LEARN Linking Environment & Agriculture Research Network

IMPACT EVALUATION IN CANADIAN AGRICULTURAL LANDSCAPES: SCOPING THE SCIENCE

Robin Naidoo
Conservation Science Program, WWF-US, and
Institute for Resources, Environment, and Sustainability, University of British Columbia

Peter Boxall and Wiktor L. Adamowicz
Department of Resource Economics and Environmental Sociology, University of Alberta

Project Number: **RP-02-2011**

Research Paper
September 2011

Canada

UNIVERSITY OF ALBERTA
DEPARTMENT OF RESOURCE ECONOMICS
AND ENVIRONMENTAL SOCIOLOGY

Growing Forward

Abstract

Impact evaluation methods such as experimental and quasi-experimental designs are rigorous ways of evaluating the impact of a policy or program on intended or unintended outcomes. We discuss in basic terms how to design and conduct such evaluations, with specific reference to agricultural landscapes. We then discuss an impact evaluation of China's Grain for Green program. Our hope is that this paper is able to help stimulate rigorous evaluations of a variety of planned and existing programs on Canada's agricultural lands.

Keywords: Experimental design, quasi-experimental, statistics, outcomes, impact evaluation, conservation, agriculture

JEL Classification: Q15, Q20, Q57

I. Background

The current economic downturn has put into sharper focus the increased interest on the part of donors, government agencies, and non-governmental organizations in evaluating the impact of their programmatic activities on expected outcomes. For example, a donor might be interested in knowing the impact that a $5 million gift to an environmental organization has had on the conservation of biodiversity in a particular region or country. A government agency responsible for public health might want to evaluate how their 2009-2010 activities did or did not reduce infection rates of a particular disease in targeted areas. There are many other examples that could be used. A shared goal in every case would be a desire to assess the degree to which some level of effort or spending has or has not contributed towards a stated outcome of the program in question.

As one might imagine, there are a wide variety of philosophies and methodologies that could be employed to answer the question "What has been our impact on *X* given our effort *Y*"? A thorough review of the field of program evaluation is beyond the scope of this paper. Instead, we will focus on describing a particular type of program evaluation called *impact evaluation*. We will briefly describe its use to-date, particularly with regard to the environmental field, and the advantages and disadvantages that this method offers. We will then describe the nuts and bolts of the methodology while attempting to minimize jargon and assuming a minimal background in statistical analysis. The final section will examine in detail an evaluation conducted on China's Grain to Green program, a scheme in which rural landholders are paid to retire cropland on sloping lands so that environmental benefits may be realized. This examination will expose some of the methodological details described in the previous section. Throughout we provide references that the interested reader may wish to consult for more detail in particular areas.

II. Impact Evaluation: a Brief Introduction

Experiments are generally regarded as being able to provide the strongest evidence when testing hypotheses regarding the way the world works, and as such are a pillar of scientific inquiry. Experiments have a long history in the agricultural and environmental sciences (e.g., assessment of crop productivities under different fertilization regimes) as well as in other areas

of the natural and social sciences. Yet, when it comes to environmental or agricultural *policy*, there is a notable lack of use of experimentation to inform our decisions. In a recent and influential opinion piece, Ferraro and Pattanayak (2006) argued strongly for an increase in experimental evaluations of environmental programs, saying that rigorous empirical evaluations in the environmental field lag far behind other policy domains such as health and education. This does indeed appear to be the case, although a number of environmental and development organizations seem to have heeded the call and are now actively engaged in promoting impact evaluations of a subset of their projects (e.g., World Wildlife Fund, Global Environmental Facility).

When there is no possibility of conducting an experiment (due to logistical or ethical reasons, for example), scientists must use other means of investigation. Statistical analyses can attempt to mimic the experimental principles of randomization, replication and control, although they cannot demonstrate causality in the same way as experiments can. Nevertheless, so-called "quasi-experimental" evaluation techniques have an important role to play in evaluation because there are many instances when experimentation is simply not possible. In such cases, as in the common circumstance when an *ex-post* evaluation of a program is desired because none was designed at program inception, there is no other alternative but to use quasi-experimental statistical techniques to evaluate the impact the program has had.

Note that for both experimental and quasi-experimental approaches, the key objective is the same: the quantification of a scenario that answers the question "what would have happened to the outcome variable of interest in the absence of our intervention?" This "counterfactual", as it is called, is the reference point against which outcomes under the intervention are compared. In the experimental situation, the counterfactual is created by randomly assigning replicate units to either a treatment group or a control group prior to the beginning of the experiment. This ensures that there is no reason to believe that units getting the treatment are in any way different from those not getting the treatment. In the *ex-post* situation, units were not assigned randomly at the beginning of the intervention, therefore statistical methods are used to create a comparison group that, statistically at least, appears similar to the group of units that received the treatment.

III. Experimental Methods

Imagine a researcher is interested in evaluating the effect of a particular type of fertilization on the growth of a crop. A rigorous yet relatively straightforward way to do this would be to design an experiment along the following lines:

i) select a field or patch of land that is similar in growing condition across its length and breadth;
ii) divide the field into say 100 equally-sized parcels;
iii) randomly assign each parcel to either receive the fertilizer, or not;
iv) randomly assign seeds to each parcel and plant them;
v) measure crop growth in all parcels; and
vi) compare the mean growth of crops in treated parcels versus those in untreated parcels.

While this example ignores many of the complications inherent in designing even simple experiments, it is enough to illustrate the three main principles associated with experimentation: *control, replication,* and *randomization.* The use of *control* parcels allows the growth of plants in the treatment parcels to be compared to those that are not receiving the treatment but that are otherwise growing in identical conditions. Adequate *replication* ensures that any observed similarities or differences in growth in treatment versus control parcels are not due simply to a small number of plants or plots possessing strange characteristics; larger numbers ensure the mean of the outcome variable accurately reflects the mean of the population from which individuals were drawn. Finally, and most importantly, *randomization* of treatment versus control parcels allows the researcher to be sure that any effects of the treatment vis-à-vis the control are not due to observed or unobserved differences in the environmental conditions of the parcels.

These simple principles of experimentation can be also applied to policy-making contexts. Consider a circumstance where a government department wishes to be able to evaluate a soon-to-be-implemented program for payments to farmers to conserve remaining waterfowl habitat on their lands. The budget for the program must be used both to fund the payments, and to do monitoring and evaluation of the program. Given these budget constraints, managers in the

department are asked to submit three scenarios for designing and evaluating the program, for which they come up with the following.

Option 1: Monitor 80 farms that will receive payments. Start the waterfowl habitat monitoring after the farmers have started receiving payments, and choose farms to receive payments based on connections and personal contacts with farmers in the target area (thereby reducing transaction costs).

Option 2: Monitor 70 farms that will receive payments. Measure waterfowl habitat in each farm before the payments start, and then continue to monitor each farm after payments have started. Choose farms to receive payments based on connections and personal contacts with farmers in the target area (thereby reducing transaction costs).

Option 3: Randomly choose 40 farms to receive the payments, and also randomly select 20 other farms in the target area. Measure waterfowl habitat in each farm before the payments start, and then continue to monitor each farm after payments have started.

All three program options have the same total cost, but have very different implications for being able to rigorously evaluate the impact of the payment scheme. Option 3 is of course the best option from this point of view, as the others have deficiencies regarding lack of baseline data against which to measure change (Option 1) and lack of an appropriate control group against which to evaluate the payment scheme (Options 1 and 2). On the other hand, it is also apparent that fewer farmers are able to receive payments if a rigorous experimental evaluation is implemented at the beginning of the program, since more effort must be expended to sample both pre- and post-, as well as treatment and control, farms. This illustrates one of the tradeoffs when considering different types of evaluation schemes; rigorous empirical evaluations tend to be expensive and require much time, effort, data, and expertise to conduct. As such, even its strongest proponents agree that not all programs should be evaluated in this way (Ferraro and Pattanayak 2006).

IV. Quasi-Experimental Methods

Quasi-experimental methods refer to statistical techniques that are used to mimic, to the degree possible, the conditions that experiments produce. Because these are typically implemented in an *ex-post* fashion, after the program has already been begun (or even ended), they are considerably more complex in terms of statistical methodology than experimental designs. This is the price that must be paid if it is not feasible or practical to design and implement an experimental evaluation at the outset of the program. Done correctly, however, they can produce evidence that is comparable in robustness to that produced from experimental evaluation.

As with experiments, the goal of quasi-experimental methods is to allow the program to be compared with the counterfactual, i.e., what would have happened in the absence of the program. In this section we focus on the conceptual issues associated with quasi-experimental designs while keeping the technical details to a minimum. The case study example in the following section reveals more of the technicalities of quasi-experimental methods, and readers interested in delving even further into these may wish to consult several of the following references: Pattanayak (2009), which this section draws heavily on, Joppa and Pfaff (2010), Andam et al. (2010) and Andam et al. (2008).

Consider the same example above, but in this case imagine that the program on payments to farms for conserving waterfowl habitat has been running for the last 5 years. At the time of inception, the government agency was unable to implement any monitoring or evaluation activities, but they are now interested in assessing how successful the program has been. Since no data has been collected by the agency itself, scientists responsible for assessing the program are tasked with evaluating the utility of existing data that was originally collected for other purposes, and conducting the relevant analyses.

As discussed above, the three main principles of experimentation are *randomization, replication,* and *control*. With regard to evaluating a program that is already up and running, it is immediately obvious that randomization of the study objects into control and treatment groups is not possible, since the farms participating in the program were selected five years ago. The

question then becomes how to create a comparison group of farms that mimics, to the extent possible, the type of comparison that would have occurred if farms had been assigned at random to either receive or not receive payments (and then been monitored in the same way).

Although possible, it is unlikely that most *ex-post* impact evaluations will involve situations where units (such as farms, villages, etc.) were randomly selected for inclusion into the program. In most cases there will have been specific reasons why a unit was included in the program. Understanding what those factors were is a key component of an ex-post impact evaluation, as these reasons are often predictable and therefore introduce bias with respect to a completely random process of inclusion. In particular, it is critical to consider whether the likelihood of being included in the program is in some way related to the outcome variable of interest, as this would introduce serious bias in any evaluation if not accounted for.

For example, in Costa Rica it has been shown that estimates of the effect of protected areas on reducing deforestation, which are derived from comparing forest clearance in protected areas versus a *random selection* of unprotected areas, are biased upwards. This is because protected areas have been placed in locations where the land is undesirable from an economic point of view, and therefore are at below-average risk of being deforested, even if not under formal protection (Andam *et al.* 2008). This non-random placement of protected areas means that a random sample of unprotected areas is an inappropriate comparison group when evaluating the impact of protection on forest clearance.

What then *is* the appropriate comparison group? Recall the counterfactual: what would have happened to the outcome variable in the absence of the program? Since we cannot observe the same farms both with and without the program (they either receive payments or do not), the goal is to identify a set of non-payment farms that are, in all important ways, otherwise identical to those that received the payment. If this can be done and the outcome variable can be measured in the same way in both groups, we can be quite confident that any differences in outcomes are in fact due to the payments program.

Naturally, the devil is in the details of constructing the comparison group, and in this regard the statistics can get quite complex. This is especially true because a number of different variables could have simultaneously influenced the likelihood that a farm was selected into the program. For example, imagine a situation where the only thing that had influenced a farmer's decision to participate in the program was how big their farm was; only farmers with farms of a certain size decided to enroll in the program. If this were the case, then if we identified all farms of that size class that were NOT enrolled in the program and randomly selected a set of these farms, we would have an appropriate comparison group against which the impact of the program could be measured.

In reality, however, it is likely that a number of factors were interacting to produce the set of farms that actually became enrolled in the program. These could include such things as where the farm was located in relation to major cities and towns, the amount of remaining wetland cover on the farm, the wealth of the household, and so on. When a large number of variables interact to affect the likelihood of program enrolment, it is necessary to resort to statistical techniques to quantify these impacts. Prior to doing this, however, a detailed understanding of the theory and practicalities of how farms were selected is necessary, so that the relevant explanatory variables can be identified and quantified. Without such an understanding important variables may be omitted or irrelevant variables included in the modeling process, both of which may lead to inefficient and/or biased models of farm enrollment.

One common technique by which the likelihood of farm enrollment is modeled is called *propensity score* modeling. The propensity score is the probability of a farm having been enrolled in the program, and is a function of the variables described above. As mentioned, the specific set of variables depends on the analysts' knowledge of the system and the factors likely to have governed enrollment, and will vary on a case-by-case basis. The statistical technique used to create this function and to model the probability is called logistic regression (though what's called a "probit" model could also be used here). The output of a logistic regression model is a predicted probability of enrollment for a set of farms that were indeed enrolled, and a set that were not.

The details of logistic regression analysis will become clearer in the example to follow, but a rough step-by-step procedure might be as such:

1) For the program in question, identify those farms that enrolled in the program and, in the same general area/region, identify as many farms as possible that *did not* enroll in the program. For each farm, quantify the change in wetland cover at a point in time before the program began, and a point in time after the program's implementation.
2) Based on theory and knowledge of the area, the program, and the factors likely to affect program enrollment as well as the outcome variable of interest (change in wetland cover), choose and quantify variables that will enter into the statistical modeling procedure. These should be collected for all of the farms identified in (1). Variables could include biophysical, socioeconomic, institutional, and geographic variables that adequately characterize farms.
3) Using the dataset produced by (1) and (2), estimate a logistic regression model that produces, for each farm, a predicted probability of having been enrolled in the program. The dependent variable is whether the farm was indeed enrolled in the program (1 if "yes", 0 if "no"), and the explanatory variables are those from (2).
4) Based on the resultant probabilities, match each farm that was enrolled in the program with 1 to 3 non-enrolled farms that have the *closest predicted probability* (i.e., the closest propensity score). These matched farms form the comparison group.
5) Calculate the mean of the outcome variable (i.e., change in wetland cover on farms) for the farms that enrolled in the program, and for the matched set of farms identified in (4). The difference between the two is a robust and non-biased estimate of the effect of the program on the outcome variable of interest.

The preceding is a stripped-down summary of the most common way of constructing a matched comparison group. There are other ways of doing so, including simpler methods ("pipeline" matching, where the comparison group consists of, e.g., farms who have expressed interest and/or intend to participate in the program, but have yet to do so (Chase 2002), and more complicated methods, a few of which we turn to below. In addition, the propensity score technique has a variety of technical twists to it that we have ignored for the sake of simplicity.

Readers interested in both of these issues may refer to Pattanayak (2009) and other references noted at the end of this document for more detail.

For the propensity score approach to provide a valid estimate of the effect of a program, a critical assumption that must be met is that called "conditional independence". What this means is that conditional on the variables that are entered into the regression model, the expected probability of a farm having enrolled in the conservation program is the same for those farms that actually did so as for those that did not. In other words, it is assumed that there are no differences between farms in the program versus the matched comparison group not in the program that would affect the outcome variable of interest, because these have been controlled for by the variables that make up the propensity score.

How valid is this assumption? To some degree it depends on how well the model captures all the relevant variables that are expected to determine a farm's participation in the program. However, even the best models will almost inevitably lack some potentially relevant information, because some variables are simply difficult or impossible to accurately measure. Despite our inability to measure them, these "unobservable" variables may nonetheless be playing an important role in determining whether farms are part of the conservation program. For example, a farmer's attitude or inclination towards conservation-friendly land management may have a big influence on their decision on whether to participate in the program, however quantifying this very specific individual characteristic would require a specialized survey administered to all participating and non-participating farmers in the region, and so is unlikely to be available.

This example forces us to be aware of things outside our model that can still lead to biased estimates of the effect of a program. How can these factors be accounted for if they cannot be observed? There are several options. "Difference-in-differences" (DID) methods quantify the difference between the treatment group and the comparison group prior to the intervention, as well as after the intervention, and subtract the effect size of the former from the effect size of the latter to get an estimate of the program impact. This controls for any unobservable effects by removing any trend due to unobservable effects that are not captured in

the propensity score. It's assumed that that the trend in control farms is equivalent to that in treatment farms, and that the effect is time-invariant (i.e., the unobservable variables affect farms in the same way both before the treatment has been established, and after). Note that DID methods are conceptually the same as the Before-After-Control-Intervention experimental design ideal.

Another way of controlling for unobserved effects involves the method of "instrumental variables". Instrumental variables are variables that are highly correlated with the treatment (i.e., whether a farmer participates in a program) but not with the outcome variable (i.e., change in waterfowl habitat on farms). In other words, instrumental variables affect the outcome variable of interest only through their indirect effect on explanatory variables. They are useful in cases where there is considerable uncertainty in whether all the relevant variables have been included in a model that predicts treatment effects on an outcome variable, and in situations where the causality or directionality of the model is in doubt. As an example, consider again the program mentioned above. A simple model to estimate the effects of the program would be a linear model that considers the lone explanatory variable "program" coded as "1" if a farmer entered the program, and "0" if not. As we have already discussed, this is a pretty poor model of the effects of the program since participants were not randomly assigned, and therefore comparing the difference between these two groups will likely result in a biased estimate of the program effect. There are likely to be a number of other variables affecting the outcome in addition to whether a farmer was in the program, and it is also possible that the degree of remaining waterfowl habitat may have influenced farmers' decisions on whether to participate or not.

An instrumental variable that is highly correlated with the decision on whether to participate in the program or not, but is uncorrelated with remaining waterfowl habitat, controls for unobserved effects because it essentially acts as a natural agent of randomization for the treatment variable. By controlling for this variable we can therefore isolate the effect of the treatment on the program outcome. Typically, a two-stage modeling procedure is used; in the first step, the treatment variable is regressed on the instrumental variable, and in the second step, the predicted values from step 1 are used to model the program outcome. Despite seeming like

an excellent solution to the problems mentioned above, in practice it can be exceedingly difficult to identify variables that are indeed correlated with the treatment but uncorrelated with the outcome. More details on instrumental variables can be found in Gangl (2010) and Newhouse and McLellan (1998).

Although we have described a variety of fairly complex statistical techniques that are used to measure the impacts of programs in cases where randomized experiments are not possible, none of these analyses can go forward without a wealth of data. The collection of the appropriate variables requires a detailed knowledge of the theory behind the program in question and the socioeconomic, geographical, and biophysical context in which it is occurring. Because much of the data that will be used will have been generated independently of the program that is being evaluated, a good knowledge of possible sources of existing information, as well as the analytical, statistical, and GIS skills that might be needed to derive them, will be helpful. These technical and data requirements bring up a key question to be answered at the outset: should an impact evaluation of a particular project be attempted? In addition to the points raised above, other questions to consider in answering this question cover the following:

- *Cost/resources*. Impact evaluations cost money to conduct. Are there sufficient resources available?
- *Novelty*. How novel is the intervention that is being proposed? The more innovative the program, the greater the value added and the greater the likely benefits of conducting an impact evaluation.
- *Broader impact*. An impact evaluation will have greater weight if it considers a question that is scalable or replicable in contexts other than the one being considered.
- *Policy relevance*. What appetite is there for the results of an impact evaluation? They should be conducted in situations where the results are likely to help change the way we do business, in terms of program design and implementation.

V. Case study: Effects of China's Grain for Green program on Rural Households

Our case study involves an examination of the effects of China's Grain for Green program (GFG). The GFG, also known as the Sloping Land Conversion program, was initiated in 1999 and is a voluntary program (though there are reports of farmers in certain instances being

"strongly encouraged" to participate) that pays farmers to retire and plant trees on a portion or all of their sloped croplands. The environmental benefit of the program is to increase tree cover on sloped lands, thereby, it is thought, reducing soil erosion. The degree to which these environmental benefits have actually been achieved remains uncertain, but prior to the study we discuss here, there was little evidence of the side effects of the GFG on rural households participating in the program. This is an important question because there are alternative possible impacts. The program may have a positive impact on rural household livelihoods because payments per-acre are relatively high, and the retirement of cropland may allow labour to be allocated to other more profitable economic sectors. On the other hand, if political factors lead to the already well-off gaining preferential access to GFG payments, the poorest rural households may not be receiving funding that in theory should be directed to them (the GFG does state that one goal of the program is poverty alleviation).

While we focus here on the Grain to Green program, other selected impact evaluations related to agriculture include, among others, studies on the impact of technological change and the wellbeing of rural farmers in Bangladesh (Mendola 2007), the effects of agricultural technology adoption on income levels and poverty in rural China (Wu *et al.* 2010), the effects of water conservation and intensification on net returns to rice farmers in Ghana (Faltermeier and Abdulai 2009), and the intended and unintended impacts of USA's Conservation Reserve Program on rural households and communities (Sullivan *et al.* 2004). For a longer list of impact evaluation studies related to agricultural programs, see the website of the International Initiative for Impact Evaluation's website (www.3ieimpact.org), which contains a database of a variety of evaluations, especially as relates to programs in the developing world (shortened here to http://tinyurl.com/3o7h9vn).

Question of interest

A recent study used quasi-experimental matching methods to assess the impact that the GFG has had on participating rural households in China (Uchida *et al.* 2007). More specifically, the research addressed the following question:

- To what extent has the GFG impacted participating households' income, assets, and labour allocation?

Research methodology

The authors conducted a household study in 2003, surveying 359 households in three Chinese provinces that had been participating in the GFG since 2000. Of the households surveyed, 75% participated in the program and the remaining 25% were non-participants. Further sampling details can be found in Uchida *et al*. (2007). One important point to note is that as there were no baseline data available on households prior to GFG implementation, the authors asked survey respondents to recall the state of their household in 1999 (prior to participating in GFG) and in 2002 (the previous year).

Similar to what we have already seen, a fundamental problem with evaluating the question above is that the participation in the GFG might be non-random with respect to variables that affect both the outcome variables of interest and the likelihood of program participation. In the case of participating versus non-participating households, the data of Uchida *et al*. (2007) show that a number of socioeconomic and demographic characteristics are significantly different among the two groups (Table 1). In particular, it appears that participating households are poorer and have fewer assets (except for land) than non-participating households. These differences among groups must be controlled for if attempting to isolate the impact of the GFG, as opposed to intrinsic household characteristics, on rural households.

The authors do this by using some of the quasi-experimental techniques we described earlier. Specifically, the authors first use a logistic regression (logit) model to evaluate factors that may have contributed to participation in the GHG. These factors cluster into four broad types: (1) environmental factors, such as slope and distance to waterway; (2) wealth factors (such

as land, income, and asset holdings); (3) costs of implementation (such as proximity to transportation network); and (4) other household socioeconomic and demographic characteristics. The authors find that, despite its importance when examined in isolation, income is not a predictor of GHG participation when other factors are controlled for. Statistically significant variables (Table 2) included those related to transportation costs and household demographics, as well as a plot's slope; as would be expected, landholders whose plots were on more sloping land (i.e., the targets of the program) had a much higher chance of participation. The authors do not explain the fact that all the coefficients on the transportation costs variables were positive, a counterintuitive result that suggests that all else equal, participants whose enrollment costs were higher were more likely to participate than those with lower costs.

Table 1. Socioeconomic characteristics of participating and non-participating households in the Grain for Green program in one region of China. Adapted from Table 1 in Uchida *et al.* (2007).

Variable	Participating	Non-participating
Household size	4.8	4.5
Land size (mu)	13.8	10.2
Distance of house to farmed plot (m)	1029	760
Per-capita income (yuan)	1404	1850
Agricultural per-capita income (yuan)	648	869
House value (yuan)	13659	20066
Consumer durables (yuan)	569	930
# households in sample	253	86

Table 2. Direction of predictor variable effect on a household's probability of participating in China's Grain for Green program. Adapted from regression results in Table 3 of Uchida *et al.* (2007).

Variable	Coefficient sign
Max. slope on property	+
Migrants in household	+
Distance from road (m)	+
Distance of house to farmed plot	+
Household size	+
Household head age	+

The authors then use a matching approach based on the outputs of the above logistic regression model that predicts participation in the program based on the 1999 information. These outputs are essentially the Propensity Score described earlier, i.e., the probability that a respondent participates in the program, conditional on the variables that enter into the regression model. In this case the authors match each participating household with one non-participating household. One design issue here is that because there are only one-third as many non-participating as participating households (86 vs. 253), the same non-participating household can be a match for a number of different participating households. The authors are unclear on how many of the 86 non-participating households were ultimately used as matches; it is likely only a subset of the 86 but this is not specified. Additionally, it would have been useful for the authors to present a table of the logit regression covariate means and standard deviations for both the participating and matched non-participating comparison groups. The differences in variables shown in Table 1 should in theory have disappeared, although in practice matching typically reduces, but does not entirely eliminate, imbalances in covariates among treated and non-treated groups (e.g., Andam *et al.* 2008, Andam *et al*. 2010).

As mentioned above, the assumption of conditional independence when using the propensity score approach is a strong one and unobservable effects on the outcome variable of interest may yet be present. Uchida *et al*. (2007) recognize this and therefore take one of the steps discussed above: they use a Difference-in-Difference approach in which the difference between means in variables of interest between 2002 and 1999 in the matched comparison group of households is subtracted from the difference in means of the same variables for the households participating in the GHG. As shown in Table 3, the results are for the most part similar, although there is one variable for which the interpretation is significantly different when both the 1999 and 2002 data are used, rather than just the 2002 data: off-farm jobs would have been interpreted as having decreased as a result of the GHG, whereas the matched analysis shows they have in fact increased. It would also have been useful for the authors to show differences among participating and non-participating households when a (naïve) random sample of non-GHG households were used as the comparison group. For a non-agricultural example

that compares how comparison groups constructed from matching methods produce different results as compared to naïve random samples or other methods, see Andam *et al.* (2008).

Table 3. Estimated impacts of China's Grain for Green program on various outcomes using two different evaluation methods. Adapted from Table 4 of Uchida *et al.* (2007).

	Difference in means	
Variable tested	Matched (no DID)	Matched (DID)
Crop per-capita income (yuan)	-172.21	-167.14
Other agricultural per-capita income (yuan)	171.99	168.02
Off-farm jobs per household	-0.04	0.045
House value (yuan)	485.8	521.8
Livestock value (yuan)	180	220

VI. Conclusion

There is a growing recognition among researchers and practitioners that we should be as rigorous in evaluating the impacts of policies or programs as we are in testing academic hypotheses in the natural and social sciences. Impact evaluation methods, both experimental and quasi-experimental, provide a set of tools that can help quantify the degree to which a program or policy has affected various outcomes of interest. Both types of impact evaluation methods are likely to be costly and data-intensive for agricultural landscapes. This is especially true for quasi-experimental evaluations, where information on participants and non-participants relating to the outcomes (e.g., wetland area), the biophysical characteristics of the areas (soil type, vegetation, parcel size, etc.) and the characteristics of the people (age, income, etc.) needs to be collated from a variety of sources before the evaluation can proceed. Nevertheless, certain situations and policy interventions lend themselves well to impact evaluation. In such cases, our understanding of the effects of the program can be greatly enhanced by using the experimental and quasi-experimental techniques we have described here. By building up a library of impact evaluation studies in particular contexts, our general understanding of programs and policies will be enhanced and allow us to manage Canadian agricultural landscapes using the best available understanding of how and why various programs work.

VII. Literature Cited

Andam, K. S., P. J. Ferraro, A. Pfaff, G. A. Sanchez-Azofeifa, and J. A. Robalino. 2008. Measuring the effectiveness of protected area networks in reducing deforestation. Proceedings of the National Academy of Sciences 105:16089-16094.

Andam, K. S., P. J. Ferraro, K. R. E. Sims, A. Healy, and M. B. Holland. 2010. Protected areas reduced poverty in Costa Rica and Thailand. Proceedings of the National Academy of Sciences 107:9996-10001.

Faltermeier, L., and A. Abdulai. 2009. The impact of water conservation and intensification technologies: empirical evidence for rice farmers in Ghana. Agricultural Economics 40:365-379.

Ferraro, P. J., and S. Pattanayak. 2006. Money for nothing? A call for empirical evaluation of biodiversity conservation investments. PLoS Biology 4:e105.

Gangl, M. 2010. Causal inference in sociological research. Annual Review of Sociology 36:21-47.

Joppa, L., and A. Pfaff. 2010. Reassessing the forest impacts of protection: The challenge of nonrandom location and a corrective method. Annals of the New York Academy of Sciences 1185:135-149.

Mendola, M. 2007. Agricultural technology adoption and poverty reduction: a propensity-score matching analysis for rural Bangladesh. Food Policy 32:372-393.

Newhouse, J. P., and M. McClellan. 1998. Econometrics in outcomes research: the use of instrumental variables. Annual Review of Public Health 19:17-34.

Pattanayak, S. 2009. Rough guide to impact evaluation of environmental and development programs. South Asian Network for Environmental and Development Economics, Kathmandu, Nepal.

Sullivan, P., D. Hellerstein, L. Hansen, R. Johansson, S. Koenig, R. Lubowski, W. McBride, D. McGranahan, M. Roberts, S. Vogel, and S. Bucholtz. 2004. The Conservation Reserve Program: economic implications for rural America / AER-834. Economic Research Service/USDA, Washington, DC.

Uchida, E., J. Xu, Z. Xu, and S. Rozelle. 2007. Are the poor benefiting from China's land conservation program? Environment and Development Economics 12:593-620.

Wu, H., S. Ding, S. Pandey, and D. Tao. 2010. Assessing the impact of agricultural technology adoption on farmers' well-being using propensity-score matching analysis in rural China. Asian Economic Journal 24:141-160.