for Microfinance Impact Evaluations" by Duvendack and Palmer-Jones replicates the papers of Chemin (2008) and Pitt and Khandker (1998) that estimate the impact of microfinance in Bangladesh. My paper replicates the Duvendack and Palmer-Jones replication and finds so many serious errors in their code and misrepresentations of the methods described in their paper that I conclude that their results are spurious and provide no evidence about the validity of either the papers of Chemin or Pitt and Khandker or on the effectiveness of microfinance.*

"High Noon" is the name of a classic Western movie in which the showdown between rivals is scheduled for precisely noon and is a term that has come to describe a confrontation that definitively resolves an ongoing conflict. In this case, Duvendack and Palmer-Jones (henceforth DPJ) challenge the results of Pitt and Khander (1998) and Chemin (1998) and promise an analysis that will finally decide whether or not microfinance has positive effects on the lives of the poor. They are to be saluted for taking on such a daunting task.[1] However, this paper (named after another classic Western "Gunfight at the OK Corral") shows that the DPJ findings are far from OK. There are so many serious errors in their code and misrepresentations of the method applied that their results have no meaning or interpretation and thus provide no evidence about the validity of either the papers of Chemin or Pitt and Khandker or on the effectiveness of microfinance.

[1] The authors are among the authors of the well-publicized DFID report titled "*What is the evidence of the impact of microfinance on the well-being of poor people?"* (Duvendack *et. al*, 2011) that claims to have screened 2643 research papers and reports on the impact of microfinance, and reviewed 58 of these studies in detail. The DFID study protocol states that both Duvendack and Palmer-Jones were responsible for statistical analysis and replication in that report. Their study concludes that "impact evaluations of microfinance suffer from weak methodologies and inadequate data." The Pitt and Khandker (1998) study receives particular criticism in the DFID report, with the authors arguing that it "is not robust based on the replication of PnK done by RnM [Roodman and Morduch] and Duvendack and Palmer-Jones (2011) [the DPJ paper]," and that they (Duvendack and Palmer-Jones) have "replicated the key studies related to PnK and applied PSM as well as sensitivity analysis concluding that PnK's original findings cannot be confirmed." Indeed, a number of papers of Duvendack alone or in co-authorship with Palmer-Jones are cited often throughout the DFID report to back up the claim that other studies, particularly those by authors Pitt, Pitt's students, or Khandker, are deficient. So an examination of the quality of the keystone DJP replication, which I attempt here, surely also reflects the quality of that DFID report and its conclusions.

DPJ replicates both the papers of Chemin, published in this *Journal*, and Pitt and Khandker. In the very first page of their paper, DPJ correctly note that replication is an unappreciated and poorly rewarded task in the social sciences and that "the returns to finding problems in papers could be high for society; policies which are legitimated in large part by iconic studies which are subsequently shown to not to lead robustly to [sic] the conclusions for which they are known, could lead to different research or policy conclusions with high social benefits" and then declare in the very next sentence that the "iconic" study they are referring to is Pitt and Khandker: "In this article we present evidence that undermines an, if not the, iconic study which legitimated for much of the past two decades the belief that microfinance (MF) is beneficent for the poor, especially when targeted on women." Specifically, DPJ claim that they "fail to confirm PnK's [Pitt and Khandker] original findings of beneficent outcomes caused by MF [microfinance] using PSM [propensity score matching] with these data." Pitt and Khandker's most cited result is that women's participation in group-based credit programs for the poor has a large and statistically positive effect on total household consumption expenditure, while men's participation has no effect, in a population where women were the vast majority of credit program participants. The propensity score matching evidence DPJ present surely contradicts Pitt and Khandker by claiming that microfinance "participants are worse off than individuals in control villages spending 0.4 per cent to 6.5 per cent less" (p. 9). In the summary of their replication, DPJ state that their results "do not corroborate PnK or Chemin," suggest that Pitt and Khandker of used "sophisticated analytical methods to compensate for weak research design," and offer the advice that "policymakers would have been well advised to have placed less reliance on PnK."[2]

This paper is a replication of their replication. It is not meant as a defense of Pitt and Khandker. Nothing in this paper should be seen as making a case that Pitt and Khandker is either correct or incorrect. Rather, this paper's objective is simply to examine whether the paper by DPJ is a reasonable application of propensity score matching to these data, and whether it should be accepted as contrary evidence to the Pitt and Khandker results. I offer no general critiques of the propensity score matching method, however sorely I am tempted. I examine only the quality of the research methods and data manipulation that DPJ use in doing propensity score matching. As noted, I have discovered so many serious errors that I can only conclude that the DJP results are entirely spurious. I lay out their errors in some detail in the form of 18 points so that readers can judge for themselves the quality of the DPJ paper. As DPJ so rightly point out: "Doubts can be cast on the creativity of those who replicate, and also on their motivation, which might include casting doubt on the integrity or ability of the original authors."[3,4]

---

[2] DPJ offer criticisms of the propensity score matching approach they choose to employ but apparently do not find those criticisms sufficiently important to lead them to refrain from drawing the numerous negative conclusions about Pitt and Khandker that are quoted here. In public forums describing their replication work on microfinance, Duvendack and Palmer-Jones have extended their critique as follows:

> *We also comment on the inappropriate prestige and spurious scientificity awarded to micro econometric analyses which replication can do much to re-calibrate.*
> *http://eadi.org/gc2011/default_104.html*

[3] Richard Palmer-Jones is particularly aware of the issues and ethics of replication. In his online biography posted as part of the *3ie-LIDC Seminar Series - Replication Studies Symposium,* Palmer-Jones notes that he is engaged in

Specific points

(1) This paper, as well as that of Chemin, treats treatment choice as individual-specific in estimating treatment propensity, but, unlike Chemin, does not exclude from the estimation sample those with no empirical probability of treatment. DJP do not mention either in the paper or as part of Table 1 that their estimations of the probability of microfinance participation are at the individual level. Moreover, nothing is said about who was included in the sample, and thus there is no discussion of what inclusion criteria make sense. As it turns out the author's sample of 9,397 is everyone in the full sample except those with missing data. It includes newborns and a 98-year old, plus those for whom the probability of treatment is zero because the option of treatment (credit program) did not exist in their village.

Consider the issue of age. In the DPJ sample a total of 4,199 individuals, or 44.7 percent of the estimation sample, are under the age of 16, and not a single one of these individuals is in the treatment group. This sample includes 464 individuals zero or 1 year of age. The empirical probability of treatment for those under 16 is zero, and consequently they cannot help identify any of the treatment model's parameters. It is not possible that if a one year old had enough livestock or a low enough wage,

---

examining "the ethics of quantitative and qualitative research conduct and publication." The International Initiative for Impact Evaluation (*3ie*) has commissioned Palmer-Jones (who also sits on the *3ie* Replication Program Advisory Group) and Vegard Iversen to write a detailed history of and best practices for replication. The day-to-day manager of the Replication Programme, Ben Wood, has written that *3ie* will "rely on their (forthcoming) working paper to define the stages of internal replication for our program: pure, statistical, and scientific." Palmer-Jones is also co-organizer of '*As well as the subject: additional dimensions in development research ethics*', a funded project that claims on its web site that

> "The questions that motivate this initiative derive in part from concerns with reporting of quantitative work where errors in data processing or analysis and/ or data 'massaging' are difficult to detect in the review process. This concern is embodied in the growing, but not always popular, practice of replication of quantitative analyses **requiring timely availability of data and code from authors in order to understand the details of what has been done** [my emphasis]." (http://www.uea.ac.uk/dev/ethicalanalysis)

The responsibilities of Palmer-Jones are of some interest here not only because it seems to afford him some credibility as a replicator, but also because he and Duvendack have refused my request for the code that generated the dataset used in the estimation of the DJP paper so that I could make use of it in writing this replication of their work.

[4] There also concerns here are about the process of replication that Duvendack and Palmer-Jones have followed not just with this paper but other papers reporting replications of my work. I knew that they had written at least one such replication, but not because they ever sent me any of their manuscripts, or otherwise communicated with me on any matter concerning this paper or any of their other papers that replicate my work. (The only email exchange we have had prior to this was a request for data from Duvendack that I responded to, but the data that I sent her has not been incorporated into a paper as far as I am aware.)

Very recently and quite by chance I became aware that this replication paper was forthcoming in *the Journal of Development Studies*, was accepted for publication, and had been published on-line since April. I contacted the Administrative Editor asking why I had not been consulted for a reply or as a referee. One of the Managing Editors wrote me to say that the Managing Editor handling this manuscript (there are four Managing Editors, one of whom is Palmer-Jones), had not requested any input from the authors of the two papers that the Duvendack and Palmer-Jones paper replicate, neither as referees or authors of a reply. In response to my concerns, that Managing Editor has withheld the Duvendack and Palmer-Jones paper until this reply and that of Chemin were received.

he or she could be induced to join the Grameen Bank. If the key variable *age (in years)* were introduced as age ranges, i.e., 0-7, 8-15, 16-23, etc., the age effects for almost half the sample would go to minus infinity in a logit or probit regression; that is, they would not be identified. Clearly, nothing can be learned about the propensity for treatment from those who cannot be treated. This is analogous to including females of all ages, including infants, in estimating the determinants of recent fertility. In this fertility example, adding powers of age in years may seem to give a very good fit for the binary outcome "birth in the past year," but it is not meaningful because children and old women are not at risk for a birth. Note that in the DPJ preferred specification in col(2) of Table 1, there are only 5 variables out of 59 that vary across individuals within a household, and 3 of these 5 are age and polynomials in age. The others are the sex dummy, and marital status. Absent age, sex, and marital status, everyone in the household would have the same propensity score, and the nearest neighbor match for any treated person would be another person in the same household. As we explain below, it does not matter whether members of a treated household are actually matched with members of the same household who are considered controls; rather, what matters is whether members of treatment households are used as controls at all.

Then there is the issue of treatment choice. One possible benefit of the propensity score matching (PSM) method is that it does not require the use of eligibility criteria based on landownership in order to identify the effect of treatment, as in Pitt and Khandker (1998). However, leaving aside eligibility, if there is no treatment offered in a village, and village residents can only participate in groups in their village, then there is a deterministic zero propensity to participate irrespective of the value of the independent variables in the logit participation equation. There are 1,509 individuals in the sample who reside in control villages, which were purposely drawn in the original sampling scheme because there were no credit programs and no one in the village participated in any of the credit programs. As with one year olds, these individuals have a zero empirical probability of treatment. All of the estimates of Table 1 claim to include village dummy variables, which is puzzling, given that the village dummy variables for any of the 15 villages with no individuals accepting treatment will go to minus infinity. In addition, only relatively few villages have credit groups for men. Indeed, 42.7 percent of males in the sample (2,057 individuals) reside in villages without a male credit program that they can choose to join, and 27.6 percent of females (1,264 individuals) reside in a village without a female credit program.[5] Two out of 2,057 males and 14 out of 1264 women still report themselves members of a credit program. Chemin recognizes the problem and restricts the sample to "individuals with less than 0.5 acres" and "on the sample of individuals in villages with microfinance." If this is a replication of Chemin, not using his sample selection rule invalidates any comparison.[6] More importantly, treating thousands of

[5] Pitt and Khandker (1998) allow credit treatment effects to vary by sex and by program (there are three programs), for a total of six treatment effects. The key finding of Pitt and Khandker is the large difference between female and male treatment effects. For the key measure of household welfare total consumption expenditure, Pitt and Khandker find that female credit treatment effects are positive and statistically significant but male credit treatment effects are not different from zero. The propensity score matching estimates of Chemin and those of Tables 2 and 3 of DPJ do not differentiate by the gender of the person treated and are thus possibly confounding positive female credit effects with zero male credit effects.

[6] When comparing their model to that of Chemin, DPJ state that "The number of observations differs as well for reasons we cannot explain."

observations that lack choice to join a credit program as if they do have choice will seriously bias the DPJ results.

In addition, Pitt and Khandker use household-level data for both the probability of treatment and the outcome, so neither Chemin's nor DPJ's approach can possibly be a replication of Pitt and Khandker. Chemin claims that his propensity score model in his Table 1 (reproduced in DPJ as Spec. 1) is that of Pitt and Khandker, and DPJ also refer to their "Spec. 1" as the "PnK" model. Pitt and Khandker's participation equation (Pitt and Khandker, Table 1, page 979) clearly states that the sample size is 1,195 households in the female credit equation and 895 households in the male credit equation, as compared to 4,215 observations in Chemin and 9,397 observations in DPJ.[7] In addition, as the Pitt and Khandker credit equations are at the household level, there are no variables "Sex" and "Age" in their specification. Chemin added those (and DPJ followed) on his own initiative yet still referred to it as the Pitt and Khandker specification. Finally, not all of the variables in the actual Pitt and Khandker specification made it into Chemin and subsequently DPJ. In short, this credit participation specification barely resembles that of Pitt and Khandker.

(2) The most serious repercussion of the DPJ approach to using individual data in examining the impact of treatment on household variables is that it wrongly assigns individuals to the control group when they are members of treatment households. Every one of the outcome variables that DPJ examine are measured at the level of the household, not the individual. In the DPJ paper, the propensity score matching assigns individuals, rather than households, to either treatment or control even though households have multiple members, with a mean household size of about 6.4 members.[8] That means that if one household member is treated the others are treated as controls even though the value of every single outcome variable in the study (such as per capita expenditure and children's school enrollment) is exactly the same for each of them. As an example, consider the hypothetical case where treatment raises household per capita expenditure by 100 taka. The per capita household expenditure of every member of the household is, by definition, higher by 100 taka. However, DPJ consider all of the members of the household *except* the person who belongs to the credit programs as not having been treated (that is, as controls), and these individuals are then matched with individuals in other households who have been treated and received the 100 taka treatment effect. Since both the treatment and control individuals, as defined by DPJ, have had their household per capita expenditures raised by 100 taka (because in both households someone has been treated), the treatment effect estimated by DPJ is zero whereas it is in reality 100. In the sample as a whole, there are 3,441 observations that DPJ assign to the control group that actually are members of a treatment household, making up 40.6 percent of all controls and almost 45 percent of the poorest two strata in the DPJ analysis. More importantly, all of these observations are, by construction, in villages with a credit program. In treatment villages these incorrectly assigned observations make up 74.2 percent of

[7] Roodman and Morduch, whose prior paper replicating of Pitt and Khandker is cited by DJP more than two dozen times, clearly understood that households are the appropriate units of observations.

[8] 89 percent of the households that have a treated member have only one treated member, and only 0.71 percent have more than two treated members.

controls. This latter figure is the most relevant since, as we show below, DPJ seriously mis-specify the propensity score for individuals residing in the 15 control villages.

(3) DPJ inaccurately claim that they include village fixed effects in the estimation of the treatment propensity equations (Table 1). In fact, they only include fixed effects for thanas (upazilla). There are three villages in each of the 29 thanas in the sample. In the Stata estimation code (*.do file*) that Duvendack sent me, the dummy variables included in the logit equations for microfinance participation are labeled with "thanaid" plus the thana number, and there are 23 of them. The data set that they sent me has the village identifier in it (variable named *upzvill*) and their footnote 13 suggests they know the difference between a thana and a village. In any event, in the rest of this paper I will refer to "thana fixed effects" rather than "village fixed effects" since the former is what was actually done, but even then not in all matching specifications.

(4) DPJ arbitrarily treat all observations from thanas 25 through 29 as actually being located in thana 1, the first thana in the sample, when estimating the logit equations for microfinance participation (Table 1), thus making invalid the propensity scores used for matching. The issue is that it is impossible to estimate thana (or village) fixed effects for thanas in which not a single individual participates in the credit program treatment. Those coefficients on the thana dummy variables would go to minus infinity. Recognizing this, Chemin does not include any households from control thanas in his propensity score estimation sample. However, DPJ include all 1,509 individuals from these thanas and assign them the thana-specific intercept of thana 1. They do this by including dummy variables for thana 2 through thana 24. If the estimation sample only includes the treatment thanas (thanas 1 through 24), then, as is well known, one of the 24 thana dummy variables needs to be dropped from the set of independent variables and the thana-specific intercept of the omitted thana is the constant term. The results do not depend on which one is dropped, and when Stata is asked to form the dummy variables with the *xi* command, it automatically drops the first indicator dummy variable, which in this case is the dummy variable for thana 1.[9] As none of the thana dummy variables ever have the value "1" for any of the observations from thana 1 and thanas 25 through 29, they all share the constant term as their thana-specific intercept. This is, of course, completely arbitrary, but it is not noted in the paper. There is absolutely no reason to believe that the three thanas (15 villages) that are control thanas have the same thana intercept as the arbitrarily ordered first thana of the treatment group. Moreover, as noted above, observations from these control thanas cannot contribute anything to estimating the propensity score equation.[10]

[9] This is the exact command used by DPJ: xi i.thanaid. This command creates thana dummy variables for all thanas 2 through 24. They are clearly aware of this since they subsequently list out the dummy variables included in the logit equations as:
global treatvilldumm "_Ithanaid_2 _Ithanaid_3 _Ithanaid_4 _Ithanaid_5 _Ithanaid_6 _Ithanaid_7 _Ithanaid_8 _Ithanaid_9 _Ithanaid_10 _Ithanaid_11 _Ithanaid_12 _Ithanaid_13 _Ithanaid_14 _Ithanaid_15 _Ithanaid_16 _Ithanaid_17 _Ithanaid_18 _Ithanaid_19 _Ithanaid_20 _Ithanaid_21 _Ithanaid_22 _Ithanaid_23 _Ithanaid_24"

[10] The analogous equation might be an individual propensity to use an intra-uterine device. Putting men in such an equation does not assist in identifying the determinants of behavior by those at risk for the treatment, who are by definition women. Moreover, a dummy variable for "male" would have a coefficient of minus infinity.

Does this error matter? First, adding observations without choice of treatment (that is, those whose treatment status is deterministically zero) necessarily biases the coefficient of any logit or other regression. To get a handle on how large this bias might be, consider the special case where the observations having treatment choice and those with (deterministic) lack of choice (such as those from the control villages or children) are both randomly drawn from a population. The bias in this case is in proportion to the share of deterministic controls in the sample. One can see that the bias could well exceed 50 percent, given that, as explained above, 44.7 percent of the estimation sample are under the age of 16, not a single one of these individuals is in the treatment group, an overlapping 16 percent of the sample are from control villages where no one has participated, and the sample includes a huge share of males who live in villages where treatment is only available for women. The bias will not be proportional if the share of observations without choice varies across households or thanas, and where thana and household variables are important in the propensity score, as they are in DPJ. Second, inappropriately including observations from control villages into the logit estimation of the propensity scores by forcing them to look like they came from a particular treatment thana is likely to do even more violence to the estimates because it makes the bias non-proportional. For example, what happens to the mean propensity score of individuals residing in thana number 26, one of the five control thanas, calculated from the second specification of Table 1 (the preferred specification of DPJ)? If the *omitted* thana in the fixed effects estimation is thana number 6, the mean propensity score for individuals from thana 26 would be 44.5 percent larger than when the omitted thana is thana number 4.[11]

(5) The second and third specifications of the propensity score function include independent variables that are clearly caused by treatment rather than are determined prior to the treatment decision, violating a necessary condition for the validity of the PSM method. Simply put, the covariates X in the matching equation must constitute a "set of pretreatment covariates" (Dehejia and Wahba, 2002). Matching estimators require the *Conditional Independence Assumption* that given a set of observable covariates *X* which are not affected by treatment, potential outcomes are independent of treatment assignment (*unconfoundedness*). This implies that selection is solely based on observable characteristics and that all variables that influence treatment assignment and potential outcomes are simultaneously observed by the researcher. Only variables that are unaffected by participation (or the anticipation of participation) should be included in the set of propensity score covariates X.[12] These variables should either be (i) fixed or nearly fixed over time, such as education or sex; (ii) perfectly predicted, such as age; or (iii) measured prior to being offered treatment choice. Including covariates that are the consequence of treatment results in biased matching estimates.

With that in mind, consider that the main treatment provided by the three microfinance programs studied (Grameen Bank, BRAC, and BRDB-12) is the provision of credit for self-employment enterprises

[11] Even when DJP estimate a propensity score equation without the control thanas, as they apparently try to do for Table 3, they impute the thana-specific intercept of thana 1 as the intercept of the control thanas 25 to 29. Perhaps they are unaware that when one uses the Stata command *predict* after a logit or other regression, Stata calculates the predicted value over the entire dataset, not just the sub-sample that was used for the logit regression. I say "apparently" above because they include one of the five control thanas (thana 25) in that propensity score logit model.

[12] This is clearly noted in the paper by Caliendo and Kopeinig (2005) that DJP cite.

among the poor. The most common activities involve livestock, most often *milch* cows.[13] Hossain (1988) in Table 21 and 22 (pages 40 and 50) provide the breakdown of loans by purpose and use. The other broad categories besides "Livestock, poultry raising and fisheries" are "Crop cultivation," "Processing and manufacturing," "Trading and shopkeeping," and "Transport and other services." In addition, all three microfinance program have mandatory savings requirements. Considering the nature of these microfinance programs, how can it be argued that variables such as "have self-employment enterprise," "livestock value," "savings," "agricultural income," "revenue of non-farming enterprises," "dairy product sales," "expenses of non-farming enterprise," "transport assets," and "equipment assets" be considered pre-treatment? The whole point of these microfinance programs is to improve the lives of the poor by altering these very characteristics. As these variables are all measured in the data post-treatment, how is it possible that variables such as "have a non-farm enterprise," "dairy product sales" or "savings," among others, are unaffected by treatment if the nature and intent of treatment is to affect them?

How does the inappropriate inclusion of variables that are clearly the targets of the microfinance treatment affect the matching results? Consider two households prior to treatment, where one possesses the assets required to undertake a self-employment activity (a *milch* cow) and the other does not, and that operating this self-employment activity with those assets increases household expenditure. Further assume that, aside from the self-employment activity and its assets, their pre-treatment covariates are identical. Now the treatment is offered. The assetless household accepts the treatment and borrows funds to finance the same quantity of asset (a *milch* cow) as the household with the cow and operates the self-employment activity as efficiently, and as a consequence, their household expenditure rises to the level of the pre-treatment asset-holding (cow owning) household. DPJ, by using post-treatment covariates to match on treatment propensity, would consider these two households as identical as far as match covariates. They would then claim that microfinance had no impact since they have the same expenditure. The nature of the microfinance treatment is to generate the self-employment activity and the required asset, so once it is essentially netted out of the impact evaluation, why should any impact remain? DPJ would have us believe that acquiring a *milch* cow as a result of the treatment simply tells us who is likely to be treated, and otherwise has no effect on household income or other outcomes.

(6) Variables in the propensity score function that are commonly understood to be individual-specific are constructed so that they do not vary among members of a household and further contribute to the misspecification of the model. For example, the variable labeled "Education" in Table 1 of the DJP paper never varies among members of households. There are infants with more than 14 years of "education." Indeed, almost two-thirds of infants under the age of two have one or more years of "education" and more than half of them have more years of schooling than they are old. It would seem that the "education" variable in DPJ is actually the education of someone in the household. Exactly who is not clear since I do not have access to the do files that created their variables.

What about the variables "mother still alive" and "father still alive"? These variables also never vary among household members. Either everyone in a household has a dead mother, or no one does. The variables "father's education" and "mother's education" also do not vary among members of the same household. Ditto for "injury."

[13] In these data 878 of 2,089 loans from microcredit (42 percent) were for *milch* cow. The annual reports of the Grameen Bank persistently have *milch* cow raising at the top of their list of loan value by use.

The variable "highest grade completed" is a recoding of the maximum of "highest grade completed" within the household. In this recoding, the value 0 corresponds to no schooling, 1 corresponds to 1 to 5 years, 2 corresponds to 6 to 10 years, 3 corresponds to 11 or 12 years, and 4 corresponds to more than 12 years of schooling or to a missing value for schooling. Consequently, those who attended university are given the same value of "highest grade completed" as those whom a missing value was recorded, many of whom are small children. In fact, the single largest value in the data (37 percent) is for code that indicates either "attended university" or missing, with missing making up 96 percent of those.[14]

As far as I can tell, of the 59 independent variables in specification 2 of Table 1 (which is the specification subsequently used in the estimation of treatment effects), only age, sex and marital status vary at the level of the individual. This is certainly unexpected and not at all made clear or explained in the paper.

(7) Agricultural and non-agricultural wages in the propensity score function do not vary within a household but vary within villages and are often zero. These variables are in fact not wage rates but something else that I cannot determine. These variables often take the value of zero. Wage rates were collected as part of the community (village) survey component that augments this household survey, and Duvendack and Palmer-Jones include them in the data set but not in the estimation. Unable to make sense of the two wage variables, I emailed Duvendack who sent back some Stata code specific to these variables along with the statement that "It is clear that these two variables do not correspond to any meaningful variable" and are in error.[15]

---

[14] These data are not really missing. Only those who had completed schooling had schooling levels recorded in this schedule of the questionnaire. Those who were still in school had their schooling recorded in a different schedule. Thus the" missings" mostly indicate that the individual had not yet completed their schooling. As the "highest grade completed" variable was used as a linear variable and not as a set of categorical dummy variables, this error of giving the highest value to those who have attend university or are still in a school at any level, is truly a mis-specification.

[15] My desire to make sense of these variables led to my request (on 7 July 2012) for the Stata code used to build the estimation dataset. The DJP paper claims that "A complete set of our Stata code is available from the authors to run with the data that can be downloaded from the World Bank together with additional data we can supply, and instructions on how to organise the data" on page 5. Duvendack replied the next day on behalf of herself and Palmer-Jones saying that they "prefer not to share the full data preparation code until *High Noon* is in print in its final form." Their paper was published online 27 April 2012, two and one-half months prior to my request. There is a striking inconsistency between advocating for the "timely availability of data and code from authors in order to understand the details of what has been done" (http://www.uea.ac.uk/dev/ethicalanalysis) as Palmer-Jones (and Duvendack) do, and denying the request of someone who they know is preparing a replication for publication in the very same issue in which their paper will be published. Publication of DJP was delayed by one of the Managing Editors of the *Journal* just so my and Chemin' s comments on DJP would appear in the same issue as the DJP paper. DJP are fully aware that providing me their Stata code after the print publication date of their paper will mean that anything I learn about it will not appear in this replication paper. In addition, I cannot make use of the recently posted data construction code used by Roodman and Morduch whose estimation variables DJP seek to "triangulate," and for the same reason DJP say they cannot. The Roodman and Morduch code is in Microsoft SQL binary format which the result that DJP could not examine it, saying in footnote 17 "we were not tempted (able) to borrow code from RnM or take their results as correct" which is curious because elsewhere Palmer-Jones claims

(8) The p-values shown below the estimates of Table 1 are not based upon clustered standard errors. Clustering should be done at the level of the household, as the mean household size exceeds six and none of the members of a household is excluded from the estimation sample. It is not credible that the unobserved variables affecting who in a household joins a credit program are independent across its members. Chemin notes that he uses "robust t-statistics," which is understood to mean Huber-White heteroskasticity-robust standard errors, although he does not say whether his t-statistics have been clustered at the household level. DPJ use neither robust t-statistics nor clustering. This difference between Chemin and DPJ would help explain why DPJ claim to find 10 more variables "significant" in their logit propensity score estimation than Chemin did.

(9) DPJ wrongly define strata in estimating the impact of credit program participation on household expenditure adding yet another source of misspecification to their methods. The idea of PSM is to compare treated observations with untreated observations whose propensity scores are close. The idea of partitioning the *common support* of the estimated propensity score in order to define groups for matching originates with the seminal paper of Rosenbaum and Rubin (1983), who referred to it as sub-classification but which has subsequently been referred to most often as stratification matching. The average treatment effect on the treated which should then be calculated as the weighted (by the number of treated) average of the strata-specific treatment effects, which are computed as the difference in average outcomes of treated and controls within the same strata. There is no doubt that when the literature refers to stratification matching, it is referring to strata defined by the propensity score. It is this kind of stratification that Chemin used.[16] However, DPJ create strata based upon the outcome, total household consumption expenditure, rather than on the propensity score. I am unaware of other PSM studies in the social sciences that use the outcome as the index for stratification. Moreover, the method that they implemented does not compute the strata-specific treatment effects as the difference in average outcomes of treated and controls within the same strata, but rather estimates a separate matching equation for each stratum and then computes the treatment effect with single nearest-neighbor matching algorithm within the outcome-based strata. This very peculiar method of impact evaluation is not described at all in the paper, is not what Chemin did, and does seem to make any sense.[17] I can only presume that even the authors are unaware that this is what they have

---

that with respect to the Roodman and Morduch code "I can attest, both code and data have been and are readily and fully available and easy to use – indeed RM are exemplary in this respect." (http://blogs.cgdev.org/open_book/2011/03/response-to-pitts-response-to-roodman-and-morduchs-replication-of-etc.php)

[16] That Chemin understood stratification correctly is clear from his statement on page 475 that "three strata are used for the Stratification technique: 20, 10, and 5 (the propensity score being between 0 and 100)."

[17] Forming strata from the outcome variable when all of the outcome variables are household-specific, that is when they do not vary among members of the same household, necessarily means that all of the members of a household are in the same stratum. Consequently, the difference in the means of all outcome variables between the treated and the controls within a stratum is a comparison of treated individual with all of the other individuals within the same household.

done, since their code seems to be an adaptation of the example code that is provided in *Stata*'s *psmatch2* help document.[18] [19]

The problem with stratifying on outcomes rather than on pre-treatment covariates is that if the microfinance treatment is effective in increasing expenditure, treated households will be moved into higher expenditure strata and so will look poorer in comparison to the controls in that higher expenditure strata. As a rather extreme example, if a poor person is "treated" with a lottery prize of $1,000,000 they will move into the much-discussed "1 percent." However, even $1,000,000 will not make them richer than the average of their new peers if that average income is $10,000,000. The DPJ matching method would declare that the treatment effect of winning the lottery is to reduce income, in this hypothetical example by $9,000,000.

(10) Tables 2, 3, and 4 claim that the significance levels of the estimates of average treatment effects were obtained from bootstrapping, but they were not. The *Stata* code did not do any bootstrapping. When I emailed Duvendack about this she responded that this was an error that carried over from a previous draft[20], saying that they decided to drop their bootstrapped statistics in favor of non-bootstrapped estimates because they became aware of the paper by Abadie and Imbens (2008). That paper argues that bootstrap standard errors are not valid as a basis for inference with simple nearest-neighbor matching estimators with replacement and a fixed number of neighbors when matching on a covariate. Although the Abadie and Imbens result did not include matching using a kernel density or stratification, or the case of matching on a propensity score, it does cast some doubt on bootstrapping as a means of obtaining the asymptotic variances of matching estimators. However, it is also clear that the significance tests provided by DPJ are not consistent because they do not take into account the nonzero variance of the propensity score, the very issue that bootstrapping is intended to deal with. In brief, the DPJ significance tests are not valid and the alternative tests based on bootstrapping are also probably not valid. There is no discussion of any of this in the paper.

[18] The example code in the *psmatch2* documentation is for the case where the strata are defined by a grouping variable such as regions or states, racial groups, or sex, where there is reason to believe *a priori* that treatments effects vary. For example, if the propensity to accept contraception or educate girls varies systematically with ethnicity or religion, as does the treatment effect, then the evaluation of a program made available to an entire population should disaggregate the analysis by ethnicity or religion but can still provide a national average treatment effect by calculating the weighted average of the ethnicity- or religion-specific treatment effects. That is what the example code provided by the Stata command *psmatch2* is intended to do. It will also do the correct form of stratification as a matching algorithm if the grouping variable is based on the propensity score and differences in mean outcomes within strata are calculated.

[19] If DPJ stratified by total household consumption expenditure because they actually believed that the structure of propensity scores and the magnitude of treatment effects varied by expenditure classes, it would be expected that they would then show how treatment effects vary by strata. They do not.

[20] The DJP paper published online by the *Journal of Development Studies* lists July 2011 as the date of receipt of the final version by the *Journal*. Other papers by these authors written after that date, one as recently as 3 months ago, also claim that the test statistics that they present were obtained by bootstrapping, suggesting they may have forgotten the Abadie and Imbens result. The method used to compute test statistics should be consistent across papers.

(11) The t-statistics on weighted average treatment effects estimated from stratification matching are done incorrectly, and vastly overstate the statistical significance of the results. DPJ do not provide standard errors or t-statistics in their tables reporting estimates of treatment effects, just asterisks indicating significance at the 10, 5 and 1 percent levels. Some of the ones that are calculated with their *Stata do file* are spectacularly large in absolute value. For example, the t-statistic is -11.57 on the estimated treatment effect of microfinance on household expenditure when treated individuals are compared to non-participants in treatment villages (third row and third column of Table 2) and is -19.04 when treated individuals are compared to individuals in control villages (fourth row and third column of Table 2). The p-values associated with these t-statistics are so small that they challenge the machine precision of personal computers -- statistically significant indeed. Unfortunately, the calculation is completely wrong.[21] They calculate the t-values for the stratification estimates after first removing all of the variation within each strata – that is, they treat the average treatment effect within each strata as having zero variance and assign the strata average effect to every treated person in the strata. They then calculate the t-statistic for the reported average treatment effect for the entire sample as the t-statistic on the constant term of a regression of the treatment effect on only a constant term. However, there are exactly 5 values in the data used in the regression, one value for each strata, consisting of 921 individual observations (the number of treated individuals in the sample) as everyone in a strata is assigned the same average treatment effect. The t-statistic for each strata is thus treated as if they were infinity.

(12) The stratification estimates of Table 2 are actually based on probit propensity scores, not logit propensity scores. This should not matter much in practice. However, as only logit propensity scores are discussed in the text of the paper and only logit propensity score equations are presented in Table 1, this is a bit of an oddity.

(13) The stratification estimates comparing treated individuals to individuals in control villages that are presented in Table 2, row 4, have little resemblance to the methods described in the paper and have no apparent logic. Leaving aside the issues of stratification on the outcome rather than the propensity score, the use of separate propensity score estimation equations for each strata, and the inappropriate construction of the t-statistics, the DPJ estimates for the last row of Table 2 are based on propensity score models that are unlike the propensity score models presented in Table 1, unlike the models estimated to calculate the treatment effects reported in row 3 of Table 1 where the controls are drawn from treatment villages[22], and unlike those of Chemin. The strata-specific samples get very small in this setup. There are only 2,069 total observations comprised of 921 treated individuals from treatment villages and 1,148 control observations from control villages. This sample is divided into as

[21] This error in constructing the standard errors would have been obvious to readers of the paper if standard errors or t-statistics had been part of the results reported in the table. A set of asterisks can never be thought of as an adequate replacement of either standard errors or t-statistics when reporting econometric estimates.

[22] Strictly speaking not all of the controls in Table 2 rows(3) and (4) are from treatment villages as DJP claim. Inexplicably, the three villages in thana 25, one of the five control thanas, are considered as treatment villages in row(3) , and they are assigned the same dummy variable as thana 1 in the propensity score estimation.

many as 20 strata.[23] The resulting strata have as few as 65 valid observations. The probit models used in the comparison in which the controls come from the treatment villages specify 37 independent variables but most of these variables have coefficients that are not estimable with these data. Importantly, DPJ drop the thana fixed effects from all of the strata-specific probit propensity score models, and make no mention of it.[24] Other coefficients were unidentified even after dropping the 23 thana fixed effects. Consider the probit propensity score equation results presented in Table 1 for the poorest strata and the fifth poorest stata when 20 strata are defined. A quick glance at the table reveals that something is very wrong. Some of the standard error/t-statistics are dots ".", meaning they are not estimable. All the estimated t-statistics are 0.00 except for one that is 0.01. Many of the coefficients are clearly outlandish – a unit change in the variable "No adult male in HH" increases the normal index z by over 85 standard deviations. Finally, the pseudo-R^2 both equal 1; that is, the independent variables completely explain the treatment/control dependent variable. Yet it is on the basis of these fatally flawed probit regressions that DPJ calculate propensity scores and the average treatment effect.

(14) Contrary to the statements of the DPJ paper, DPJ's stratification treatment effects are not comparable to the kernel matching treatment effects reported in Table 2. The kernel matching treatment effects do not in fact use Chemin's specification 3 for the propensity score equation as DPJ state at the bottom of Table 2. Instead, DPJ use Chemin's specification 2 (DJP's specification 3) but leave out all of the thana fixed effects without informing readers.[25] Plus, the propensity score equation used in kernel matching is a logit while that used in stratification matching is a probit.

(15) The comparison of female and male treatment effects suffers from additional design issues. In the female treatment effects propensity score model, not all of the treated are females, although most of them are. The propensity score (logit, this time, and again without the thana fixed effects) has a dummy variable for sex that would go to minus infinity if all of the treated were female, but it is does not. The same is true for the male treatment effects propensity score model.

(16) No attempt is made to allow for the choice-based nature of the sampling frame used with these data, nor is this issue even discussed. The sampling frame for the dataset is choice-based. This is explicit in Pitt and Khandker as well as in the subsequent replications of Roodman and Morduch. Credit program participants were oversampled relative to their frequency in the population of persons eligible for microfinance. Under choice-based sampling, the weights, which are available in the dataset, are generally required to consistently estimate the probabilities of program participation. When the weights are unknown, Heckman and Todd (2009) show that with a slight modification, matching methods can

[23] The strata are not of equal size since DPJ use all 9397 individuals in forming the strata based on total household expenditure. Those equal size strata are no longer of equal size after all the control individuals from treatment village, 78 percent of the sample, are dropped.

[24] Indeed, they had no choice. None of the thana fixed effects are identifiable even for an infinitely sized sample given the sample selection scheme that they use. In the sample used, there are only treated but no controls in the first 24 thanas, and only controls but no treated in the last five thanas. This lack of identifiability is simply a consequence of the stratification procedure that they adopted.

[25] However, the kernel estimates of Table 3 do include the thana fixed effects. Model specification varies from table to table and from one estimate to another within a table and these variations are neither described or defended, as if the this were part of a specification search.

still be applied, because the odds ratio (P/(1 − P)) estimated using a logistic model that ignores the choice-based nature of the sample is a scalar multiple of the true odds ratio, which is itself a monotonic transformation of the propensity scores. Therefore, matching on the odds ratio (or of the log odds ratio) will eliminate the need for weights. In the special cases of nearest neighbor or stratification matching, it does not matter whether matching is performed on the odds ratio or on the propensity scores (estimated using no weights), because the ranking of the observations is the same. This nice result does not apply to all kernel matching methods, or to local linear regression, since in many kernel density implementations the absolute distance between propensity scores matters.

(17) DPJ do not follow Chemin's strategy of removing the village fixed effects from total household consumption expenditure (and other outcomes) from a prior regression, as Chemin explains in the quote below, making their replication attempt even less comparable. Chemin says on page 475:

> *To make sure that results do not come from systematic differences across villages, the logarithm of per capita expenditure is removed from village effects by regressing this quantity on village dummies from both the programme and control villages only, and then estimating the residual arising from this regression. This quantity is termed the 'pure' logarithm of per capita expenditure since it is now freed from any village level effects.*

(18) Finally, do all of the issues raised above matter for measuring the impact of microfinance using propensity score matching? Although wrongly assigning individuals to the control group when they are members of treatment households will clearly negatively bias the measured impacts, there are so many other significant errors it is impossible to predict the sign or magnitude of the full set of errors. Keeping in mind that the paper you are reading is not meant to be a defense of Pitt and Khandker but rather a replication of Duvendack and Palmer-Jones, I have nonetheless done propensity score matching to estimate the female credit treatment effect on total household consumption expenditure using my version of the data (posted online as a Stata file on my Brown University website in early 2011) using the same specification that was used in the Pitt and Khandker 1998 paper. No new variables were introduced or taken away. In order to include village fixed effects and have a sample in which the empirical probability of treatment was not zero, only villages with a credit program for females are included. Both kernel density and stratification (on propensity scores) matching was performed using a logit model. "Common Support" was imposed on all of the matching estimates. In addition, alternative estimates in which the top and bottom 1 percent and 2 percent of propensity scores were trimmed from the kernel estimates so as to reduce the influence of outliers. In addition, to deal with sampling weights, a model using the odds ratio as the propensity score was used in the untrimmed kernel density case. The results using the odds ratio are identical to those where probabilities are used, since, it would appear, the default Epanechnikov kernel (as well as the normal kernel) in *psmatch2* is based on ordering of neighbors and not distance in the propensity score metric. No sensitivity tests were carried out (aside from the trimming) as the point here is simply to see what would happen if DPJ had replicated Pitt and Khandker without the myriad errors that they have introduced into their paper. The t-ratios were obtained from 100 bootstrap replications, although the Abadie and Imbens (2008) result noted above is likely to apply. The results in Tables 2 and 3 indicate positive and highly significant results of women's

participation in microfinance in every one of the matching estimators. The results are also very robust to changes in bandwidth, trimming, and the number of strata, as well as dropping the last two rounds.[26]

Conclusion

This replication of DPJ has identified a number of serious errors in their work and identifies some differences between the way they have represented their methods in the paper and the methods that were actually used. The extraordinarily large number of mistakes and misrepresentations in the DJP results demonstrate that their work completely lacks credibility and calls into question their other papers using the same methods with the same data. My experience with DJP leads me to offer one suggestion about assessing the validity of empirical papers in economics. While it is certainly very helpful that authors of papers make their code and data available, something that has very recently become the norm in economics, it is even more important that authors fully and accurately describe their methods in their paper. Drawing examples from DJP, if a paper says that logit is used, then it should not use probit some of the time; if it says that the errors are bootstrapped and village fixed effects are used, then it should be careful that that is the case; if it uses data for all individuals zero to 98 years old, that should be mentioned in the discussion of the data as it is not evident from means and standard deviations; and if variables in a dataset of individuals are generally considered attributes of individuals (such as "education" or "mother still alive") then it should be noted when they are not. In any paper, and particularly in a replication, clarity within the paper about the rules for sample inclusion and what methods are actually used are critical if one is to suggest that the work of others does not stand up. One should not have to judge the quality of an empirical paper only by working through the computer code of its authors. As this paper demonstrates, a sober assessment of the DPJ estimation code (my request for data construction code was refused) reveals serious discrepancies between the "pure, statistical, and scientific" approach to replication and what this paper delivers.

[26] Data and *Stata* do files can be found at http://www.brown.edu/research/projects/pitt/.

References

Abadie, Alberto and Imbens, Guido W. (2008) On the Failure of the Bootstrap for Matching Estimators. Econometrica, 76(6), pp. 1537-1557.

Caliendo, M. and Kopeinig, S. (2005) Some Practical Guidance for the Implementation of Propensity Score Matching. Forschungsinstitut zur Zukunft der Arbeit (IZA) Discussion Paper No. 1588, May.

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

Dehejia, R. H. and Wahba, S. (2002) Propensity Score-Matching Methods For Nonexperimental Causal Studies, *The Review of Economics and Statistics*, 84(1), pp. 151-161.

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

Duvendack, M., Palmer-Jones, R., Copestake, J.G., Hooper, L., Loke, Y. and Rao, N. (2011) *What is the Evidence of the Impact of Microfinance on the Well-being of Poor People?* (London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London).

Heckman, J. and Todd, P.E. (2009) A Note on Adapting Propensity Score Matching and Selection Models to Choice Based Samples*. Econometrics Journal*, 12(s1), pp. S230-S34.

Pitt, M.M. (2011a) Overidentification tests and causality: a second response to Roodman and Morduch. Accessed from: http://www.brown.edu/research/projects/pitt/

Pitt, M.M. (2011b) Response to Roodman and Morduch's 'The impact of microcredit on the poor in Bangladesh: revisiting the evidence'. Accessed from http://www.brown.edu/research/projects/pitt/

Pitt, Mark and Shahidur Khandker (1998)"The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter? *Journal of Political Economy*,106(5), pp. 958-996.

Roodman, D. and Morduch, J. (2009) The impact of microcredit on the poor in Bangladesh: revisiting the evidence. Center for Global Development, Working Paper No. 174, June.

Rosenbaum, P., and Rubin, D. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika 70:1, pp. 41-55.

Table 1. Probit propensity score regression: Controls are individuals in control village, strata 1 and 5 of 20 strata

	Estimates exactly as reported by Stata					
	Strata 1			Strata 5		
Variable name	coefficient	standard error	t-statistic	coefficient	standard error	t-statistic
Sex	4.89517	.	.	-11.3589	.	.
Age	1.266421	888.2112	0.00	5.696786	.	.
Age of HH head	2.487847	177.0743	0.01	-.6131282	.	.
No adult male in HH	85.49982	.	.	-25.72199	.	.
Education	5.822617	.	.	-.9526161	.	.
Savings	.0420196	11.46605	0.00	.1125512	.	.
Own a non-farm enterprise	28.18035	.	.	-8.369398	.	.
Livestock value	.000601	9.658245	0.00	-.0004301	12.5037	-0.00
HH size	-1.144668	.	.	2.732317	.	.
Non-agricultural wage	.2986457	94.5047	0.00	.2277563	.	.
Agricultural wage	.1127328	97.2548	0.00	-.0236376	111.2708	-0.00
Age squared	-.0450965	28.65308	-0.00	-.1323896	17.80966	-0.01
Age to the power four	9.37e-06	.0052417	0.00	.0000153	.0127934	0.00
Constant	-199.051	.	.	-36.61422	.	.
No. of observations	65			91		
Pseudo R2	1.0000			1.0000		

Table 2. Propensity score matching estimates of the effect of women's credit on measure of household consumption: kernel density matches

(bootstrapped t-ratios in parenthesis)

	Kernel bandwidth			
	0.08	0.06	0.04	0.02
Common support, no trimming	0.0515 (3.45)	0.0512 (3.63)	0.0497 (3.60)	0.0443 (3.18)
Common support, trim 1%	0.0507 (3.42)	0.0469 (3.43)	0.0504 (3.25)	0.0446 (3.31)
Common support, trim 2%	0.0511 (3.39)	0.0515 (3.54)	0.0511 (3.66)	0.0458 (3.29)

100 bootstrap replications

Table 3. Propensity score matching estimates of the effect of women's credit on measure of household consumption: stratification matches

(bootstrapped t-ratios in parenthesis)

Number of strata		
5 strata	10 strata	20 strata
0.0476 (3.07)	0.0476 (3.48)	0.0458 (2.55)

100 bootstrap replications