

USING RANDOMIZED EXPERIMENTS AND
STRUCTURAL MODELS FOR ‘SCALING UP’:
EVIDENCE FROM THE PROGRESA EVALUATION

Orazio Attanasio
Costas Meghir
Miguel Szekely



EDePo

**Centre for the Evaluation
of Development Policies**

THE INSTITUTE FOR FISCAL STUDIES
EWP04/03

Using randomized experiments and structural models for ‘scaling up’. Evidence from the PROGRESA evaluation.*

Orazio P. Attanasio[†], Costas Meghir[‡] and Miguel Szekely[§]

May 2003

Abstract

The evaluation of welfare programs and more generally government or international organisms interventions is often posed as a one off question, in that evaluators ask whether a specific intervention achieves a specific objective in a specific situation. However, recently, the more general question of whether results from a given studies can be used to predict the effect of different interventions in, possibly, different contexts has received a considerable amount of attention. The usefulness of such an exercise, if successful, is obvious. The ability to extrapolate success stories and avoid failures in different situations would obviously be highly desirable. Unfortunately, a rigorous and successful extrapolation is extremely difficult. Perhaps such difficulties should not be unexpected, given the problems that often one encounters in establishing the effects of social programs in non experimental settings. In this paper we discuss at length the issues involved with the evaluation of social interventions and with the attempts at ‘scaling them up’. In particular, we discuss the relative merits of non-parametric evaluation strategies that rely on (possibly experimental) exogenous variation to estimate the impact effects and of more structural approaches. The difference between the two approaches is particularly relevant when one comes to the issue of ‘extrapolation’ and ‘scaling up’. In principle one could consider two types of extrapolation and scaling up. First, one might want to predict the effects of a program that is different from the one that was evaluated or the effect of changing some aspects of the program evaluated. Second, one might want to predict the effect of exporting an existing program from a population where its effects were evaluated (evaluation population) to a different population (implementation population). In what follow we focus on the latter problem and discuss the former only marginally. After considering extensively the conceptual and technical issues involved with this type of exercises, we apply

*Paper prepare for the ABCDE conference, Bangalore May 21-22.

[†]UCL, IFS and NBER

[‡]UCL and IFS

[§]Sedesol, Mexico

the ideas we discuss to the results from the evaluation of PROGRESA, a large welfare program in Mexico, for which a randomized evaluation sample is available and has been extensively studied. In particular, we divide the seven Mexican states in which the evaluation was carried out in two groups and check to what extent the results in one group can be extrapolated to the other. The advantage of such a strategy is that one can compare the extrapolation results (based on a structural model) with the actual 'ex-post' evaluation that can be carried out either by simple comparison of means or by structural methods.

1 Introduction

The evaluation of welfare programs and more generally interventions by government or international organisations is often posed as a one off question, in that evaluators ask whether a specific intervention achieves a specific objective in a specific situation. However, recently, the more general question of whether results from given studies can be used to predict the effect of different interventions in, possibly, different contexts has received a considerable amount of attention. The usefulness of such an exercise, if successful, is obvious. The ability to extrapolate success stories and avoid failures in different situations would obviously be highly desirable. Unfortunately, a rigorous and successful extrapolation is extremely difficult. Perhaps such difficulties should not be unexpected, given the problems that often one encounters in establishing the effects of social programs in non experimental settings.

In this paper we discuss the issues involved with the evaluation of social interventions and with the attempts at 'scaling them up'. In particular, we discuss the relative merits of non-parametric evaluation strategies that rely on (possibly experimental) exogenous variation to estimate the impact effects and of more structural approaches. The difference between the two approaches is particularly relevant when one comes to the issue of 'extrapolation' and 'scaling

up'. In principle one could consider two types of extrapolation and scaling up. First, one might want to predict the effects of a program that is different from the one that was evaluated or the effect of changing some aspects of the program evaluated. Second, one might want to predict the effect of exporting an existing program from a population where its effects were evaluated (evaluation population) to a different population (implementation population). In what follow we focus on the latter problem and discuss the former only marginally.

After considering extensively the conceptual and technical issues involved with this type of exercises, we apply the ideas we discuss to the results from the evaluation of PROGRESA, a large welfare program in Mexico, for which a randomized evaluation sample is available and has been extensively studied. In particular, we divide the seven Mexican states in which the evaluation was carried out in two groups and check to what extent the results in one group can be extrapolated to the other. The advantage of such a strategy is that one can compare the extrapolation results (based on a structural model) with the actual 'ex-post' evaluation that can be carried out either by simple comparison of means or by structural methods. The extrapolation is based on a structural model. Rather than building a new model, we use the one recently used by Attanasio, Meghir and Santiago (2001) to evaluate PROGRESA.

The rest of the paper is organized as follows. In Section 2, we discuss the conceptual and technical issues related to the 'scaling up' of evaluations. In Section 3, we describe the structural model we propose to use and the essence of the exercise we perform on the Mexican data. In section 4, we describe some details of the program and of the evaluation sample. In section 5 we report the results of our estimation and of our main simulation exercise. Section 6

concludes the paper.

2 Issues in Scaling up

The basic discussion that follows considers a context where the impact of a policy varies by observable and possibly by unobservable characteristics. If the impact is constant many of the issues we discuss here become much simpler. When considering the possibility of applying the findings from one evaluation to a different area, we need to consider a number of factors. These factors have to do with the way that the areas differ as well as with the way that the evaluation took place originally, i.e. the way it was designed and the extent to which it has features that make it generalizable.

In terms of the differences across areas we need to consider the following factors: First, the distribution of observable characteristics may be different. It may for example be the case that the area where the program was evaluated is wealthier, or contains more educated individuals than the area in which we wish to implement the policy. Second preferences and unobserved characteristics may differ. Third, the institutions may be different in the two areas. For example there may be different laws governing child labour, or the existing laws may be implemented with greater vigour in one area *vis a vis* the other.

The design of the original evaluation will define in many ways what can be learned about other setups. Randomized experiments, natural experiments and matching based evaluations may all identify different parameters that are more or less relevant for other contexts. In what follows we discuss each of these issues. We argue that the best chance in practice of a reliable scaling up can be obtained if a reliable and well identified structural model is available. It is very

hard to transfer results from one setting to another without reference to any theoretical context. Having said that, however, we will also argue that, at least conceptually and with enough data, one could obtain many of the evaluation results obtained from a structural model, using a non-parametric approach and many randomized experiments. The main problem with this strategy is the fact that such a wealth of data is typically impossible to obtain.

To inform the discussion and establish notation, consider the following simple problem: Suppose we are considering the impact of a given policy on an outcome variable Y . Such a policy could be a conditional cash transfer, such as the ones distributed by PROGRESA and the outcome variable the probability that a child enrolls in school. Suppose that an individual i has outcomes Y_i^1 under the policy and the same individual has outcome Y_i^0 if she is not exposed to the effects of the policy. An evaluation will at best estimate some aspects of the distribution of the gains $Y_i^1 - Y_i^0$. What precise aspects of this distribution will be identified will depend very much on the structure of the evaluation. We also introduce the characteristics X_i and an assignment rule for the policy $D_i = 1$ or $D_i = 0$. Given these, examples of parameters that are the objects of evaluation are the average treatment effect (ATE) $E(Y_i^1 - Y_i^0)$, the impact of treatment on the treated $E(Y_i^1 - Y_i^0 | D_i = 1)$ and versions of these conditional on characteristics X . (see Heckman, Lalonde and Smith, 1999)

Differences in the Distribution of observable characteristics. Suppose that a program has been evaluated in a particular area or country, for example by a randomized experiment, and ATE has been estimated. Randomization of program assignment, of course, gives the possibility of estimating the

counterfactual conditional and unconditional distributions of the outcome variable Y . We can therefore directly compute $E(Y_i^1 - Y_i^0)$ or $E(Y_i^1 - Y_i^0|X_i)$. Now suppose that we can characterize individuals by a vector of (outcome relevant) characteristics X_i whether observable or unobservable. In this case ATE can be written as

$$ATE = \int E[Y_i^1 - Y_i^0|X_i]dF(X_i)$$

where the expectation is taken over some random noise which is assumed identical across all areas. Central to the argument is the distribution of characteristics in the evaluation area, $F(X_i)$. If $F(X_i)$ differs from the distribution of characteristics in the area where we now want to implement the policy, and if the impacts vary with X_i then the ATE parameter we have estimated from the evaluation has little to say about the impact of the program in the new area. There are two main issues relating to this: First the question is whether the distribution of characteristics in the evaluation area and the new area have the same support, i.e. whether all types of people characterised by X that can be found in the evaluation area can also be found in the implementation area. If the implementation area includes individuals who do not exist in the evaluation area, the evaluation cannot say much about the impact of the policy on these people, except through some form of extrapolation, based on parametric assumptions. In other words we can only hope to predict the impact if we have some form of credible model capable of extrapolating in different circumstances. In any case such results will nearly always be steeped in controversy.

A second less serious problem is one that arises when the distribution of characteristics is different over the common support. In this case knowledge of the ATE in the evaluation area is insufficient to estimate the impact in the

implementation area.. Here we need to know $E[Y_i^1 - Y_i^0|X_i]$. An implementation area ATE can then be estimated by averaging using the distribution of characteristics in the implementation area (over the common support). While this is easy to say, it is often the case that these conditional $ATEs$ are not well estimated if anything because X can be multidimensional and because the evaluation sample sizes may not be large enough. Again here some form of parametric assumptions may prove to be very useful.

The distribution of unobserved characteristics may differ. As discussed above, if the distribution of characteristics is different and if this is relevant for the impact, the outcomes must be reweighed to match the distribution in the implementation area. However, this is not directly possible when some of the characteristics are unobservable. In this case knowledge of $E[Y_i^1 - Y_i^0|X_i]$ (call it $ATE(X_i)$) for observable X_i in the evaluation areas is not sufficient for knowledge of $ATE(X_i)$ in the implementation area because

$$ATE(X_i) = \int E[Y_i^1 - Y_i^0|X_i, u_i]dF(u_i|X_i)$$

and hence the result depends on the distribution of $F(u_i|X)$ which could be area specific. Overcoming this problem is of course very difficult. A parametric structural approach may allow us to identify $F(u_i|X_i)$ and $E[Y_i^1 - Y_i^0|X_i, u_i]$ separately in the evaluation area. However, we may still have no way of identifying the distribution of unobservables in the implementation area.

Institutions and aggregate conditions may differ. Technically this problem is similar to the one where the distribution of unobservables are the same. From a practical point of view it may be possible to identify what we

believe are key institutional differences and inform policy on the basis of judgement. However, this is a key problem for scaling up, when the evaluation and the implementation areas are quite different. No obvious formal solution can be found, unless, we can characterize the institutional differences with a small set of observable aggregate variables and we then have at hand evaluations over a large set of areas with sufficient variability in these characteristics to be able to identify similar environments to the one we now wish to implement the policy. This is of course a similar problem to the support problem mentioned earlier.

In addition to institutional differences, we need to consider differences in aggregate macroeconomic conditions. Wages and generally labour market opportunities may affect the outcomes of a policy. Consequently implementation in an environment with different macroeconomic conditions may give rise to different impacts. This of course concerns both the transfer of the policy from one area to another as well as from one time period to another.

The precise nature and intensity of the policy may differ. It is unlikely that an identical policy will be implemented in a new area. For example if the policy concerns a school subsidy different levels of the subsidy may be envisaged as well as different ways of means testing. In fact the evaluation itself may give certain hints towards improving the design to increase the impacts or to reduce costs. However the evaluation itself will usually yield only a very specific result, relating to the particular rules and intensity envisaged in that case. To go beyond the confines of the particular program evaluated we need to combine the evaluation data with some structural parametric model, that will allow us, subject to assumptions, to glean information from other sources that

may be indirectly informative about the impact of the policy.

If one does not want to rely on the assumptions necessary to estimate and implement a structural model, to predict the likely effects of changes in the program, one needs exogenous variation in all the dimensions of interests. For instance, in the case of PROGRESA, beside the evaluation of the effect of the particular grant on enrollment of children of a certain age, one might be interested in evaluating the effect of changing the level of the grant for different groups of children. If the localities in the PROGRESA sample had been randomly assigned to different versions of the program, one could have evaluated in a fully non parametric way the effect of changes in the program. Similarly, one could think of estimating the effects of different changes in the program by the appropriate randomization scheme. Of course, the difficulty is that it is unlikely to have data that would allow the evaluation of any aspect of interest. The use of a (carefully constructed) structural model is one way of making a parsimonious use of any exogenous variation available in data to extrapolate in different dimensions. We come back to this issue below. At this point, however, it is worth stressing that at least conceptually one could use appropriately constructed experimental data to estimate any elasticity that one would infer from a structural model. The limitation of this strategy, however lies in the availability of data.

The nature of the evaluation. The precise nature of the original evaluation and which parameter has been identified is a central issue in all this. Quite frequently evaluations are designed to estimate the parameter, treatment on the treated, i.e. $E(Y_i^1 - Y_i^0 | D_i = 1)$ or the local average treatment effect (*LATE*),

rather than the average treatment effect that we have been supposing up to now. The former depends on the precise assignment rule to the policy, while the latter depends on both the assignment rule and which precise policy is being considered. Attempting to learn about implementation in the new environment from such parameters can be very hard indeed because we would have to model the way in which the assignment rules differ. Thus scaling up is more likely to be successful with an evaluation design that allows the estimation of average treatment effects, subject of course to the provisos mentioned above.

The ethics and political economy of evaluations. As we mentioned above, a big obstacle to the exclusive use of non-parametric methods in the evaluation of policy intervention is the availability of experimental data in which the assignment of individuals (or localities) to programs (and possibly to different kinds of programs) is random. Randomized data sets collected explicitly for evaluation purposes are few and far between and the proposal of their construction is typically met with strong resistance from politicians and administrators. The reasons for this reluctance are many. Obviously there are some important ethical issues. Excluding individuals from programs that researchers believe to be effective in some important dimensions is obviously problematic. And these difficulties are typically amplified in the political process. Additional problems arise from the short horizon that seems to be relevant for many politicians.

However, experience has shown that there are some dimensions that can be used to overcome these resistances. Typically, large programs take time to have full coverage. The expansion phase can then be used so that, instead of randomizing in terms of ‘who gets the program and who does not’, one ran-

domizes on the timetable of expansion, so that some individuals or communities are randomly assigned to the ‘beginning of the line’, while others are placed at the ‘end of the line’. The PROGRESA evaluation constitutes a good example of this strategy. The control communities were not excluded from the program for ever, but they were put among those communities were the program, because of budgetary limitations, arrived late in the expansion program, roughly two years after the program was first implemented in the ‘treatment’ communities included in the evaluation sample.¹ Of course, this strategy leads to a different set of problems: if individuals in the ‘control’ sample know that they are going to get the program and react to this type of information, this could contaminate the evaluation. This type of anticipation effects among control individuals or communities constitute another element in favour of the use of structural models: one can explicitly introduce the information about the future implementation of a program into the structural model. This is not an issue in completely static problems.

An alternative strategy is to work on pre-program pilots that could be used to design the detail of a specific program. This type of studies are particularly interesting for at least two reasons. First, from a political point of view it might be easier to introduce exogenous or random variation in the implementation of the program in a small set of areas or for a small set of individuals than for the population at large. Second, it might be easier, within a pilot and with the explicit purpose of fine tuning the details of a specific program, to experiment

¹Another distinguishing feature of the PROGRESA evaluation was that the randomization was done at the locality rather than the individual level. In that particular situation this procedure was preferable for at least two reasons. First, by randomizing at the community level one can, in principle, estimate spillover and general equilibrium effects induced by the program. Second, from a political point of view, the random exclusion of a number of individuals in small communities where the program was operating, might have been even harder to sustain.

with various version of the program itself. The main limitation of these studies, is likely to be the short horizon over which they need to be performed.

A final interesting situation is the one where there are important budgetary limitations that prevent the allocation of the program to all applicants. In this case, the random allocation of the program to individuals may arise as the fairest and most efficient way of allocating the program. This was the case, for instance, in a voucher program in Colombia that was recently analyzed by Angrist et al. (2003).

To make the discussion in this section concrete we now move to the analysis and evaluation of a specific welfare program: PROGRESA in Mexico. As we mentioned, PROGRESA has been the subject of an extensive evaluation and, given the availability of a large and good quality data base, has been studied extensively. The aim of the following sections is not to provide an additional evaluation. This has been done, using a variety of techniques, in many different papers. Rather, in what follows we take PROGRESA as a specific example to illustrate the various issues we discussed and in particular how, under what conditions, and what which limitations, one can use a structural model to evaluate a welfare program and to scale up the results obtained to different situations.

3 A structural model for evaluation and Scaling up

The setting we consider is one where the program to be evaluated is a school subsidy program such as PROGRESA in Mexico. We will take it that there has been a randomize trial which has split up the relevant population to a treatment group and a control group. As has been the practice, one treatment with its

entire set of rules is compared to no treatment. Thus the randomized trial can provide the impact of the policy (seen as a whole) on average and by sub group depending on the age of the child or the school grade and so on. As such there is little we can learn for other but the most similar settings, i.e. those that can be considered as (stratified) random samples from the identical population. To go further we need to use a model based on assumptions about behaviour. This will be identified, partly thorough the randomized experiment and partly through extra assumptions about the validity of certain cross section assumptions. This model will also indirectly highlight the areas where extra randomized evaluation could usefully inform policy.

We use a simple dynamic school participation model that was developed in a recent paper by Attanasio, Meghir and Santiago (2001) (AMS01 from now on). In this paper we provide only one version of the model without discussing or justifying extensively the many assumptions made along the way. These discussions and alternative specifications of the model can be found in AMS01.

Each child (or his/her parents) decide whether to attend school or to work taking into account the economic incentives involved with such choices. We assume that children have the possibility of going to school up to age 17. All formal schooling end by that time. In the data, almost no individuals above age 17 are in school. We assume that children who go to school do not work and viceversa. We also assume that children necessarily choose one of these two options. If they decide to work they receive a village/education/age specific wage. The model we consider is dynamic for two main reasons. First, the fact that one cannot attend regular school past age 17 means that going to school now provides the option of completing some grades in the future: that is a six

year old child who wants to complete secondary education has to go to school (and pass the grade) every single year, starting from the current. This source of dynamics becomes particularly important when we consider the impact of the Progresa grants. Second, we allow for state dependence: The number of years of schooling affects the utility of attending in this period. We discuss this issue at length below.

3.1 The basic framework

The structure of the model is as follows. In each period, going to school involves pecuniary and non-pecuniary costs, in addition to losing the opportunity of working for a wage. The current benefits come from the utility of attending school and possibly, as far as the parents are concerned, by the child-care services that the school provides during the working day. As mentioned above, the benefits are also assumed to be a function of past attendance. The costs of attending school are the costs of buying books etc. as well as clothing items such as shoes. There are also transport costs to the extent that the village does not have a secondary school. For households who are entitled to Progresa and live in a treatment village, going to school involves receiving the grade and gender specific grant.

As we are currently using a single cross section, we use the notation t to signify the age of the child in the year of the survey. Variables with a subscript t may be varying with age. Denote the utility of attending school for individual i in period t who has already attended ed_{it} years as

$$u_{it}^s = \mu_i + a'z_{it} + b ed_{it} + 1(p_{it} = 1)\beta^p x_{it}^p + 1(s_{it} = 1)\beta^s x_{it}^s + \varepsilon_{it}$$

where z_{it} relates to a number of taste shifter variables, including parental back-

ground and age. The variable $1(p_{it} = 1)$ denotes attendance in primary school, while the variable $1(s_{it} = 1)$ denotes attendance in secondary school. x_{it}^p and x_{it}^s represent factors affecting the costs of attending primary school and secondary school respectively. These factors may interact with other characteristics, such as age or parental education; The term ε_{it} represents a logistic error term which is assumed independently and identically distributed over time and individuals. Notice that the presence of ed_{it} introduces an important element of dynamics that we discuss below. Finally, the term μ_i represents unobservables which we assume have a constant impact over time. As we discuss below, we will be assuming that μ_i is a discrete random variable whose points of support and probability distribution we estimate.

The utility of not attending school is denoted by

$$u_{it}^w = \vartheta_i w_{it}$$

where w_{it} are (potential) earnings when out of school. The wage is a function (estimated from data) of age and educational attainment as well as village of residence. ϑ_i is a random variable, representing heterogeneity in the sensitivity of child i 's decision to the wage. When we consider this additional form of heterogeneity, we assume that ϑ_i is a discrete random variable whose point of support and probability distribution we estimate along with those of μ_i . These are the unobserved characteristics which, as mentioned earlier, may or may not have the same distribution in the new implementation area.

The Progresa grant can be easily added to this framework. Let's $g(ed_{it}, z_i^p, s)$ denote the grant a child in grade ed_{it} receives if he is a beneficiary ($z_i^p = 1$) and goes to school ($s = 1$). Then the utility of going to school will be:

$$u_{it}^s = \mu_i + a'z_{it} + b ed_{it} + 1(p_{it} = 1)\beta^p x_{it}^p + 1(s_{it} = 1)\beta^s x_{it}^s + \theta g(ed_{it}, z_i^p, s) + \varepsilon_{it}$$

where the parameter θ reflects the effect that the grant has on the relative choice between school and work. The relative size of this parameter and the one on the wage is of some interest. A model with completely selfish parents would predict a coefficient on the grant of the same size as the one on the wage. Notice that in the absence of the exogenous variation in the availability of the grant induced by the randomization, one would be forced to estimate the effect of the program through the coefficient on the wage. This is, for instance, the strategy followed by Todd and Wolpin (2003) and would be the only possible alternative if one wanted to estimate the effect of the program before its implementation. The availability of the randomization allows us to estimate a richer structural model in that it allows for differences between the effect of the grant and that of the wage.

After age 17, we assume individuals work and earn wages depending on their level of education. However, the number of choices open to the individual after school include working in the village, migrating to the closest town or even migrating to another state. Since we do not have data that would allow us to model these choices (and schooling as a function of these) we model the terminal value function simply as a quadratic function of years of schooling, with the parameters to be estimated alongside the other parameters of the model.²

Since the problem is not separable over time, schooling choice involves com-

²We have used some information on urban and rural returns to education at the state level along with some information on migration in each state to try to model such a relationship. Unfortunately, we have no information on migration patterns and the data on the returns to education are very noisy. This situation has motivated our choice of estimating the returns to education that best fit our education choices.

paring the costs of schooling now to its future and current benefits. The latter are intangible preferences for attending school including the potential child-care benefits that parents may enjoy.

There are two sources of uncertainty in our model. The first is an iid shock to schooling costs, modelled by the (logistic) random term ε_{it} . Given the structure of the model, having a logistic error in the cost of going to school is equivalent to having two extreme value errors, one in the cost of going to school and one in the utility of work. Although the individual knows ε_{it} in the current period, she does not know its value in the future. Since future costs will affect future schooling choices, indirectly they affect current choices. Notice that the term μ_i , while known (and constant) for the individual, is unobserved by the econometrician.

The second source of uncertainty originates from the fact that the pupil may not be successful in completing the grade. If a grade is not completed successfully, we assume that the level of education does not increase. We assume that the probability of failing to complete a grade is exogenous and does not depend on effort or on the willingness to continue schooling. We allow however this probability to vary with the grade in question and with the age of the individual and we assume it known to the individual.³ We estimate the probability of failure for each grade as the ratio of individuals who are in the same grade as the year before at a particular age. Since we know the completed grade for those not attending school we include these in the calculation - this may be important since failure may discourage school attendance. We denote by $I \in \{0, 1\}$ the random increment to the grade which results from attending school at present. If successful, then $I = 1$, otherwise $I = 0$. We denote the probability of success

³Since we estimate this probability from the data we could also allow for dependence on other characteristics.

at age t for grade ed as $p_t^s(ed_{it})$.

Thus the value of attending school for someone who has completed successful ed_i years in school and is of age t already and has characteristics z_{it} is

$$V_i^s(ed_i, t|z_{it}) = u_{it}^s + \beta \{ p_t^s(ed_i + 1) E \max [V_i^s(ed_i + 1, t + 1), V_i^w(ed_i + 1, t + 1)] + (1 - p_t^s(ed_i + 1)) E \max [V_i^s(ed_i, t + 1), V_i^w(ed_i, t + 1)] \}$$

where the expectation is taken over the possible outcomes of the random shock ε_{it} . The value of working is similarly written as

$$V_i^w(ed_i, t|z_{it}) = u_{it}^w + \beta E \max \{ V_i^s(ed_i, t + 1), V_i^w(ed_i, t + 1) \}$$

The difference between the first terms of the two equations reflects the current costs of attending, while the difference between the second two terms reflects the future benefits and costs of schooling. Finally the parameter β represents the discount factor. In practice, since we do not model savings and borrowing explicitly this will reflect liquidity constraints or other factors that lead the households to disregard more or less the future.

3.2 Estimation

In terms of estimation, the problem in the absence of unobserved heterogeneity ($\mu_i \equiv \mu, \forall i$) other than through the iid shock ε_{it} , is relatively simple. The likelihood function is based on the probability of attending school that takes the form:

$$P(Attend_{it} = 1 | z_{it}, x_{it}^p, x_{it}^s, ed_{it}, wage) = F \{ u_{it}^s - u_{it}^w - \beta [E \max \{ V_i^s(ed_i + I, t + 1), V_i^w(ed_i + I, t + 1) \} - E \max \{ V_i^s(ed_i, t + 1), V_i^w(ed_i, t + 1) \}] \}$$

where the expectation is taken over both ε and I where relevant.

The difference between the (current) values of going to school and working will reflect both the pecuniary trade-offs (the effect of the wage and the cost of going to school) and other relevant factors, such as the dis-utility of work and (possibly) the utility of going to school. Notice that the most general version of our model allows these effects to be heterogeneous across individuals through the terms μ_i and ϑ_i . The difference in square brackets reflects the difference between the future value function implied by the current choice.

Assuming the unobserved preference shock ε_{it} is logistic, when the discount factor (β) is zero our model collapses to simple logit model. With a positive discount factor, instead, the model needs to be solved at each iteration to compute the future value functions V_{it+1}^s and V_{it+1}^w . In our case these computations are relatively simple since the expected value of the value functions can be computed analytically, because of the distributional assumption we make. Given assumptions on the terminal value function for each final grade, the expressions in equation (1) can be computed by backward recursion..

As mentioned above, in the presence of unobserved heterogeneity, we assume that the constant μ_i (and possibly ϑ_i) is a discrete random variable, distributed independently of all characteristics $z_{it}, x_{it}^p, x_{it}^s$, and the $wage_{it}$.⁴ However, given the structure of our model and the fact that we use a single cross section, we have an important initial conditions problem because we do not observe the entire history of schooling for the children in the sample. That is, we cannot assume that the random variable μ_i (and ϑ_i) is independent of past school decisions. as reflected in the current level of schooling ed_{it} .

To solve this problem we specify a reduced form for educational attainment

⁴In practice dependence with the wage rate can be allowed for. However, the wage data is not rich enough to estimate a joint model of school participation and wages.

up to the current date. We assume that conditional on unobserved heterogeneity κ_i the level of schooling achieved up to now follows a Poisson distribution with mean $\exp(h'_i\zeta + \kappa_i)$ where h_i includes variables reflecting past schooling costs such as the availability of secondary schools in pre-experimental years. The probability of the stock of schooling and of attending school in this period are conditionally independent (given $z_{it}, x_{it}^p, x_{it}^s, h_i, wage_{it}$, and the unobservables $\mu_i, \vartheta_i, \kappa_i$). Hence we can write the probability of $ed_{it} = e$ and of child i attending school as

$$\begin{aligned} P(ed_{it} = e, Attend_{it} = 1 | z_{it}, x_{it}^p, x_{it}^s, h_i, wage_{it}, \mu_i, \vartheta_i, \kappa_i) = \\ P(Attend_{it} = 1 | z_{it}, x_{it}^p, x_{it}^s, wage_{it}, ed_{it}, \vartheta_i, \mu_i) \\ P(ed_{it} = e | z_{it}, x_{it}^p, x_{it}^s, h_i, wage, \kappa_i) \end{aligned}$$

The endogeneity of the stock of schooling is captured by the potential dependence of ϑ_i, μ_i and κ_i . Thus we assume that we can model this joint distribution as

$$F(\mu_i = m, \vartheta_i = s, \kappa_i = k) = p_{msk}$$

for $m \in M, s \in S$, and $k \in K$ where M, S and K are the set of points of support for μ, ϑ and κ . Hence for an individual with observable characteristics $z_{it}, x_{it}^p, x_{it}^s, h_i, wage$ the observed probability of attending and having reached a level of schooling e is

$$\begin{aligned} P(ed_{it} = e, Attend_{it} = 1 | z_{it}, x_{it}^p, x_{it}^s, h_i, wage_{it}) \\ \sum_{m \in M} \sum_{s \in S} \sum_{k \in K} p_{msk} \{ P(Attend_{it} = 1 | z_{it}, x_{it}^p, x_{it}^s, wage_{it}, ed_{it}, \vartheta_i = s, \mu_i = m) \\ P(ed_{it} = e | z_{it}, x_{it}^p, x_{it}^s, h_i, wage, \kappa_i = k) \} \end{aligned}$$

The number of points of support as well as the values that m, s and k can take and the probabilities at these points can be estimated as suggested in Heckman and Singer (1983).

3.3 Using the model for addressing issues in scaling up

The model we presented above is one of many possibilities. We chose it because it incorporates some key issues in educational choice, namely the trade-offs between costs and benefits, that are likely to characterize behaviour in a broad set of circumstances. It also incorporates the intertemporal trade-offs that are central to PROGRESA. The model provides a way of evaluating the impact of changing the parameters of the program, such as the amount offered and the way the amounts vary by age. Because of the forward looking nature of the model it also allows one to distinguish the impact of the program *vis a vis* a no program state as opposed to the randomized experiment which provides an impact of having the program compared to expecting to receive it in 18 months time.

However, all these advantages do not come for free. They come because of a number of assumptions that we have made which allow us to identify a rich behavioural model, combining the randomized experiment with further cross sectional variation. In particular, over and above the exogenous variation that is induced by the experimental design, we also use the structure of the grant across different age groups to identify how the impact changes with the amount of the grant and also how individuals react to the promise of a future payment. Obviously a richer evaluation framework that would have generated variation in the amounts and possibly in the age structure would have led to a model that would be identified using much fewer assumptions.

Given the model, we can immediately deal with the issue of the distribution of observable characteristics, by applying the model to a random sample drawn from the implementation area. This will ensure that when we aggregate the

impact we have applied the right weights. This process of course is possible also without a structural model. In terms of dealing with issues relating to the support, this is just a matter of comparing the support of the Xs in the evaluation data to the support in the implementation data. This can be done by using the device of the propensity score, which here is defined as the probability of being in the implementation data, given the Xs . If there is lack of common support, one then has to decide if extrapolation is to be used based on the parametric model.

The key difficulty though is the treatment of unobserved heterogeneity. We have shown in the estimation section how one can estimate the distribution of unobserved heterogeneity in a parametric setting. The distribution of unobservables can have a large impact on the impact of the policy. However, we know nothing of this distribution in the implementation area. In practice the only choice is to assume that the distribution does not vary across these areas; this of course is not satisfactory and can be a source of errors in the *ex ante* evaluation.

The model allows the examination of different structures of the program such as variation in the age rules and in the amounts. It also allows us to take into account differing aggregate conditions, so long as these vary within the evaluation areas. This is not the case for PROGRESA to any important degree, except perhaps as far as some variation in the wage is concerned. Finally note that the model can technically predict both the average treatment effect and the treatment on the treated. Of course it is important to note that the randomized experiment in the first place has allowed identification of ATE without further assumptions that would be required to extract ATE from treatment on the

treated.

4 The PROGRESA program and its evaluation sample

In this section, we start by describing the main features of the program PROGRESA. We then move on to discuss the evaluation sample and some of the results have been obtained from its analysis. We then move on to split the sample into two parts and describe some of the features in each of the two groups.

4.1 The program

In 1997, the Mexican government started a large program to reduce poverty in rural Mexico. PROGRESA, the programme introduced by the Zedillo administration was innovative in that introduced a number of incentives and conditions with which participant households had to comply to keep receiving the programme's benefits.

PROGRESA is the spanish acronym for 'Health, Nutrition and Education', that are the three main areas of the program. The health component consists of a number of initiatives aimed at improving information about vaccination, nutrition, contraception and hygiene and of a program of visits for children and women to health centres. Participation into the health component is a pre-condition for participating into the nutrition component that, in addition to a basic monetary subsidy received by all beneficiary households, gives some in kind transfers to households with very young infants and pregnant women. The largest component of the program is the education one. Beneficiary households with school age children receive grants conditional on school attendance.

The size of the grant increases with the grade and, for secondary education, is slightly higher for girls than for boys. In addition to the (bi) monthly payments, beneficiaries with children in school age receive a small annual grant for school supplies. Finally, all the transfers are received by the mother in the household. Before giving additional details on the education component of the program, we discuss how the program targets communities and households.

The Program first targeted the poorest communities in rural Mexico. Roughly speaking, the two criteria communities had to satisfy to qualify for the program were a certain degree of poverty (as measured by what is called an 'index of marginalization', basically the first principal component of a certain number of village level variables routinely collected by the government) and access to certain basic structures (schools and health centers). The reason for the second criterion is the conditional nature of the program: without some basic structures within a certain distance, beneficiary households could not comply with the basic conditions for retaining the beneficiary status (participation in vaccination and check-up visits for the health and nutrition components and school attendance for the education component)

Once a locality qualifies, individual households could qualify or not for the program, depending on a single indicator, once again the first principal component of a number of variables (such as income, house type, presence of running water, and so on). Eligibility was determined in two steps. First, a general census of the Progresá localities measured the variables needed to compute the indicator and each household was defined as 'poor' or 'not-poor' (where 'poor' is equivalent to eligibility). Subsequently, in March 1998, an additional survey was carried out and some households were added to the list of beneficiaries.

This second set of households are called 'densificados'.

For logistic and budgetary reasons, the program was phased in slowly but is currently very large. In 1998 it was started in less than 10,000 localities. However, at the end of 1999 it was implemented in more than 50,000 localities, covering about 2.6 million households, or 40% of all rural families. The program has now a budget of about 1 billion US\$ and is by far the largest welfare program in Mexico. It is the first program of its nature to survive a change of administration. Although its name recently changed to Oportunidades, the Fox administration decided not only to continue it, but to expand it to poor urban areas. The program has received a considerable amount of attention and publicity and similar programs are currently being implemented in Honduras, Nicaragua, Colombia, Turkey and Argentina. A detailed evaluation of various aspects of the program is contained in IFPRI (2000).

The program represents a substantial help for the beneficiaries. The nutritional component was 100 pesos per month (or 10 US dollars) in the second semester of 1998, which corresponds to 8% of the beneficiaries' income in the evaluation sample.

We report the details of the educational grant in Table 1. All the figures are in current pesos, and can be converted in US dollars at approximately an exchange rate of 10 pesos per dollar. As mentioned above, the grants are conditional to school enrolment and attendance of children, and can be cumulated within a household up to a maximum of 625 pesos (or 62.5 dollars) per month per household or 52% of the average beneficiary's income. The average grant per household in the sample we use was 348 pesos per month for households with children and 250 for all beneficiaries or 21% of the beneficiaries income.

To keep the grant, children have to attend at least 85% of classes. Upon not passing a grade, a child is still entitled to the grant for the same grade. However, if the child fails the grade again, it loses eligibility.

Table 1: PROGRESA bi-monthly monetary benefits				
Type of benefit	1998 1st sem.	1998 2nd sem.	1999 1st sem.	1999 2nd sem
Nutrition support	190	200	230	250
Primary school				
3	130	140	150	160
4	150	160	180	190
5	190	200	230	250
6	260	270	300	330
secondary school				
1st year				
boys	380	400	440	480
girls	400	410	470	500
2nd year				
boys	400	400	470	500
girls	440	470	520	560
3rd year				
boys	420	440	490	530
girls	480	510	570	610
maximum support	1,170	1,250	1,390	1,500

4.2 The evaluation sample

Before starting the program, the agency running it decided to start the collection of a large data set to evaluate its effectiveness. Among the beneficiaries localities, 506 were chosen randomly and included in the evaluation sample. The 1997 survey was supplemented, in March 1998, by a richer survey in these villages, located in 7 of the 31 Mexican states. All households in these villages were interviewed, for a total of roughly 25,000 households. Using the information of the 1997 survey and that in the March 1998 survey, each household can be classified as poor or non-poor, that is, each household can be identified as being entitled or not to the program.

One of the most interesting aspects of the evaluation sample is the fact that it contains a randomisation component. The agency running PROGRESA used the fact that, for logistic reasons, the program could not be started everywhere simultaneously, to allocate randomly the villages in the evaluation sample to 'treatment' and 'control' groups. In particular, in 320 randomly chosen villages of the evaluation sample were assigned to the communities where the program started early, that is in May 1998. The remaining 186 villages were assigned to the communities where the program started almost two years later (December 1999 rather than May 1998).

An extensive survey was carried out in the evaluation sample: after the initial data collection between the end of 1997 and the beginning of 1998, an additional 4 instruments were collected in November 1998, March 1999, November 1999 and April 2000. Within each village in the evaluation sample, the survey covers all the households and collects extensive information on consumption, income, transfers and a variety of other issues. For each household member, including each child, there is information about age, gender, education, current labour supply, earnings, school enrolment, and health status. The household survey is supplemented by a locality questionnaire that provides information on prices of various commodities, average agricultural wages (both for males and females) as well as institutions present in the village and distance of the village from the closest primary and secondary school (in kilometers and minutes).

The evaluation sample has been extensively studied. In addition to several reports produced by IFPRI, efficiently summarized by Skoufias (2001), several papers have looked at various outcomes, including Behrman and Todd (1999), Schulz (2000), Gertler (2000), Santiago (2001), to mention a few. The Behrman

and Todd (1999) paper is particularly important because it looks at differences between treatment and control localities in pre-program variables. By and large, the randomization was successful in that, with a few exceptions, there are no apparent differences between treatment and control villages.⁵

4.3 Two groups of states and their features

As we mentioned above, to illustrate the issues involved with scaling up, we divide the PROGRESA evaluation sample in two parts, conduct the 'evaluation' in one and use the other part of the sample to perform an 'ex-ante' evaluation of the program. Of course we can then compare the results of these evaluation with the results obtained in the 'ex-post' evaluation. The seven states included in the evaluation sample were divided into two groups. The first group of states, are the poorest four: Guerrero, Puebla, Veracruz and Hidalgo. The second group is formed by slightly more dynamic states: Michoacan, Queretaro, San Luis Potosi. In Table 2, we report sample means and standard deviations for a number of variables that are likely determinants of the outcome of interest, as well as pre-program outcomes. Among the first group of variables we include child level variables (completed years of schooling) household level variables (income, mother education, ethnicity), and community level variables (agricultural wage, presence of secondary school and distance from the closest secondary school). All these variables, which are likely to affect the effectiveness of the program, were included among the determinants of schooling choices in the theoretical model.

⁵Of course, one would expect 5% of rejections at the 5% level. Unfortunately, one of the few pre-program variables that turned out to be statistically different between treatment and control villages is school enrollment! This difference, whose origin is not clear, has motivated the use of diff-in-diff estimators. In the structural model we proxy for it with a 'treatment' dummy.

The differences between the sample localities in the two groups of states are remarkable. As expected the localities in the first group are considerably poorer than those in the second group. Every single indicator, from household income to agricultural wage, to the percentage of children belonging to beneficiary families points in that direction. Particularly remarkable is the percentage of children belonging to indigenous families, which is 35% in the first group and less than 4% in the second group. Pre-program enrollment is also quite different: in Group 1 75% of the boys aged 6 to 18 are in school, while in group 2 this percentage is 86%.

Of course, with so many dimensions, it is difficult to summarize the differences between the two groups of states in terms of these conditioning variables. For such a purpose, we estimate a simple probit model where the probability of a child being in the first or second groups is estimated as a function of the conditioning variables included in Table 2 (obviously excluding the outcome variables, such as pre-program schooling). Most of the variables inserted in the regression turned out to be statistically significant, indicating systematic differences in the distribution of these variables in the two groups of states.

For all observations in our sample we can then compute the propensity score (for belonging to the first group) which we use as a summary statistic for the difference in the distribution of the dependent variables between the two groups. In Figure 1, we plot the distribution of propensity scores in the two groups of states. Not only is the distribution quite different (as could be inferred from the fact that most variables were significant in the Probit regression), but we also see that for a sizeable proportion of observations, the support of propensity scores in the two samples does not overlap. 6% of the observations in group 1

have a value of the propensity score higher than the highest value in group 2 and 5% of the observations in group 2 have a propensity score lower than the lowest level observed in group 1.

For scaling up, this result implies that, even if unobserved heterogeneity does not constitute a problem, we can use the results of the evaluation in group 1 for an 'ex-ante' evaluation of the program in group 2 states only for 95% of the observations. Moreover, 28% of the observation in Group 1 have a propensity score between -2 and -3, while only 0.3% of the observations in Group 2 have a propensity score less than -2. These issues will impose important limits to the scaling up exercise, unless, of course, the independent variables we have been analyzing here turned out to be irrelevant for the effectiveness of the program.

Table 2: Differences between two groups of states				
	Group 1		Group 2	
	Mean	St.dev.	Mean	St. dev.
pre-program enrollment	0.748	0.419	0.856	0.370
completed years of educ.	5.03	2.74	5.05	2.58
Mother education				
less than compl. primary	0.377	0.485	0.295	0.456
less than compl. sec.	0.345	0.475	0.403	0.491
secondary or more	0.278	0.448	0.302	0.459
Household income	637	948	814	1056
Agricultural male wage	25.5	6.6	37.5	11.6
Distance from sec. (mins)	79.12	99.7	55.7	61.8
% of indigenous	0.349	0.477	0.038	0.192
% of program beneficiaries	0.859	0.348	0.801	0.348
Boys older than 5 and younger than 18.				
Group 1: Puebla, Guerrero, Veracruz, Hidalgo; N.obs: 16905				
Group 2: Queretaro, Michoacan, San Luis Potosi; N.obs: 3655				

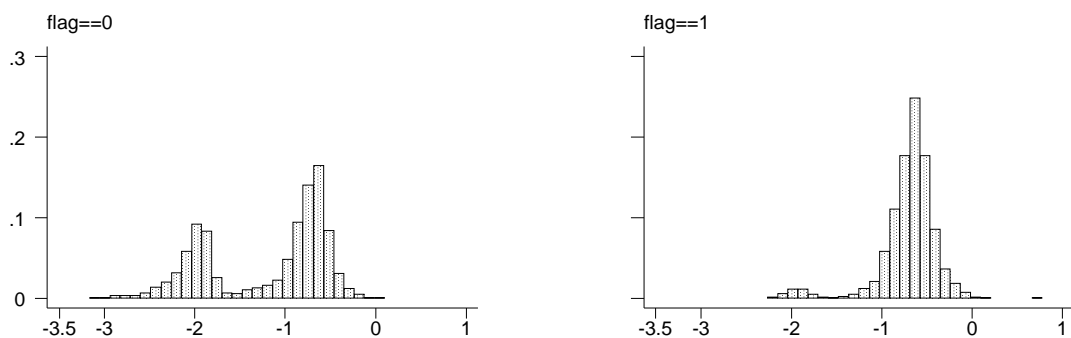


Figure 1: Propensity scores distribution in the two groups

4.4 The effect of PROGRESA in the two groups of states estimated by diff in diff.

Before presenting the results of the estimation of the structural model discussed above, we briefly discuss the effects of the program as estimated applying a diff in diff estimator in the two groups of states. It is now well documented that most of the effect of PROGRESA was on slightly older children (see Schulz, 2000). For that reason, here, and in what follows, we focus on boys ages 10 to 18. We use a diff. in diff. estimator that controls, however, for a few determinants of school choice, in order to improve efficiency.⁶ In particular, we control for the boy's age, mother education, distance of secondary schools and town average cost of secondary schools.

The results indicate a remarkable difference in the estimated effect of the program in the two groups of states. In Group 1, the effect is estimated at around 2.3% (with a standard error, which takes into account cluster effects at

⁶As the diff. in diff. estimator uses explicitly the randomization of the program between treatment and control localities, which is by construction uncorrelated with all independent variables, the only reason to 'control' for such variables is to improve efficiency.

the municipality level, of 0.012). In Group 2, the effect is estimated at 7.4%, (with a standard error of 0.017). The differences between the two effects is statistically significant (p-value 1.3%).

Two observations are in order. First, in the presence of anticipation effects in the control localities, the effect of the program is likely to be underestimated. This, following the discussion in AMS01, is our maintained assumption. We report the size of the effect of the program implied by our structural model and the assumption that the program is expected to be implemented in the control towns a year after the November 1998 measurement below. Second, to the best of our knowledge, this dramatic difference in the effect of PROGRESA in different states, had not been documented before. While the study of these differences is not the main goal of this paper, we cannot help stressing the difference in the effectiveness of the program in raising enrollment rates between the poorest states and the rest.

5 Estimation and simulation of a structural model

In this section, we first estimate the model described above in both groups of states. This set of results will allow to quantify the differences in the distribution of unobserved heterogeneity in the two groups of states. Any difference will prevent a straightforward use of the evaluation in on groups of states of 'scale up' to the other group. In the second part of the section, however, we ignore these differences, as well as the potential differences in the coefficients on the observable variables, and apply the model estimated in the first group of states to 'scale up' the program to the second group of states. This 'ex-ante' evaluation is then compared to the one that can be obtained from the evaluation sample

itself.

5.1 Estimation results

Table 3 reports the estimation results for the two groups of states. The specification of the model is identical for the two groups of states. In both cases, we performed a grid search on the discount factor and concluded that a value of around 0.9 maximizes the likelihood function. In both cases it was assumed that the program is perceived to reach the 'control' communities a year after the implementation in the treatment communities. This assumption is discussed in AMS01. Among the three scenarios tried in estimation (the program is never implemented in the control municipalities, the program is implemented after two years and after one year), is the one that yields the largest likelihood function.

The first two rows of the table report the estimated points of support for the discrete random variable μ_i : the intercepts of the value of going to school. All the coefficients are expressed as costs of going to school, so that the probability of going to school decreases with the size of this coefficients. In both sets of states we can identify two groups of children, the first of which is much more likely to go to school. However, in Group 2 both points of support are much lower, reflecting the overall larger enrollment of these children. At the bottom of the table we report the estimated distribution of the unobserved heterogeneity. Considering the marginal probabilities for the intercepts (reported in the third column), we see that in both groups of states, the group of children more likely to go to school accounts for over 80% of the sample. The differences in the level of these coefficients obviously constitute important problems for scaling up. Below we check how important these problems are in this particular example.

In rows three and four, we report the estimated coefficients on the wage. Here

the differences between the two states are also quite evident. Children in Group 2 seem to be much more sensitive to wages: both coefficients are positive and significant for group 2 (indicating that an increase in children wages decreases the probability of attending school), while for group 1, one of the two coefficients is not significantly different from zero. The marginal distribution is also different between the two groups. In Group 1, the two points account for 76 and 23% of the sample (notice that 76% are therefore insensitive to the wage), while in Group 2, the split is 46/54.

Going down the rows, we notice that all coefficients have the expected sign and that there are, once again, marked differences between the two groups. These differences constitute another potential problem for extrapolation and scaling up. Not only is the distribution of background variables different among the groups of states, as documented in the previous section, but the effect of these variables on the outcome of interest seems to be different. It is therefore difficult to use the results of one evaluation to extrapolate to a different sample.

Finally, notice the different (this time not too large), in the estimated coefficient of the grant. It should be noticed, however, that given the differences in all other coefficients, and the non linearity of the model, it is difficult to relate the size of this particular coefficient to marginal effect. It is to evaluate this type of effects that we now turn to simulations of the model.

Table 3: Estimates of structural model						
	Group 1			Group 2		
	Parameter est.	St.err.	Parameter est.	St. err.		
intercept p1	-26.12824	2.85159	-36.05473	3.90073		
intercept p2	-19.27025	2.31334	-24.36846	2.7138		
wage p1	-0.08644	0.06442	0.17190	0.08117		
wage p2	0.64341	0.12377	0.87642	0.13170		
grant	0.67240	0.11736	0.77147	0.20139		
distance from sec. school	0.09381	0.01325	0.19183	0.02608		
cost of sec. school	0.00793	0.00164	0.00463	0.00176		
age	3.24929	0.34110	3.91940	0.46625		
years of education	-2.55861	0.28772	-2.74321	0.42009		
father: primary	-0.19454	0.14100	-0.45956	0.22372		
father: incompl. sec	-0.42453	0.16610	-0.77902	0.27938		
father: secondary or more	-0.99106	0.37631	-1.39751	0.65107		
mother: primary	-0.15970	0.14841	-0.42837	0.22448		
mother: incompl. sec	-0.38623	0.17568	-0.70545	0.26758		
mother: secondary or more	-2.08134	0.53666	-1.08697	0.72461		
non-indigenous	-0.55368	0.14295	-0.95029	0.38256		
distr. of unobs. het.	0.6396	0.1855	0.8251	0.3463	0.5238	0.8701
(intercepts:rows	0.1276	0.0473	0.1749	0.1173	0.0126	0.1299
slopes: columns)	0.7672	0.2328	1.0000	0.4637	0.5363	1.0000

Boys older than 9 and younger than 18. The specification also includes state dummies, and dummies for 'beneficiaries' and for 'treatment localities'. The discount factor is 0.9.
Group 1: Puebla, Guerrero, Veracruz, Hidalgo; N.obs: 16905
Group 2: Queretaro, Michoacan, San Luis Potosi; N.obs: 3655

5.2 Simulation results

We now proceed to use the structural model to see how it performs in estimating the potential effect of the policy. The counterfactual simulations use the data from our chosen “implementation area” (Group 1) with a number of different combinations of parameters. For all counterfactual simulations we use the distribution of unobserved heterogeneity as estimated in the “evaluation area” (Group 2).

Before presenting the counterfactual simulations, however, we report the ef-

fect of the program as estimated by the model estimated in the “implementation area”. This is the kind of exercise that would not be possible to do ‘ex-ante’ and that should be replaced by ‘scaling up’.

The effect we report in the first column of Table 4 is the one predicted by the model as compared to the case where no policy is expected to be in operation. As mentioned above, the effect, estimated at 0.08, is larger than the one predicted by the randomised experiment, since the anticipation effect we estimate for the control group is netted out.

In the next column, we carry out another simulation experiment which in practice is not feasible. We use the estimated parameters from the evaluation area but use the state dummies estimated in the implementation areas - the latter would not normally be known in a genuine ‘scaling up’ exercise. This shows how close one can get, just by using the observable characteristics, if one knew the contribution of the unobserved area effects. The impact we get is about a quarter less than the baseline one. The difference is attributable to unobserved area effects.

In the third, fourth and fifth columns we present feasible predictions based on the coefficients from the evaluation area, the data from the implementation area and using in turn the state dummy from each of the three states in the evaluation areas. This should also be accompanied (not done yet) by an analysis that would show which state is likely to be most similar to the implementation area. From the results we see that none of the effects are particularly close to the one estimated using implementation area data. In one case the effect is less than half. (standard errors to follow)

Where does this leave us? At the moment we are still in the awkward

Table 4: Effect in Group 1 states.				
Predicted Model	Effect	Effect w.	Effect w.	Effect w.
Effect	with own State	Queretaro	SanLuisPotosi	Michoacan
	dummies	dummy	dummy	dummy
	0.080	0.061	0.039	0.058
				0.110

position of having to accept that we are not well enough equipped for a scaling up exercise that is reliable. We know that there is a very strong tradeoff here between the scope of the original evaluation and the modelling assumptions one is prepared to make. Clearly our model may be far from perfect. It makes strong behavioural assumptions. Moreover, one can argue that it does not perhaps have a rich enough specification. However, it does seem to fit reasonably well the data. What we do show in this paper is that knowledge of observed individual characteristics are unlikely to be sufficient in practice to extrapolate the effects of a policy.

6 Conclusions

In this paper we discuss issues to do with scaling up, i.e. using knowledge obtained from the evaluation of a policy in one area, for predicting the effectiveness of this policy in another. We start from the premise, which is empirically supported, that effects we are interested in vary substantially both in the observed characteristics and the unobserved characteristics dimension. Although one can achieve something close to ideal with an elaborate experimental design, it is unlikely that in the near future we will have the resources or even the will to carry out such complicated experimentation in all the required directions. We can go

some of the way if we have very large samples in the kind of experiments available now. This would at least allow us to correct carefully for the differences in the distribution of unobserved characteristics. But even there we are some way off having enough to produce reliable predictions, suitably reweighted. Thus the next best thing is to combine a structural model with the data we have. This allows us to fill in the gaps in a theoretically coherent way and offers a framework for redesigning policies. Inevitably this requires assumptions, but at least they are made in a coherent and transparent way. However, even there, in our first attempt, we show that our ability to predict the actual effects is limited, particularly by the lack on knowledge, of aggregate area effects. This points to the need to collect data at the area level and perhaps to design evaluations in such a way that more variation is induced at that level and not only at the individual level data.

7 References

Angrist, J. Bettinger, E., Bloom, E., King E. and M. Kremer (2002): “Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment,” *American Economic Review*, 92, pp. 1535-1558.

Attanasio, O.P., Meghir C. and A. Santiago (2001): “Education Choices in Mexico: using a structural model and a randomized experiment to evaluate a welfare program”, UCL Mimeo.

Behrman, J. and P. Todd (2000): “Randomness in the Experimental Samples of PROGRESA”, IFPRI, Mimeo

Gertler, P. J. (2000): “Final Report: The Impact of PROGRESA on Health”,

Mimeo, IFPRI.

Heckman, J., Lalonde, R. and J. Smith, (1999): “The evaluation of labor market policies”, Ashenfelter, O. and D. Card (eds) Handbook of Labor Economics, North Holland.

Schulz, T.P. (2000): “School subsidies for the poor: Evaluating a Mexican strategy for reducing povrety”, IFPRI, Mimeo.

Skoufias, E. (2001): “PROGRESA and its Impacts on the Human Capital and Welfare of Households in Rural Mexic: A synthesis of the Results of an Evaluation by IFPRI”, IFPRI, Mimeo.

Todd, P. and K. Wolpin, (2003): “Using Experimental Data to Validate a Dynamic Behavioural Model of Child Schooling and Fertility: Assessing the Impact of a School Subsidy Program in Mexico”, Mimeo, University of Pennsylvania.