Aremenia WTM Training Evaluation

1. Given the evaluation design and evidence in the report, what can we conclude about the impacts of the project?

The evaluation of this dimension of the project is professional and thorough and unfortunately provides almost no evidence that the MTM training (and coupled credit) program had an important impact on farming practices, output, income or welfare.

2. Can we conclude whether or not this project was successful in achieving its objectives? What would we need to do differently in order to draw conclusions?

The evaluation was designed to test the null hypothesis of no impact of the project. Along most relevant dimensions, this null hypothesis cannot be rejected in most relevant. Certainly if the report presented a multiple outcome test of this null hypothesis it would not be rejected. Of course this does not mean that we can conclude that the project had no effect; it may well have had an impact that is too small to detect. It might be useful for the report to use the observed variation in the data to report power along key dimensions. For output, the point estimate of the impact is large and the standard errors are huge (Table II.9), so it is possible that the project had a very large effect. For consumption (Table II.16), I think the chance of there having been a large undetected impact is very, very small. The preponderance of the evidence is that there was not an important impact of the project.

3. What lessons learned/conclusions from the design and implementation of the project and evaluation can be used for future MCC and MCA decisions?

I'll bring up a few details below, but the design of the evaluation appears to have been strong. The main conclusion has to be that the project provides no support for the idea that the kind of training program – coupled with improved access to credit -- implemented in Armenia has a strong medium-term impact on the farming practices or outcomes of those who received the training. This is a learning opportunity. An appropriate response would be to go back to the initial design documents to see why it seemed that this intervention would be effective, in order to avoid repetitions of this experience. Perhaps knowledge was not as constraining a factor as it appeared to be during the design phase; perhaps there were more significant problems of implementation than are apparent in this report; or perhaps the delays in irrigation infrastructure provision were fatal.

Additional Comments:

1. The most important technical problem of the evaluation is the fact that a proper sample frame was not constructed and used. As a consequence, we do not properly understand the population for which the evaluation applies. Households were "selected at baseline for FPS interviews based on their likelihood of participating in training if it were offered in their community, as assessed by mayors using criteria provided by the survey team." It is sensible to focus the survey on households most likely to participate to increase the probability of detecting an impact. But it would be far preferable to understand how this population relates to the general population in order to deduce population effects.

This would have been a first order concern had positive effects been detected, in which case we would have wanted to know more about the impact on the general population. It's not much of a concern now, because the high participation by sample households in the treatment group indicated that the mayors were on-target, so this is indeed a sample in which we would have expected to see an impact were any to exist.

2. What happened with the 2nd survey (of the 3)? Why was it done? Is it not used because implementation was delayed? It could still serve as valuable data to nail down baseline values more precisely.

3. The delay in infrastructure rehabilitation may be an important factor in the lack of impact. Of course, these delays are not random, so we cannot interact treatment and an indicator that infrastructure was completed on time. I wish I had a suggestion for steps forward to investigate this. Is there any evidence on the causes of the delays? Perhaps something that can serve as an instrument?

4. When there are multiple categories in a table (like the specific practices tables), provide a joint test of the significance of the impact.

5. On the practices, it would be useful to explore two differing reasons for the lack of change. First, is there any information on knowledge of the practices? Did the training succeed in generating new knowledge? It would be interesting to know if if the training was effective in transmitting knowledge, but it just was the case that farmers didn't want to adopt, or was the training ineffective?

6. I understand the decision to censor the dependent variables, given the noise that invariably exists in this kind of survey. But the uncensored results should be presented as well, at least in the appendix.

7. Provide standard errors for impact estimates in the graphs and tables; it makes it a lot easier for the reader to estimate confidence intervals than the p-values that are provided…

8. The finding of a fairly large but not significant difference in production between treatment and control is the one indication that something might be going on. However, I agree with the reports interpretation that this is best interpreted as noise. There is no change in practices or cultivated area. Table 11.11 is perhaps the most discouraging table – with some precision it seems that investment in farming didn't increase. And we see later on in the consumption table that there is no change there, either. These consumption results should receive more emphasis: there is a lot more precision in these measures. The weight of the evidence is that little changed.

9. A minor point: I don't have any problem with the use of the linear probability model in this context; it's well-defined here. But the argument that it should be used to ease automation of the estimating procedure is really a poor argument.