INSTITUTE OF FOOD AND RESOURCE ECONOMICS
UNIVERSITY OF COPENHAGEN

FOI Working Paper

# A Comparison of Model-based and Design-based Impact Evaluations of Interventions in Developing Countries

*Henrik Hansen*
*Ninja Ritter Klejnstrup*
*Ole Winckler Andersen*

**2011 / 16**

**FOI Working Paper 2011 / 16**

A Comparison of Model-based and Design-based Impact Evaluations of Interventions in Developing Countries

Authors: Henrik Hansen, Ninja Ritter Klejnstrup, Ole Winckler Andersen

Institute of Food and Resource Economics
University of Copenhagen
Rolighedsvej 25
DK 1958 Frederiksberg  DENMARK
www.foi.life.ku.dk

# A Comparison of Model-based and Design-based Impact Evaluations of Interventions in Developing Countries

Henrik Hansen[*]
Institute of Food and Resource Economics, University of Copenhagen

Ninja Ritter Klejntrup
Evaluation Department, Ministry of Foreign Affairs of Denmark, Danida

Ole Winckler Andersen
Evaluation Department, Ministry of Foreign Affairs of Denmark, Danida

*Abstract*
We argue that non-experimental impact estimators will continue to be needed for evaluations of interventions in developing countries as social experiments, for various reasons, will never be the most preferred approach. In a survey of four studies that empirically compare the performance of experimental and non-experimental impact estimates using data from development interventions, we show that the preferred non-experimental estimators are unbiased. We try to explain the reasons why the non-experimental estimators perform better in the context of development interventions than American job-market interventions. We also use the survey as a source for suggestions for implementation and assessment of non-experimental impact evaluations. Our main suggestion is to be more careful and precise in the formulation of the statistical model for the assignment into the program and also to use the assignment information for model-based systematic sampling.

*Key words*
Development, impact, non-experimental, social experiment, within-study

*JEL-Classification*
C21, C93, H43, O22

[*] The paper was written while Henrik Hansen was Visiting Professor at the Division of Nutritional Science, Cornell University. The hospitality of the Division and in particular of Per Pinstrup-Andersen is gratefully acknowledged

## Introduction

The increasing interest in the effects of development assistance, as reflected in the adoption of the Millennium Development Goals as well as at the high level meetings in Paris (2005), Accra (2008) and Busan (2011), has led to a call for more impact evaluations of interventions in developing countries. In particular after the publication of the report from the Evaluation Gap Working Group: "When Will We Ever Learn? Improving lives through Impact Evaluation" (Savedoff et al., 2006), there has been a surge in donor funded impact evaluations, establishment of a new entity for funding quantitative impact evaluations (3ie) and a network for impact evaluation networks and groups within development (NONIE).

Alongside the call for more impact evaluations of development interventions there is an ongoing debate about evaluation designs and more generally about what constitutes rigorous quantitative evidence of causal impact. The core of the discussion centers on the role and design of social experiments as several researchers argue that only randomized controlled trials (RCTs) can deliver rigorous proof of causal effects of interventions. Other researchers maintain that even though randomization provide robust results it is not necessary for rigorous evidence of causal impact because statistical methods, such as matching, OLS- and IV-regression, can be used to effectively adjust non-experimental data.[1]

One argument for RCTs is the theoretical purity of the statistical model for causal impact, known as the Rubin causal model (RCM), which has the experiment with random selection into participation as the benchmark (Rubin, 1974, 1978, and Holland, 1986). Yet, in practical work within governmental and donor organizations the preference for the RCT is probably more grounded in the perception that non-experimental impact evaluations are generally biased. This perception is in turn undoubtedly founded on the many empirical studies showing that estimates of causal effects from observational data are different from estimates based on experimental data (see, e.g., the meta-study by Glazerman, Levy and Meyers, 2003).

In this paper we show that for development interventions the perception may be unfounded. Four recent studies use interventions in developing countries to test if impact estimates based on observational data are close to impact estimates obtained from experiments. In all four cases we find good agreement between the two estimates indicating that the model-based estimators, using observational data, are unbiased.

Even though four studies is a very small sample it is striking to find four successes out of four trials, in particular in light of the previous results. Moreover, we argue that there are good explanations for the results. First of all, the successes are not independent of the modeling; they occur when information about the selection into program participation is both well described and well used in the statistical modeling. Second, in most cases the outcome of interest is a binary variable, say child school participation, which is easy to measure, rather than a complex, hard to observe variable such as income or consumption expenditure.

[1] This debate is not limited to evaluations of development interventions. It is a discussion across a broad spectrum of sciences. However, our focus is on interventions in developing countries.

Needless to say, we also try to reconcile the new results with previous studies and our three suggestions for the new results are (i) the understanding and modeling of the selection into participation, (ii) the smaller variance in the outcome variable, and (iii) the failure of an economic model (the economics of job-market interventions) which has been confused with failure of the statistical model (model-based impact evaluation).

In addition to being a contribution to the debate about quantitative evaluation methods we hope also to contribute to the practical aspects of impact evaluations. We do this by giving a few suggestions for how to implement and assess model-based impact evaluations. We believe this to be an important object because, despite their vastly increasing number in recent years, social experiments will never become the main tool in impact assessments in developing countries.

The reasons are both economic and political. Good randomized controlled trials are both very time consuming—because they interfere with the program implementation—and expensive. The costs may in many cases be simply too high relative to the benefits, not least due to the limited external validity of many RCTs.[2] Some researchers try to address the cost issue by designing a number of small RCTs. However, this solution runs into the classical statistical problem of ensuring samples of sufficient size such that the power of the statistical tests of the impacts is acceptable.[3]

RCTs may also be politically–and ethically–infeasible due to their direct implications for the design of the development interventions. Even though researchers are exceedingly creative in their designs of RCTs, donors and national policy makers have to sacrifice part of the program targeting in order to use experiments for program evaluation.[4] Thus, they can be difficult to align with local priorities and government officials may not have an interest in a random assignment of participants. Hence, if the goal is to have not only rigorous but also relevant evaluation of development effectiveness we need to learn more about when and how to design and assess model-based impact evaluations in practice.

The four studies we present and discuss are all using direct comparisons of estimates from randomized experiments (design-based estimators) and non-experiments (model-based estimators). Thereby they build on the tradition of testing the accuracy of model-based estimators using within-study comparisons that started with LaLonde (1986). Cook, Shadish and Wong (2008) also review recent within-study comparisons of impact estimators and they also find that model-based estimators are often unbiased. The main difference between Cook et al. (2008) and the present study is that we focus only on development interventions while they

[2] The limited external validity of RCT-based results is a source of critique against categorical preferences for experiments in evaluations that are meant to have a formative purpose. For discussions in the context of development economics, see e.g., Deaton (2010) and Rodrik (2009). For a discussion with departure in the philosophy of science see Cartwright (2007)

[3] See Gerter et al., (2011, Chapter 11) for a non-technical discussion of the relationship between sample size and statistical power and Cochrane (1977, Chapter 4) for a more technical description.

[4] This issue is discussed more at length in Ravallion (2007). Other ethical issues associated with experimentation with human subjects in the context of development interventions are discussed by Barrett and Carter (2010).

discuss 12 different studies covering US job training and education programs in addition to the development interventions. Moreover, Cook et al. have a different focus as they emphasize good designs of quasi-experiments while we emphasize understanding and modeling the selection procedure as it has been decided and designed by policy makers. Relatedly, Cook and Steiner (2010) also compare and review within-studies. Their focus is on the importance of the functional form (regression versus propensity score matching) relative to covariate selection and on the special role of the pretest values of the outcome variables. We build on their results when making suggestions for design and appraisal of model-based impact evaluations.

Several other studies also discuss when and how to apply model-based estimators—see e.g, Heckman and Vytlacil (2007a, b), Ravallion (2008), Todd (2008), and Imbens and Wooldridge (2009)—but these studies are quite technical as they tend to focus on the theoretical aspects of the model-based estimators in order to scrutinize the applicability from a theoretical viewpoint. We do not discuss the technical aspects of individual estimators but focus on empirical results and practical suggestions.

The paper is structured as follows. First we briefly lay out the difference between the model-based and the design-based approaches and explain the way these approaches are compared in within-studies. Next, we survey the four within-studies that empirically compare the performance of the approaches using data from development interventions and we try to explain the reasons why the model-based estimators perform better for the development interventions than for the many American job-market interventions where earlier literature has found them to be biased. The survey is subsequently used as a source for our suggestions for implementation and assessment of model-based impact evaluations. Finally, we offer a few concluding remarks.

## Design-based and model-based impact evaluations and empirical comparisons

As a point of departure in this paper we wish to note that the current practice, by which quantitative impact evaluations are classified as being either experimental or non-experimental (observational), is somewhat unfortunate because it blurs the way in which the non-experimental evaluations should be approached and assessed. We would argue that it is preferable to make a distinction between design-based and model-based evaluations where the design-based are (normally) RCTs, while the model-based evaluations use statistical analysis of quasi- or non-experimental data. The notions of design-based and model-based sampling are well known in statistics (see e.g., Binder and Roberts, 2003). We stress the relationship between impact evaluation and the terms to emphasize that both methods are useful and that they are based different assumptions leading to different strengths and weaknesses.

Classical statistical modeling (the model-based approach) is based on the notion of a probability model as the hypothetical data generating process. The probability model is a formalization of *a thought experiment*, sometimes in a highly stylized form, but no actual experiment needs to be conducted. The probability model illustrates the relationship between the central parameters of the model and the data, and when necessary it also describes incidental parameters and confounding factors. When describing the population in the classical context, it is simplest to think of the notion of a super-population. An advantage of the model-based approach is that extrapolation to other samples (i.e., external validity) is straightforward prediction based on the statistical model.

A well-known and influential model in the impact evaluation literature is Heckman's sample selection model (Heckman, 1979). This model explicitly incorporates non-random selection of individuals and the stylized model is used for derivation of estimators and statistical inference (Heckman and Robb, 1985). This is a prototype example of a model-based approach to impact evaluation. It is obvious that impact evaluations based on this approach are conditional on the untested assumption of a valid underlying statistical model. If the model assumptions are wrong we get something different from what we expect. The difference between what we get when the model is valid and when it is invalid is the bias.

In the context of sampling from existing populations there is a different use of probability as the probability calculations are related to the sampling procedure used by the investigator in the planning of the sampling. The simplest case is random sampling but other designs are often more efficient (see e.g., Cochran, 1977). By this approach the randomness follows from the sampling procedure, hence, in relation to statistical analysis the probability model is a function of the design of the sampling, which is why it is called a design-based approach. The main advantage of the design-based approach, of which the randomized controlled trial is an example, is that it leads to estimates and measures of uncertainty for the population means of the variables of interest without any assumptions about the distribution of the variables. In that sense the design-based approach is robust; if the implementation of the sampling is in accordance with the planning, then estimated means will be unbiased. A weakness of design-based estimation is that extrapolation to other samples cannot in any way be based on the statistical analysis of the sample at hand.

Rubin's causal model (RCM) (Rubin, 1974, Holland, 1986) is often seen as a development of the design-based approach to statistical inference as the model does not make direct assumptions about the outcome variables because they are modeled as fixed measures for any given individual. Further, Rubin stresses that a major difference between the RCM and Heckman's sample selection model is that the causal effect (the change in the outcome variable caused by the intervention) is separated from the probability model describing the assignment mechanism. This focus on the selection mechanism and its properties and the emphasis on matching of samples prior to the estimation of the impact naturally lead to the randomized controlled trial as a benchmark model because this is the simplest model in which the assignment mechanism is ignorable.

Still, with observational data the assignment into program participation is treated as random in the RCM and since it is not (completely) determined by the investigator, the assignment model is described by a probability model. And when the estimators and the statistical analysis are derived from (classical) assumptions about the assignment model, the RCM is a classical statistical model in the sense that it is a model-based approach to estimation and inference (see Rubin 1990, 1991). It follows that the derived estimators are unbiased when the assignment model is valid, while they are biased when the crucial assumptions of the model fail.

Although Rubin prefers a Bayesian approach to statistical inference, this does not change the fundamental point that analysis of the RCM with observational data is a model-based approach because one has to make untestable assumptions about the assignment mechanism. Likewise, if researchers choose to follow Heckman and formulate a probability model for the outcome variable as well, this does not change the fundamental notion that it is a model-based approach. The main reason why we stress this (well-known) model dependence of the RCM is that we feel it has been somewhat neglected in the recent highly technical search for an omnipotent matching estimator.

Now, since the design-based impact estimator is not model-dependent it is robust to model errors and as such it can be used to evaluate the performance of the model-based estimators whenever an experiment has been performed. The impact estimates can be compared empirically using either between-study or within-study comparisons.

Between-study comparisons look into the differences by comparing the estimated impact of similar but different interventions. An example is the study by Glewwe et al. (2004) that compares the impact of using flip-charts on exam results of 6-8 graders in schools in Kenya. The study finds substantial differences between design-based and model-based estimates.[5]

A serious problem in between-study assessments is that the contrast of interest, the bias from nonrandom selection in model-based studies, is confounded with other differences between the sample designs, specifically differences in treatment details. In the study of flip-charts we observe such treatment differences as Glewwe et al. (2004) compare a controlled experiment in which a fixed number of flip-charts and wall maps are randomly assigned to schools about one year prior to the exam to a sample of schools having a varying number of flip-charts, but no wall maps, during some (unspecified) time before the exam. Hence, in the study the interventions are similar but they are clearly not identical.

Within-study comparisons aim at solving the confounding problem by comparing estimates of the exact same intervention. The causal effect of an intervention is first estimated by contrasting the treated group and a randomized control group. Subsequently the impact is re-estimated using the same treated group and one or more non-experimental comparison groups, which are adjusted statistically by means of regression and/or matching to generate the model-based estimates.

Even though the within-studies solve the problem of confounding bias and differences in treatment details they are not without problems. First of all, as the experimental estimate is used as the benchmark, the experiment must obviously meet all criteria for technical validity.

[5] See also Lipsey and Wilson (1993) for an early meta-analysis of between-study comparisons within psychological, educational and behavioral treatments, and Greenberg, Michalopoulos, and Robins (2006) for a meta-analysis of labor market training programs in the US.

Second, design-based and model-based estimators do not, in general, estimate the same causal effect. Clearly for the within-study bias comparisons to make sense the average treatment effect estimated by the experiment must be evaluated at the same point as the local average treatment effect estimated using, say, matching and RD designs.

A final critical issue is the old problem of specification searches (Leamer, 1983). As within-studies have a fixed, known estimate from the experiment as the benchmark while the choice of model-based estimator and, in particular, the choice of control variables is determined by the researcher, the analysis is open for specification searches, possibly leading to reports of overwhelmingly positive results for the model-based estimator or the reverse, of very poor results for one or several estimators—even when the results are based on identical data sources. Within-studies of earning impact from labor market programs in the US show that the problem of specification searches is a very real concern one should bear in mind when evaluating individual within-studies (see Smith and Todd, 2005a, b and Dehejia, 2005). In the absence of double-blind comparisons in which the researchers estimating the impact using model-based estimators do not know the design-based estimate in advance, a feasible solution is to make sure the within-studies report results illustrating the robustness of the model-based estimators to small changes in the model and sample setting. The four studies we describe all report results of different samples, different model-formulations and, for three of them, different estimators. We are therefore confident that the results are not seriously biased by data mining and specification searches.

The performance of model-based estimators vis-à-vis the design-based estimator can be evaluated in several ways. The simplest and most straightforward approach is to inspect directly if the estimates have the same signs and lead to the same conclusion regarding the statistical significance of the impact. However, differences will occur by chance even if both estimators are unbiased and, as noted in Cook et al. (2008), if we consider an intervention evaluated by two independent experiments for which we assume a power of 80 percent for each experiment, then the expected probability of the two evaluations resulting in similar significance patterns is only 68 percent.[6] Clearly, as the power increases for each estimator the power also increases in the comparison of the estimators, but the simple calculation shows that comparisons of model-based estimators with design-based results requires reasonably large samples for both estimators.

Another approach is to test directly if the difference between the design-based and the model-based impact estimates is zero. This approach is an application of Hausman’s specification test (Hausman, 1978) but the variance of the estimated difference is unknown. McKenzie, Gibson and Stillman (2010) solve this problem by bootstrapping the distribution of the test. Another, less formal solution, which we will indirectly adhere to a few times, is simply to take the design-based estimate as given and test if this estimate is within the confidence bound of the model-based estimator. This procedure leads to a conservative test but it provides a simple metric illustrating that seemingly large differences between estimates are often

[6] The probability of significant findings in both experiments is 64 percent (100×0.8×0.8) while the probability of an insignificant finding in both studies is 4 percent (100×0.2×0.2).

to be expected when the sampling uncertainty of the model-based estimator is taken into account.

A third approach, suggested by Heckman et al. (1998), is to ignore the group of participants and instead estimate the difference in outcomes between the control and the comparison group. The random assignment in the experiment ensures equality of the participant and control group in terms of observed and unobserved attributes and since the control group does not participate in the intervention there should be no difference between control and comparison groups. Consequently, the model-based impact estimate, substituting the control group for the participants, is a direct estimate of the selection bias in the model-based estimator and tests of significance are readily available because this is a significance test completely analogue to the "true" model-based test for causal impact.

The papers we review present a mix of all three types of assessments.

## Within-study evaluation of model-based estimators

In this section we present results from four different within-studies, all using interventions in developing countries as their base for comparison. The four studies are all quite comprehensive, which is why we only focus on their main results. Moreover, three studies report results for broad range of different matching estimators. In the present context this is information overload and we therefore only present results for one of the many matching estimators. We have chosen to report the so-called nearest-neighbor with replacement matching estimator. Results for this estimator are given in all of the three studies.[7] We start, however, with results from a Regression Discontinuity Design (RDD) study.

### The impact of Progresa estimated by RDD

A study by Buddelmeyer and Skoufias (2003) evaluate the performance of the RDD estimator using the data set from the Mexican conditional cash transfer (CCT) programme, PROGRESA. The intervention is a widely known CCT programme because it already from the initial phase was designed to form a social experiment. Further, the history and results of the program is referred in most recent text-books and popular writings about impact evaluations and evidence based development interventions (see, e.g., Gertler et al. 2011, Banerjee and Duflo, 2011, and Karlan and Appel, 2011). The programme began in 1997 and initially targeted poor rural households, providing cash benefits conditional on school enrollment of children, attendance by a household adult in monthly health seminars, and attendance by all household members at scheduled health check-ups. Eligibility for the programme was determined in three stages. First, programme managers selected eligible localities on the basis of a locality-level marginality index, constructed by statistical modeling of the expected poverty levels on the basis of national census data. Second, within each eligible locality, poor households were deemed eligible for the programme on the basis of a discriminant score (another estimate using a statistical model). Third, the resulting list of potentially eligible households was presented at a community assembly for ratification. Results in Buddelmeyer and Skoufi-

[7] In Diaz and Handa (2006) and Handa & Maluccio (2010) matching is on propensity scores while in McKenzie et al. (2010) it is on multiple covariates.

as indicate strongly that the final step made only marginal changes to the list of eligible households selected in the second step, which was purely mechanical.

A randomized controlled trial was carried out to evaluate the programme effects. A total of 506 eligible localities were chosen to be part of the experiment, and of these 186 were randomly chosen to form the control group. This experiment has been studied extensively and it is generally judged to be technically valid (Skoufias, 2001).

Buddelmeyer and Skoufias utilize the very detailed description of the selection procedure and the stringent use of the eligibility criterion to set up a sharp regression discontinuity design. Specifically, they use the household level discriminant score to select households in the program, just below the location specific cut-off value and a comparison group of households, within the same localities, with discriminant scores just above the cut-off value. Subsequently, local average program effects are estimated for school attendance and work activity for boys and girls, 12-16 years old. The experimental estimates are also rescaled to the local average treatment effect in a small neighborhood below the participant cut-off point. This ensures that the model-based and the design-based estimators estimate the same causal effect.

The program effects are estimated at three different points in time: Before the program intervention (October/November 1997), about 8-10 months after program initiation (October 1998) and a year later than that (November 1999), some 20-22 months after the initial payments.

Table 1: PROGRESA impact on school attendance and child work: Impact and bias estimates

| | Boys 12-16 years old | | | Girls 12-16 years old | | |
|---|---|---|---|---|---|---|
| | Experiment | RDD | Bias | Experiment | RDD | Bias |
| *School* | | | | | | |
| Oct./Nov. 1997 | -0.001 | -0.018 | 0.044 | 0.000 | -0.025 | -0.034 |
| | (0.028) | (0.031) | (0.036) | (0.030) | (0.034) | (0.038) |
| October 1998 | **0.071** | 0.008 | 0.033 | **0.082** | 0.039 | 0.031 |
| | (0.028) | (0.033) | (0.035) | (0.029) | (0.034) | (0.037) |
| November 1999 | **0.099** | **0.069** | 0.026 | **0.099** | **0.107** | 0.011 |
| | (0.030) | (0.032) | (0.037) | (0.028) | (0.035) | (0.037) |
| *Work* | | | | | | |
| Oct./Nov. 1997 | 0.007 | -0.013 | 0.004 | 0.000 | 0.027 | -0.021 |
| | (0.029) | (0.031) | (0.034) | (0.024) | (0.021) | (0.023) |
| October 1998 | -0.007 | 0.001 | -0.005 | 0.001 | 0.002 | 0.005 |
| | (0.029) | (0.028) | (0.031) | (0.016) | (0.017) | (0.017) |
| November 1999 | -0.037 | -0.029 | -0.021 | -0.025 | -0.033 | -0.003 |
| | (0.025) | (0.027) | (0.030) | (0.018) | (0.018) | (0.021) |

*Notes:* The impact estimates for Oct./Nov. 1997 should be zero as this is a comparison before the program initiation. Payments started in July 1998. The experimental estimate is CSDIF-50 while the RDD estimate is using the Triangular kernel. Bold figures are statistically significant at the 5%-level, standard errors are in parentheses.
*Source:* Buddelmeyer and Skoufias (2003) Tables 2, 3, and 5.

The main results of the evaluation of the RDD are reported in Table 1. The first column of the Table reports the experimental impact estimates for boys' school attendance and work while the fourth column has the results for girls. The second (fifth) column reports the RDD impact estimates for comparison while the third (sixth) column shows the direct bias estimate in which the control group is used for comparison instead of the participant group. Since the first comparison is before the program start it measures the extent of pre-program selection bias for both the experimental and the RDD samples. As seen, the pre-program differences in school attendance and child work frequencies are quantitatively small and statistically insignificant. The Table also shows that the program had no impact on the extent of child work. And, importantly for the present study, the RDD estimates are in good agreement with the design-based estimates and there is no significant bias.

For school attendance there is good agreement for the November 1999 survey round, while the RDD estimator fails to find significant impacts in the October 1998 round in contrast with the design-based estimates. However, the direct bias estimates are insignificant throughout. Buddelmeyer and Skoufias spend considerable efforts explaining the discrepancy between the design-based and the model-based estimates for school attendance in November 1998, but are unable to point to a single explanation. Yet, selection does not seem to be the problem as the selection bias is statistically insignificant. Hence we follow the authors in concluding that the RDD generates estimates that are remarkably close to the design-based estimates.

**The impact of Progresa estimated by regression and matching**

Turning next to model-based estimators by means of regression and matching, a study by Diaz and Handa (2006) also use PROGRESA as the basis for comparisons. Diaz and Handa only include the second round (October 1998) and, in contrast to Buddelmeyer and Skoufias, they use outside information by choosing comparison samples from the September-November 1998 round of the national household survey ENIGH. Diaz and Handa construct two sub-samples; the first use all rural households in the 1998 ENIGH, excluding localities that were included in PROGRESA, while the second sample also excludes localities that were not eligible for inclusion in PROGRESA. The inclusion/exclusion is based on the marginality index. The second sub-sample is a very nice example of model-based systematic sampling.

Diaz and Handa investigate the performance of several model-based estimators. In this comparison they include exactly the same covariates as those used to compute the household specific discriminant score. Notice, however, that the authors do not compute the discriminant scores. In that sense they do actually not include all available information about the selection process.

The outcome variables of interest are school participation and child labor as in Buddelmeyer and Skoufias; however, the outcomes are analyzed for boys and girls jointly. In addition to these measures Diaz and Handa include food expenditure as an outcome variable. The purpose of this inclusion is to test if differences in the survey instrument have an impact on the outcome estimates. Specifically, the food expenditure modules differ in the PROGRESA and ENIGH surveys. Table 2 reports the main results from Diaz and Handa in the form of direct bias estimates.

Table 2: PROGRESA impact on school attendance, child work for pay and food expenditure, October 1998: Bias estimates

| | School enrollment | | Work for pay | | Food expenditure | |
|---|---|---|---|---|---|---|
| | Bias | S.E. | Bias | S.E. | Bias | S.E. |
| Sample 1 | | | | | | |
| Unadjusted difference | **-0.098** | (0.02) | -0.006 | (0.01) | **-494** | (13) |
| OLS | **-0.047** | (0.02) | -0.014 | (0.01) | **-271** | (18) |
| Matching | 0.022 | (0.04) | -0.028 | (0.03) | **-219** | (34) |
| Sample 2 | | | | | | |
| Unadjusted difference | -0.024 | (0.03) | -0.008 | (0.02) | **-404** | (18) |
| OLS | -0.024 | (0.04) | 0.005 | (0.02) | **-280** | (22) |
| Matching | 0.037 | (0.08) | 0.012 | (0.04) | **-169** | (67) |
| | | | | | | |
| Experimental means | **0.066** | (0.01) | -0.005 | (0.01) | **35.0** | (7.42) |

*Notes:* Sample 1 is the ENIGH household survey excluding PROGRESA localities. Sample 2 is a subset of Sample 1 which excludes non eligible localities. The matching estimator is nearest neighbor with replacement and common support. Bold figures are statistically significant at the 5%-level, standard errors are in parentheses.
*Source:* Diaz and Handa (2006) Table 2 and 3.

Starting with the child work estimates we find, again, that the program had no significant impact. The bias estimates are quantitatively small and statistically insignificant for both samples.

For school participation results are slightly different because sample 1 (all rural ENIGH less program localities) has a higher school participation rate than the control sample. The selection bias is reduced but not removed by the regression estimator whereas it is completely removed by the matching estimator. In the present setting the difference between the two estimators is not surprising as the selection correction for the matching estimator is a logit-regression which is closely related to the discriminant model. Hence, the household level propensity score, estimated by the logit, is in all likelihood very close to proportional to the true discriminant score the comparison households would have been given. In this way the model-based matching estimator is exploiting almost all household level information about the selection process. Moving to sample 2, in which the rich localities are omitted from the comparison sample, we find a good balance between the unadjusted comparison sample and the control sample. This results in small and statistically insignificant biases for both model-based estimators. Thus the interesting result here is that the model-based systematic sub-sampling of the national household survey is actually sufficient to get unbiased impact estimates.

For food expenditure, the situation is completely different as all of the model-based estimates are biased. The bias is reduced greatly by matching, but the bias is still large compared to the estimated impact. This result is not surprising as food expenditure is measured in two different ways in the control and the comparison samples. Diaz and Handa analyze the bias and conclude that the survey instrument difference accounts for about 75% of the total bias. Looking at the results (given in their Table 4) we are inclined to conclude that difference accounts for all of the bias because the design-based estimate is within a 2-standard error bound of all the model-based estimates once the survey instrument difference is removed.

Diaz and Handa also compare the design-based and model-based estimators when the set of control variables is reduced to a small “convenience” set of standard (household demographic) covariates. The limited use of covariates results in biased impact estimates when sample 1 is used for comparison while there is still no bias when sample 2 is used. While this is interesting it is not surprising and we disagree with the authors in their conclusion that a rich set of relevant covariates is an important determinant of the success of the model-based estimators. The point to note is that a well-specified model of the selection procedure is the important determinant. If the selection is not a function of a rich set of variables, as for PROGRESA where the location index and the discriminant score are the sufficient measures, then we may well do with a small set.

Still the study shows that model-based estimates are equal to design-based estimates when the assignment mechanism is well-specified and when outcome variables are measured in the same way in the program and the comparison samples. Further, we emphasize the substantial influence of the model-based systematic sampling which is based on location, but not household information.

**The impact of RPS estimated by matching**

The significant effect of model-based systematic sampling is also illustrated by the third paper, a study by Handa and Maluccio (2010) of the Red de Protección Social (RPS), a conditional cash transfer program in Nicaragua. The RPS was very much like the PROGRESA, although it differed in an important aspect in that eligibility was determined exclusively on the basis of location level variables. Thus, within an eligible locality all households were eligible to participate in the programme thereby leaving the households within the localities to self-select into the programme. For the first phase of the programme six out of 63 municipalities from the Central Region of Nicaragua were chosen. The selection was based on the poverty level and the municipalities’ capacity to implement the programme—judged on the basis of accessibility and coverage of health posts and schools. Within the six municipalities the 42 poorest localities were chosen on the basis of a statistically computed marginalization index—as in the case of PROGRESA. Of the 42 localities half were randomly designated to the program while the other half served as controls. The total sample size was 1453 households, 53% from the program localities. The experiment has been reviewed by Maluccio and Flores (2005) and judged to be technically valid.

Handa and Maluccio form three comparison samples from the 2001 Nicaraguan Living Standard Measurement Survey (LSMS). The first comparison sample includes all rural households in the LSMS, save those in the program municipalities. The second sample is a refinement only including rural households in non-program localities with a high marginality index score (i.e., locations that could have been selected). The third sample limits the regional coverage to rural households in non-program localities with a high marginality index score, within the Central Region. Hence, the selection of samples 2 and 3 is again a clear example of model-based systematic sampling in which knowledge of the selection process is utilized to exclude localities and, thus, households.

Table 3: RPS impact on several outcomes: Bias and impact estimates

| | Food expenditure | | Breast feeding | | Vaccination | | Health check-up | |
|---|---|---|---|---|---|---|---|---|
| Sample | Bias | Impact | Bias | Impact | Bias | Impact | Bias | Impact |
| Experiment | -- | **1224** | -- | **0.134** | -- | **0.083** | -- | **0.169** |
| | | (224) | | (0.077) | | (0.040) | | (0.039) |
| National | -200 | **607** | **0.240** | **0.256** | -0.046 | **0.202** | 0.067 | **0.346** |
| | (107) | (116) | (0.085) | (0.096) | (0.057) | (0.055) | (0.042) | (0.048) |
| National High-Priority | **-257** | **659** | **0.202** | **0.233** | -0.024 | **0.138** | **0.114** | **0.359** |
| | (102) | (116) | (0.074) | (0.089) | (0.056) | (0.040) | (0.043) | (0.050) |
| Central Region H-P | **-257** | **806** | 0.200 | 0.140 | -0.071 | 0.093 | 0.032 | **0.214** |
| | (118) | (122) | (0.138) | (0.124) | (0.047) | (0.056) | (0.052) | (0.056) |

*Notes:* All model-based estimates are using nearest neighbor matching with common support. Bold figures are statistically significant at the 5%-level, standard errors in parentheses.
Source: Handa and Maluccio (2010), Table 1.

In the model-based impact estimation, Handa and Maluccio focus exclusively on propensity score matching. In the estimation of the propensity scores, the authors include both location and household level covariates. As the program only has a specified location level selection the household level variables are probably mainly included to capture the self-selection into the program. Unfortunately, a specific model for such a self-selection is only mentioned in passing. From the selected covariates it is clear that the authors are thinking of a fairly standard poverty profile (consumption) model. Hence the covariates are household demographics, education, working members, dwelling characteristics and possession of durable goods/assets. Both impact estimates and bias estimates are reported for a range of matching estimators for no less than 12 output measures. In Table 3 we report results for 4 of the 12 measures: food expenditure, breast feeding, vaccination, and health check-up.

The interesting results in Table 3 are the changes in the estimated biases and impacts as we move from the National sample to the Central Region high-priority (i.e., eligible) sample. For the average ratio of vaccination the bias is insignificant for all samples and the estimated impact is very close to the experimental result, in particular for the Central Region sample. Likewise, for the average impact on breast feeding and health check-up, the biases decrease, becoming negligible and insignificant in the most restricted sample and the impact estimates are very close to the experimental estimates going from the National to the Central Region sample. The ‘odd-one-out’ is, once again, food expenditure for which there is a fairly constant, substantial and statistically significant bias for all three samples.[8] This result is representative for all three expenditure outcome measures included in the study. Exactly why this

[8] Notice, though, that when the authors apply the matching estimator with a Gaussian kernel the bias is significantly reduced and becomes statistically insignificant in the Central Region sample (these results are not reported here).

is so, is not clear from the study. Therefore, we follow the authors in suggesting caution when applying model-based estimators to evaluate the impact on complex outcome variables, such as (food) expenditure. We return to this suggestion in the next section.

**The impact on earnings of migration from Tonga to New Zealand**

The fourth study we discuss is very different from the three others as it analyses the average income gain from international migration and is, furthermore, based on a natural experiment instead of a well-planned social experiment. Cook et al. (2008) question the validity of the study as the experiment may not be purely random because of non-compliance bias. However, we believe the study highlights other aspects of the importance of having a good model when assessing causal impacts using model-based estimators, and this is the reason why we think of this paper as showing how a thoughtful modeling of the assignment mechanism is the key to obtaining unbiased model-based impact estimates.

Every year a set quota of Tongans are allowed to migrate permanently to New Zealand. More applications are received than the quota allows and a random ballot is used to select from amongst the applicants. This creates a natural experiment which Mckenzie, Gibson and Stillmann (2010) utilize by surveying the applicants over a three year period and estimate the average income gain from migration. Even though not all ballot winners actually did migrate (the non-compliance problem), Mckenzie et al. are able to estimate the average income gain from migration using the survey. The natural experiment estimate is compared to a range of model-based estimates that are in turn based on two different comparison samples. The first is a random sample of 180 non-applicants from the villages where the applicants came from. This survey, named PINZMS, is also conducted and administered by the authors thereby ensuring identical survey instruments. The second comparison sample is the 2003 Tongan Labor Force Survey (TLFS), which is a nationally representative survey with 3,979 individuals aged 18-45 (comparable to the applicant and the special non-applicant samples). The TLFS is much more limited in terms of relevant information compared to the smaller PINZMS survey and the authors argue that it is included in the study to mimic the prevailing practice for model-based impact estimation.

The main results of the estimator comparisons in McKenzie et al. are given in Table 4. The design-based impact estimate is an increase in weakly wages of $NZ 274 which is both economically substantial and statistically significant. An interesting result in the table is that all model-based estimators, conditioning on observables, are (more or less) biased. The relative bias for each estimator is reported in column 4 while a 90%-confidence interval for the bias is reported in the last column. The estimators that do not utilize information on past income have very substantial biases of about 30% of the estimated impact, and the biases are statistically significant. When pre-migration income is included in the information set, the bias drops somewhat and becomes statistically insignificant at the 90% level, but the biases are still about 20 percent of the experimental impact estimate. The substantial change in the bias is interesting as it highlights the importance of including the pre-intervention level of the outcome variable in the information set. The special significance of this variable was an important part of the debate between Smith and Todd (2005a, b) and Dehejia (2005) and it is also stressed in the papers by Cook et al. (2008) and Cook and Steiner (2010). Another noteworthy result is that comparable OLS and matching estimators have biases of the same order of magnitude.

Table 4: Average migration earnings: Impact estimates

| Method | Sample Size | Estimated average difference in weakly wage, ATT ($NZ) | S.E. | Percent difference compared to experiment | Percent positive bias in bootstrap replications | Bootstrapped 90%-confidence interval for difference |
|---|---|---|---|---|---|---|
| Natural experiment | 190 | 274 | 55 | -- | -- | -- |
| OLS, PINZMS | 230 | 347 | 43 | 27 | 95 | [2; 149] |
| Matching, PINZMS | 230 | 350 | 54 | 28 | 97 | [5; 140] |
| OLS, Sample TLFS | 4043 | 358 | 44 | 31 | 99 | [16; 148] |
| Matching Sample TFLS | 4043 | 359 | 47 | 31 | 99 | [16; 147] |
| Pre-migration income | 63 | 341 | 46 | 24 | 93 | [-11; 150] |
| Difference-in-difference | 219 | 334 | 44 | 22 | 87 | [-23; 131] |
| Matching , using past income | 219 | 330 | 59 | 20 | 86 | [-29; 142] |
| IV-Migrant network | 219 | 498 | 234 | 82 | 87 | [-154; 750] |
| IV-Distance to NZIS office | | | | | | |
| All islands | 219 | 309 | 90 | 13 | 68 | [-122; 202] |
| Tongatapu only | 159 | 277 | 90 | 1 | 54 | [-166; 148] |

*Notes:* Sample 1 is the PINZMS survey conducted by the McKenzie et al., Sample 2 is the Tongan Labor Force Survey (TLFS). The Difference-in-difference, matching using past income, and the IV-estimators are all based on the PINZMS. All results use (functions of) a male dummy, married dummy, age, years of education, and a born in Tongatapu dummy as controls. When using the PINZMS, sample height is also included and past income is included in the natural experiment, for the difference-in-difference, and when indicated. The matching estimator is the bias-adjusted single nearest-neighbor matching on multiple covariates.
*Source:* McKenzie, Gibson and Stillman (2010), Tables 2, 4 and 5.

It is only when additional information, which is only available in the PINZMS sample—the distance to the application office—is utilized as an instrument in IV-regressions we find "truly" unbiased model-based estimates in the sense that the point estimates are very close, statistically insignificant, and median unbiased in the bootstrap replications. In fact the last estimate, an IV-regression in which the sample is restricted to only include non-applicants from the main island Tongatapu, is spot on and the bootstrap replications shows that it is also median unbiased. Yet, given that the design-based estimate was known, this result mainly illustrates that a comparable estimate *can be* found, not that it *will be* found in practice.

McKenzie et al. state in the abstract that "non-experimental methods other than instrumental variables are found to overstate the gains from migration by 20-82%, with difference-in-difference and bias-adjusted matching estimators performing best among the alternatives to instrumental variables". We would like to offer a somewhat more general conclusion.

Basically, McKenzie et al. have a very vague model for the selection into migration from Tonga to New Zealand and the wage impact. The covariates in the model-based estimations are what the authors call standard wage equation variables; age, sex, marital status, and years of education. In addition, height is included in some regressions as a measure of health along with a dummy for being born on the main island of Tongatapu to proxy for urban skills. While it is true that this set of controls appear in most individual level wage equations it also appears in almost all regressions of almost any outcome when modeling individual behavior in developing countries. Hence, the only statistical modeling with clear reference to the problem at hand is when the authors discuss using either personal networks in New Zealand or the distance to the application office as instruments for the selection process. Now, as discussed by the authors it is hard to accept that the network, proxied by the number of relatives in New Zealand, is not correlated with the (initial) wage in New Zealand whereby this variable is much better suited as an observable control variable than as an instrument. It is easier to accept the distance to the application office as a proxy for informational migration costs which is not related to the wage levels in any of the countries. Thus, the only explicit model-based estimator for which we would accept the model assumptions is the IV-estimator, and the estimate from that is unbiased.

**The contrast to earlier within-studies**

Finding four studies out of four in which the model-based impact estimates of development interventions are unbiased contrasts with the vast number of earlier studies of job-market interventions in the US that judged non-experimental estimators to be generally biased. This calls for an explanation. What are the sources of difference between the performance of model-based estimators in the context of labor market interventions and in the contexts discussed above?

The first thing to note is that the difference is not that big. Most of the 12 studies reviewed in Glazerman et al. (2003) also find some unbiased model-based estimates. The concern in the studies by, say, LaLonde (1986) and Fraker and Maynard (1987), who both find unbiased model-based impact estimates for women under the Aid to Families with Dependent Children (AFDC) program, is that the model-based estimates are often far from the design-based and that the authors cannot find any systematic tendencies to use for guidance of model-based evaluations. These concerns are stressed and strengthened in Glazerman et al. (2003) as their review also fails to find useful systematic tendencies for the bias in the model-based estimators relating to choice of estimators and samples. Hence, the main new result for development interventions is really that we may be able to see a pattern, such that some reliable guidance can be given.

Meanwhile, two very important differences between the interventions in Mexico and Nicaragua and the twelve job-market interventions described in Glazerman, et al. (2003) is the detailed knowledge of the program assignment mechanism in the two development interventions and the coverage of the interventions. The typical US job-market intervention, say, the

National Supported Work Demonstration (NSW) analyzed in LaLonde (1986), Fraker and Maynard (1987), Dehejia and Wahba (1999) and Smith and Todd (2005a), offers job search assistance, employment-training, or rehabilitation to a fairly rare part of the American population, such as women under the AFDC program, young high-school drop-outs, ex-convicts and ex-offenders. The participation in the program is often voluntary, which is why the model of the assignment mechanism is a self-selection model based on an economic model for expected earnings differentials. Hence, the model-based evaluations of the programs are conditional on the assumption that program participants make a rational choice to participate in the program if they expect to obtain increased earnings from the participation. Considering the participants' past, and our knowledge of the importance of social norms and group identity on labor market behavior (Elster, 1989; Clark, 2003; Stutzer and Lalive, 2004), this assumption and modeling approach could be questioned. And, as the program participants are rare in the population it is actually difficult to find good comparison individuals based on simple observable variables such as sex, age, years of education and area of location—even in large databases with national coverage.

In comparison, the conditional cash transfer programs in Mexico and Nicaragua had detailed information on the assignment mechanism and, importantly, the programs had broad coverage in the target population. Hence, (i) program assignment included both a well described targeting and self-selection, and (ii) the program coverage ensured that the comparison data had good matches for the self-selection part.

The migration earnings intervention is probably the closest comparison to the US job-market interventions and the results for the model-based estimators are also compatible. The point to note is, as already highlighted, that unbiased estimates are obtained when truly model-specific information is included, namely the observation that people are more likely to enter the migration lottery (self-select into the program) when they live close to the migration office.

In sum, we think one can question what the within-studies of US labor market interventions are actually showing. In the many discussions following the initial studies by LaLonde (1986) and Fraker and Maynard (1987) the focus was always on selection and improvement of a model-based estimator, and the results were disappointing as no single estimator came out as unbiased in all cases. However, to us it is quite obvious that the economic model for participation in job-market interventions in the US is in all likelihood misspecified. If this is so, what the many studies have proven is the truism that model-based estimators are biased when the underlying models are wrong—not that model-based estimators are biased in general. If the assignment mechanisms are better understood and described in development interventions, as we believe they can be, then the model-based estimators will provide unbiased impact estimates.

## Discussion

In this section we try to transform the results of the within-studies and the lessons learnt from other surveys, such as Cook et al. (2008), Cook and Steiner (2010), and Steiner et al. (2010), to a few guiding principles for conducting and assessing model-based impact evaluations. Several recent books go more into depth with practical guidance and rather than repeating that advice we refer the readers to, say Gertler et al. (2011), and focus on relatively simple principles for model-based impact evaluations.

First and foremost, it should be clear that the model-based approach requires a model. This implies that evaluators must focus on having an explicit, transparent and well-described assignment mechanism. Such a description is the model-based counterpart to the controlled randomization of program participation in design-based evaluations. An added benefit of this focus is that donors and policy makers must describe their beneficiary assignment procedures in detail and allow evaluators to test the targeting of the program as an integral part of the impact evaluation. Hopefully, such a requirement will, over time, lead to improved targeting of interventions and also to a better integration of program planning and evaluation as the targeting should ideally be described already in the program planning phase.

Next, the evaluator must formulate an explicit model, or framework, for the specific program which includes a description of both the impact and the selection. The model must be much more specific than a list of possible covariates which are expected to correlate with both selection and outcome. Ideally, the framework should lead to a reasonable statistical model. The benefit of a well-specified model is first of all that it guides the choice of estimator: if the participation assignment can be described by a fairly simple mechanism for which observable variables are available, then “conditioning on observables” (i.e., regression and/or matching) is probably adequate. In contrast, if the program is mainly a general offer to a wide range of individuals (and families) such that participation is mainly determined by self-selection then it is probably better to look for and describe instrumental variables, i.e., variables that are expected to determine the program participation, but not the program outcome.

Obviously, only a thorough understanding and description of the intervention will lead to an adequate and well-argued choice of methods and controls/instruments. Moreover, the kitchen-sink approach to inclusion of control variables is in our view often just a replacement for precise knowledge and, therefore, not a valuable general approach—despite the many papers stating that having many control variables is better than having few.

Turning to the comparison group, two of the studies show that national surveys *can be* used, but differences in survey instruments create significant biases, indicating that evaluation specific surveys are preferable to general purpose (national) surveys. Moreover, even though matching and regression are tools for balancing the data it is more efficient, in terms of reducing selection bias, to use the assignment procedure for model-based systematic sampling within a larger comparison group to exclude irrelevant individuals prior to the matching/regression.

One of the new insights from the studies of development interventions, which has not been given much thought before, is the choice of output variables. Simple outcome variables, such as binary indicators of school attendance or regular health check-up attendance, are easier to model relative to more complex outcome variables, such as income and consumption expenditure. There may be two reasons for this. One is related to the survey instrument and questionnaires in general, as it is easier to have consistent recordings of simple measures. If this is the case, a clear recommendation for evaluations using general purpose surveys for comparison group construction is to focus on simple output measures. Another possibility is related to statistics, as the good result for the binary indicators may be because the population variances in the simple outcome variables are smaller than the variances in the more complex variables. If this is the case, the recommendation is not to prefer simple measures as such but, instead, to use outcome measures with smaller variance. Note, however, that evaluating the impact on outputs with a smaller variance should always be preferred regardless of the choice of evaluation method, be it model-based or design-based.

More research related to the choice of output measures is clearly needed. However, a simple recommendation is to include explicit power calculations as part of the evaluation. If the outcome variables have large variances and the samples are small then significance tests have little meaning (low power) as random variation may dominate the impact and, worse, slightly unbalanced comparison groups may be disastrous for the impact estimates.

Finally, regarding the choice of estimator, there is no omnipotent technique. Hence, the model-based estimator must first and foremost be chosen in accordance with the statistical model and the data at hand. If conditioning on observables is the choice then results of both regression and matching should be reported. Regression results are transparent and easy to interpret compared to the matching techniques and most studies have found that the choice between regression and matching matters little for the results relative to the choice of control variables. Thus, if the regression and matching results differ substantially these discrepancies should be investigated and explained.

## Concluding remarks

In this paper we have argued that impact evaluations of development interventions for a large part will continue to be based on observational data rather than on social experiments. The reasons are economic, political, and ethical: experiments are often too time consuming and expensive relative to their benefits and randomization is rarely a politically acceptable allocation mechanism when targeting poor people in developing countries. While this does not imply that randomized controlled trials should never be conducted it does indicate that we need to learn more about how to conduct and assess non-experimental impact assessments in practice.

We have presented and discussed four recent studies comparing experimental and non-experimental impact estimates of three different development interventions. The four studies all report a number of cases in which the non-experimental estimates are biased and also several cases where the non-experimental causal estimates are unbiased.

The most important result coming out of the four studies is that when the selection procedure is well described, and well utilized, the non-experimental estimators are found to be unbiased. We advocate for using this result in practical evaluation studies by focusing on a thorough description of the assignment mechanism and formulating a statistical model for this mechanism that utilize the detailed information. Here, it is also important to note that the current common practice of not modeling the outcome variable is not necessary a good idea. While treating the outcome variable as a fixed attribute ensures theoretical purity of the statistical model and a clean definition of causal effects the possible effects of sample selection are best discussed when a statistical model for the outcome variable is formulated jointly with the selection model.

For the outcome variable we observe that unbiased impact estimates are more often obtained for simple outcome variables, such as school attendance, than for complex variables, like food expenditure. We speculate that this could be due to better agreement between different surveys (i.e., reducing measurement error) or to a lower variance in the simple outcome variables relative to the complex outcome variables used in the studies.

Finally, we argue that if the non-experimental estimator is conditioning on observables, then both matching and regression results should be reported. The reason is that the choice of estimator matters less than the choice of controls and, therefore, the different estimators may be used as an informal specification (omitted variables) "test".

Clearly, these simple suggestions can never ensure unbiased impact estimates, and much more can be learned about modeling selection mechanisms and outcomes. Therefore, in the future we would also like to see more studies, which compare social experiments in developing countries with non-experimental evaluations of the same interventions.

## References

Banerjee, A., & Duflo, E. (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. PublicAffairs.

Barrett, C. B., & Carter, M. (2010). The Power and Pitfalls of Experiments in Development Economics. Some Non-random Reflections. *Applied Economic Perspectives and Policy, 32*(4), 515-548.

Binder, D. A., & Roberts, G. A. (2003). Design-based and Model-based methods for estimating model parameters. In R. Chambers, & C. Skinner (Eds.), *Analysis of Survey Data* (pp. 29-48). Wiley.

Buddelmeyer, H., & Skoufias, E. (2004). *An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA.* IZA Discussion Paper No. 827 Institute for the Study of Labor.

Cartwright, N. (2007). Are RCTs the Gold Standard? *BioScience, 2*(1), 11-20.

Clark, A. (2003). Unemployment as a Social Norm: Psychological Evidence from Panel Data. *Journal of Labor Economics, 21*(2), 323-351.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Wiley.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons. *Journal of Policy Analysis and Management, 27*(4), 724–750.

Cook, T. D., & Steiner, P. M. (2010). Case Matching and the Reduction of Selection Bias in Quasi-Experiments: The Relative Importance of Pretest Measures of Outcome, of Unreliable Measurement, and of Mode of Data. *Psychological Methods, 15*(1), 56-68.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature, 48*, 424-455.

Dehejia, R. (2005). Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics, 125*, 355-364.

Dehejia, R., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association, 94*(448), 1053-1062.

Diaz, J. J., & Handa, S. (2006). An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program. *Journal of Human Resources, 41*(2), 319-345.

Elster, J. (1989). Social Norms and Economic Theory. *Journal of Economic Perspectives, 3*(4), 99-117.

Fraker, T., & Maynard, R. (1987). The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs. *Journal of Human Resources, 22*(2), 194-227.

Gertler, P., Martinez, S., Premand, P., Rawlings, L., & Vermeersch, C. (2011). *Impact Evaluation in Practice.* The World Bank.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental Versus Experimental Estimates of Earnings Impacts. *The ANNALS of the American Academy of Political and Social Science, 589*(1), 63-93.

Glewwe, P., Kremer, M., Moulinc, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics, 74, 251– 268*.

Greenberg, D. H., Michalopoulos, C., & Robin, P. K. (2006). Do experimental and nonexperimental evaluations give different answers about the effectiveness of government-funded training programs? *Journal of Policy Analysis and Management, 25*(3), 523-552.

Handa, S., & Maluccio, J. A. (2010). Matching the Gold Standard: Comparing Experimental and Nonexperimental Evaluation Techniques for a Geographically Targeted Program. *Economic Development and Cultural Change, 58*(3), 415-447.

Hausman, J. (1978). Specification Tests in Econometrics. *Econometrica, 46*, 1251–1272.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47*(1), 153-161.

Heckman, J. J., & Robb, R. J. (1985). Alternatice Methods for Evaluating the Impact of Interventions. An Overview. *Journal of Econometrics, 30*, 239-267.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. In J. J. Heckman, & E. E. Leamer (Eds.), *Handbook of Econometrics* (pp. 4779-4874). Elsevier.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments. In J. J. Heckman, & E. Leamer (Eds.), *Handbook of Econometrics* (Vol. 6B, pp. 4878-5143). Elsevier.

Heckman, J. J., Ichimura, H., Smith, J. A., & Todd, P. E. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica, 66*(5), 1017-1098.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature, 47*(1), 5-85.

Karlan, D., & Appel, J. (2011). *More Than Good Intentions: How a New Economics is Helping to Solve Global Poverty*. Dutton.

LaLonde, R. (1986). Evaluating the econometric evaluation of training with experimental data. *American Economic Review, 76*(4), 604-620.

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *American Economic Review, 73*, 31-43.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *The American psychologist, 48*(12), 1181-1209.

Maluccio, J. A., & Flores, R. (2005). *Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan Red de Proteccón Social.* Research Report, International Food and Policy Research Institute.

McKenzie, D., Gibson, J., & Stillman, S. (2010). How Important Is Selection? Experimental vs. Non-Experimental Measures of the Income Gains from Migration. *Journal of the European Economic Association, 8*(4), 913-945.

Ravallion, M. (2008). Evaluating Anti-Poverty Programs. In T. P. Schultz, & J. A. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4, pp. 3787-3846). Elsevier.

Rodrik, D. (2009). The New Development Economics: We Shall Experiment, But How Shall We Learn? In J. Cohen, & W. Easterly (Eds.), *What Works in Development? Thinking Big and Thinking Small.* Brookings Institutions Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics, 6*(1), 34-58.

Rubin, D. B. (1990). Formal Models of Statistical Inference for Causal Effects. *Journal of Planning and Inference, 25*, 279-292.

Rubin, D. B. (1991). Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. *Biometrics, 47*(4), 1213-1234.

Savedoff, W. D., Levine, R., & Birdsall, N. (2006). *When Will We Ever Learn? Improving Lives Through Impact Evaluation.* Center for Global Development.

Skoufias, E. (2001). *PROGRESA and its Impacts on the Human Capital and Welfate of Households in Rural Mexico: A Synthesis of the Results of an Evaluation by IFPRI.* Mimeo, IFPRI.

Smith, J. A., & Todd, P. E. (2005a). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125,* 305-353.

Smith, J. A., & Todd, P. E. (2005b). Rejoinder. *Journal of Econometrics, 125,* 365-375.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. (2010). The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies. *Psychological Methods*, *15*(3), 250-267.

Stutzer, A., & Lalive, R. (2004). The Role of Social Work Norms in Job Searching and Subjective Well-Being. *Journal of the European Economic Association*, *2*(4), 696-719.

Todd, P. E. (2008). Evaluating Social Programs With Endogenous Program Placement and Selection of the Treated. In T. P. Schultz, & J. A. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4, pp. 3848-3894). Elsevier.