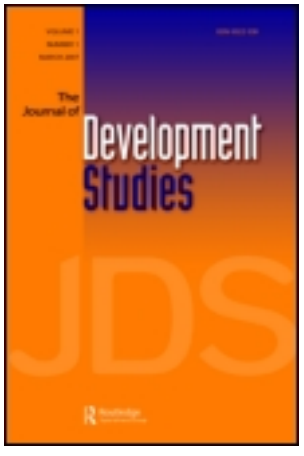

The Journal of Development Studies
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/fjds20

Gunfight at the Not OK Corral: Reply to 'High Noon for Microfinance'

Mark M. Pitt [a]
[a] Brown University, Providence, RI, USA
Published online: 18 Dec 2012.

To cite this article: Mark M. Pitt (2012) Gunfight at the Not OK Corral: Reply to 'High Noon for Microfinance', The Journal of Development Studies, 48:12, 1886-1891, DOI: 10.1080/00220388.2012.727563

To link to this article: http://dx.doi.org/10.1080/00220388.2012.727563

Journal of Development Studies,
Vol. 48, No. 12, 1886–1891, December 2012

Gunfight at the Not OK Corral: Reply to 'High Noon for Microfinance'

MARK M. PITT
Brown University, Providence, RI, USA

ABSTRACT *Duvendack and Palmer-Jones claim to replicate Chemin (2008) and Pitt and Khandker (1998) but obtain different results and hence challenge the two papers' estimates of the impact of microfinance in Bangladesh. This response details a number of reasons to demonstrate that Duvendack and Palmer-Jones is not a replication so their results provide no evidence about the validity of either of the earlier papers 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.[1] In this case, Duvendack and Palmer-Jones (henceforth DPJ) challenge the results of Pitt and Khandker (1998, henceforth PnK), and Chemin (2008) and promise an analysis that will finally decide whether or not microfinance has positive effects on the lives of the poor.[2] Pitt and Khandker's most cited result is that women's participation in group-based credit programmes 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 programme participants. The propensity score matching evidence DPJ present contradicts this finding, hence DPJ state that their results 'do not corroborate PnK or Chemin', hence their criticism of these papers. This article is named after another classic Western, *Gunfight at the OK Corral*, and shows that the DPJ findings are far from OK. There are so many unjustified differences in how they select the sample, code and measure variables and implement matching that DPJ cannot be considered a replication and thus provides no evidence about the validity of either Chemin or PnK or on the effectiveness of microfinance. This is in stark contrast to the claims about replication and criticisms of the earlier papers in DPJ.

This article is an assessment of DPJ to examine whether it is a reasonable application of propensity score matching (PSM) to the PnK data and whether it should be accepted as contrary evidence to PnK and, by implication, Chemin. It is neither a defence of PnK nor of PSM (one point on which all the authors may agree is that PSM is a method with many limitations); nothing in this article should be seen as making a case that PnK is either correct or incorrect. I lay out the errors or limitations of DPJ in the form of six substantive points and briefly mention 10 further issues so that readers can judge for themselves.

Correspondence Address: Mark M. Pitt, Population Studies and Training Center, Department of Economics, Box B, 64 Waterman Street, Providence, Rhode Island 02912, USA. Email: Mark_Pitt@brown.edu

ISSN 0022-0388 Print/1743-9140 Online/12/121886-06 © 2012 Taylor & Francis
http://dx.doi.org/10.1080/00220388.2012.727563

(1) DPJ, like 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.

The DPJ sample of 9397 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 programme) did not exist in their village. Taking the age example, 4199 individuals (44.7 per cent) of the DPJ estimation sample are under the age of 16, and cannot possibly be in the treatment group. The empirical probability of treatment for those under 16 is zero, and consequently they cannot help identify any of the treatment model's parameters. Note that in the DPJ preferred specification in column (2) of Table 1, there are only five variables out of 59 that vary across individuals within a household, and three of these 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 neighbour match for any treated person would be another person in the same household. It does not matter whether members of a treated household are actually matched with members of the same household who are considered controls, but it does matter whether members of treatment households are used as controls at all.

One possible benefit of the PSM method is that it does not require the use of eligibility criteria based on land ownership in order to identify the effect of treatment, as in PnK. 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 participation equation. There are 1509 individuals in the sample who reside in control villages, which were purposely drawn in the original sampling scheme because there were no credit programmes and no one in the village participated in any of the credit programmes. These individuals have a zero empirical probability of treatment. In addition, only relatively few villages have credit groups for men and 27.6 per cent of females (1264 individuals) reside in a village without a female credit programme. Chemin recognises 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.[3] More importantly, treating thousands of observations that lack choice to join a credit programme as if they do have choice will seriously bias the DPJ results.

(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.

The outcome variables that DPJ examine are measured at the level of the household, not the individual, so the PSM 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.[4] That means that if one household member is treated the others are assigned as controls even though the value of every outcome variable (such as per capita expenditure) is exactly the same for each of them. In effect, DPJ consider all of the members of the household *except* the person who belongs to the credit programmes as not having been treated (that is, as controls), and these individuals are then matched with individuals in other households who have been treated. As both the treatment and control individuals, as defined by DPJ, experience the same outcome (because in both households someone has been treated), the treatment effect estimated by DPJ is zero whereas in reality it may be positive. In the sample as a whole, there are 3441 observations that DPJ assign to the control group that actually are members of a treatment household, making up 40.6 per cent of all controls. More importantly, all of these observations are, by construction, in villages with a credit programme. In treatment

villages these incorrectly assigned observations make up 74.2 per cent of controls so any matching will be problematic.

(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. It would be appropriate to refer to '*thana* fixed effects' rather than 'village fixed effects' (although 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 microfinance participation (Table 1).

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 programme treatment. Recognising this, Chemin does not include any households from control *thanas* in his propensity score estimation sample. However, DPJ include all 1509 individuals from these *thanas* and assign them the *thana*-specific intercept of *thana* 1 (when Stata is asked to form the dummy variables it automatically drops the first indicator dummy variable, which in this case is the dummy variable for *thana* 1). This is, of course, completely arbitrary and there is no reason to believe that the five *thanas* (15 villages) that are control *thanas* have the same *thana* intercept as the arbitrarily ordered first *thana* of the treatment group. Observations from these control *thanas* cannot contribute anything to estimating the propensity score equation and moreover are likely to bias the results quite significantly.

(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.[5] 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 will result in biased matching estimates. For example, variables such as 'have self-employment enterprise', 'livestock value', 'savings', and 'agricultural income' can all be affected by participating in a credit programme. As these variables are all measured in the data post-treatment, they should not be included in X. The nature of the microfinance treatment is to generate the self-employment activity or the required asset, so once it is essentially netted out of the impact evaluation there will be a bias against finding any treatment effect.

(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 labelled 'Education' in Table 1 of DPJ never varies among members of households. The variables 'father's education' and 'mother's education' do not vary among

members of the same household. 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 appear to vary at the level of the individual. This is certainly unexpected and not at all made clear or explained in the paper.

A number of further discrepancies in DPJ are worthy of comment.

- 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 not wage rates but what they are is not explained in DPJ (nor were they explained in communication with the authors).
- The variable 'highest grade completed' is a recoding of the maximum of 'highest grade completed' within the household. In this recoding, those who attended post-secondary school are given the highest value of this continuous variable as are those for whom a missing value was recorded because their schooling was indicated in a different schedule of the questionnaire. The latter are mostly children.
- The *p*-values shown below the estimates of Table 1 are not based on 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 programme are independent across its members.
- DPJ define strata incorrectly when estimating the impact of credit programme participation on household expenditure. 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. There is no doubt that when the literature refers to stratification matching, it is referring to strata defined by the propensity score, as used by Chemin (2008: 475). However, DPJ create strata based upon the *outcome*, total household consumption expenditure, rather than on the propensity score. 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-neighbour matching algorithm within the outcome-based strata. This unusual method is not explained by DPJ and is likely to bias results against finding a treatment effect.
- The significance levels of the estimates of average treatment effects in Tables 2, 3, and 4 were not obtained from bootstrapping as claimed. Although Abadie and Imbens (2008) cast some doubt on bootstrapping as a means of obtaining the asymptotic variances of matching estimators, 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.
- The *t*-statistics on weighted average treatment effects estimated from stratification matching 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 levels. However, 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. There are exactly five 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 infinity.

- The stratification estimates of Table 2 are actually based on probit propensity scores, not logit propensity scores. Although this should not matter in practice it is an oddity.
- 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. Leaving aside the issues of strata and *t*-statistics above, 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 Table 1, and unlike those of Chemin. Many of the coefficients and *t*-statistics of the underlying propensity equations for each strata are implausible. The strata have as few as 65 valid observations and most of the coefficients are unidentified. Yet it is on the basis of these that DPJ calculate propensity scores and the average treatment effect.
- 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 use Chemin's specification 3 for the propensity score equation (as DPJ state at the bottom of Table 2), but rather Chemin's specification 2 (DPJ's specification 3) omitting all of the *thana* fixed effects.[6]
- 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 same is true for the male treatment effects model.
- No attempt is made to allow for the choice-based nature of the sampling frame used with these data. That the sampling frame is choice-based is explicit in PnK and the subsequent replications of Roodman and Morduch (2009), see Pitt (2011a, 2011b). Credit programme 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 programme participation. When the weights are unknown, Heckman and Todd (2009) show that with a slight modification, matching methods can still be applied.
- DPJ do not follow Chemin (2008: 475) by removing the village fixed effects from total household consumption expenditure (and other outcomes) from a prior regression, implying that they do not provide a true replication.

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 discrepancies it is impossible to predict the sign or magnitude of the full set of errors. My homepage (*c.f.* note 1) reports PSM 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 website in early 2011) and the same specification and variables as PnK. 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 programme for females are included. Alternative estimates are employed incorporating issues raised above to see what would happen if DPJ had replicated PnK accurately. The results indicate positive and highly significant effects of women's participation in microfinance in every one of the matching estimators.

This assessment of DPJ has identified a number of serious limitations that undermine their claims to either replication of or failure to confirm the findings in Chemin (2008) or Pitt and Khandker (1998). Replication can be useful but must be transparent and careful. Whilst 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 (DPJ are surprisingly inconsistent given their advocacy of replication and replicability). 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 challenge the work of others. As this response demonstrates, a sober

assessment of DPJ reveals serious discrepancies between the 'pure, statistical, and scientific' approach to replication and what the paper delivers.

Notes

1. A longer version of this response with more details and all relevant data and Stata files are available on my home page at http://www.brown.edu/research/projects/pitt/. This document also presents a replication of DPJ incorporating the points raised here and shows that their results are not confirmed (that is, not replicated) when appropriate corrections are made).
2. DPJ are among the authors of a well-publicised Department for International Development (DFID) report (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 is also critical of PnK and related studies. An examination of the quality of the keystone DJP replication, as provided here, surely also reflects the quality of that DFID report and its conclusions.
3. Note that as PnK use household-level data for both the probability of treatment and the outcome, neither Chemin nor DPJ can possibly be a replication of the credit participation specification of PnK. The PnK participation equation has a sample size of 1195 households in the female credit equation and 895 households in the male credit equation, as compared to 4215 observations in Chemin and 9397 observations in DPJ. In addition, as the PnK credit equations are at the household level there are no variables 'Sex' and 'Age' (but there are other variables not included in Chemin or DPJ).
4. 89 per cent of the households that have a treated member have only one treated member, and only 0.71 per cent have more than two treated members.
5. This is clearly noted in the paper by Caliendo and Kopeinig (2005) that DJP cite.
6. 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, for example, the propensity score equation used in kernel matching is a logit while that used in stratification matching is a probit. These variations are neither described nor defended, as if this were part of a specification search.

References

Abadie, A. and Imbens, G.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. Munich: Forschungsinstitut zur Zukunft der Arbeit (IZA) Discussion Paper No. 1588, May.

Chemin, M. (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, M. and Palmer-Jones, R. (2012) High noon for microfinance impact evaluations: Re-investigating the evidence from Bangladesh. *Journal of Development Studies*, 48(12), pp. 1864–1880.

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), S230–S34.

Pitt, M. and Khandker, S. (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.

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

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

Roodman, D. and Morduch, J. (2009) The impact of microcredit on the poor in Bangladesh: Revisiting the evidence. Washington, DC: 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.