Journal of Econometrics 145 (2008) 64–80

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Heterogeneous impacts in PROGRESA

Habiba Djebbari[a,b,c], Jeffrey Smith[d,e,b,*]

[a] Université Laval, Canada
[b] IZA, Germany
[c] CIRPÉE, Canada
[d] University of Michigan, United States
[e] NBER, United States

ARTICLE INFO

Article history:
Available online 19 June 2008

JEL classification:
C10
I38

Keywords:
Program evaluation
Heterogeneous treatment effects
Bounds
Random coefficient model
PROGRESA

ABSTRACT

We investigate impact heterogeneity using data from the experimental evaluation of the Mexican conditional cash transfer program PROGRESA. We build upon, and extend Heckman, Smith and Clements [Heckman, J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. Review of Economic Studies 64, 487–535] and recent studies of quantile treatment effects and random coefficient models. We find strong evidence of systematic (i.e. subgroup) variation in impacts in PROGRESA and modest evidence of heterogeneous impacts conditional on the systematic impacts. Our paper concludes with a discussion of the policy relevance of our findings and of heterogeneous impacts more generally.

Executive summary

In recent years, a salutary trend towards basing policy choices on compelling causal evidence has led to an increasing number of high quality experimental and non-experimental evaluations. Many, if not most of these evaluations focus exclusively on a single parameter: the mean impact of the program or policy being evaluated on those affected by it. While this parameter rightly occupies center stage in most contexts, it leaves unanswered a variety of additional evaluation questions related to the variation in the impacts across individuals and groups. The answers to these questions inform the design of program eligibility rules and of statistical treatment regimes designed to focus program resources on those most likely to benefit from them. They also illuminate the distributional consequences of a program, not least regarding the potential for programs to make some participants worse off even when helping them on average.

Building upon the recent literature, our paper investigates impact heterogeneity using data from the experimental evaluation of the Mexican conditional cash transfer program PROGRESA. As an outcome we focus on per capita consumption as measured in the November 1998 and 1999 post-random-assignment surveys. We estimate systematic (i.e. subgroup) variation in impacts as well as the idiosyncratic variation in impacts that remains after removing the systematic variation. As the experimental data do not pin down the distribution of idiosyncratic program impacts without further assumptions, we consider several sets of assumptions outlined in the literature.

We find strong evidence of systematic variation in impacts on consumption in PROGRESA. Our estimates imply that the impacts associated with extensions of eligibility to less poor villages and/or less poor individuals would be smaller than those for the population already served. They also suggest the potential value of more finely targeting the program via statistical treatment rules. Our evidence indicates substantial variation in impacts remains even after removing the systematic impact variation. This in turn suggests the value of additional data collection to try and transform this impact variation into systematic variation that can be quantified and used. Under all of the plausible assumptions we examine, we find that only a small fraction of individuals experience a fall in consumption as a result of treatment.

1. Introduction

This paper investigates heterogeneity in the treatment effects of the Mexican conditional cash transfer program PROGRESA using data from its well-known experimental evaluation. In addition to learning more about how PROGRESA affects its eligible population,

[*] Corresponding address: Department of Economics, University of Michigan, 238 Lorch Hall, 611 Tappan Street, Ann Arbor MI 48109-1220, United States. Tel.: +1 734 764 5359; fax: +1 734 764 2769.
E-mail address: econjeff@umich.edu (J. Smith).

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.05.012

we aim to provide a template for similar analyses of other programs by examining in one place a number of different methods advanced in the recent literature for learning about the distribution of impacts.

Studying heterogeneous treatment effects allows us to go beyond the simple mean impacts that dominate the literature. Looking at features of the distribution of impacts other than just the mean provides a deeper picture of how a program works and provides evidence for or against economic models that imply heterogeneous responses. Looking at subgroup variation in treatment effects allows an evaluation of the efficiency effects of actual and potential targeting rules, while estimation of the extent of variation in impacts not related to observables indicates the potential for data collection to improve program targeting. Knowledge about how program impacts vary, and how they relate to untreated outcomes, indicates the effects of the program on inequality within the eligible population, something that matters in a program, like PROGRESA, that provides transfers equal to a non-trivial fraction of untreated income among eligibles. More prosaically, politicians prefer programs that provide small benefits to many while injuring none, information not revealed by a narrow focus on mean impacts. In contrast, serious policy analysts will value the opportunity to select among programs, and program designs, based on social welfare functions that do not assume full transferability or risk neutrality.

We distinguish between systematic variation in mean impacts across subgroups and idiosyncratic variation in treatment effects conditional on any subgroup effects. Interactions between the treatment indicator and covariates suffice to estimate the former. In contrast, as pointed out by Heckman et al. (1997), getting a handle on the idiosyncratic variation poses a real methodological challenge. Experimental data (as well as observational data adjusted for selection bias) provide the marginal distributions of outcomes in the treated and untreated states, but not the joint distribution of these outcomes. Learning about the distribution of impacts and related parameters, such as the fraction with a positive impact or the impact variance requires knowledge of this joint distribution.

In this paper we adopt a variety of strategies to learn about the joint distribution of outcomes. We begin with the Fréchet–Höffding bounds, which exhaust the information about the joint distribution available from the marginal distributions. We investigate the distributions of impacts implied by the assumption of perfect positive dependence between the treated and untreated outcomes. This assumption allows the interpretation of quantile treatment effects as estimates of impacts for different untreated outcome levels; that is, it makes them impacts *at* quantiles rather than *on* quantiles. In addition, we investigate the quite different assumption made in the random coefficient model of independence between impacts and untreated outcomes. We implement multiple approaches to the random coefficient model and we describe and implement tests of both identifying assumptions. Unlike Heckman et al. (1997), much of our analysis considers systematic and idiosyncratic impact variation together, and we identify important subgroup effects in the PROGRESA data.[1]

We study PROGRESA for a number of reasons. First, the program and its experimental evaluation have had a profound influence on policy in Mexico and around the world. Similar programs now exist in a number of other countries in Latin America, inspired in large part by PROGRESA's design and positive evaluation results.

Second, the design, implementation and economics of the PROGRESA treatment suggest the likely importance of heterogeneous treatment effects. As described later on, the program provides payments conditional on school attendance that vary by the age and sex of the child. This heterogeneous treatment should lead to heterogeneous treatment effects. Implementation delays in the delivery of PROGRESA benefits in the treated villages should also generate heterogeneous treatment effects, particularly in the earlier of the two periods we examine. More generally, we would expect impacts on consumption (as well as children's school attendance and market work) to vary with household composition (including number and sex of adults and of children), initial wealth level, and tastes for education. See e.g. Skoufias and Parker (2001), Schultz (2004), Skoufias (2005) and Todd and Wolpin (2006) for more on the economics of PROGRESA.

Third, the experimental evaluation of PROGRESA provides us with high quality data on outcomes of substantive and policy interest and on conditioning variables for a large sample of individuals. Fourth and finally, because of random assignment we can proceed without worries about selection bias; this allows us to focus both our analysis and the exposition solely on the issue of impact heterogeneity.

Our empirical analysis fits into a small but rapidly growing literature that focuses on aspects of program impacts other than just the mean for active labor market and education programs. In addition to Heckman et al. (1997), who analyze the data from the U.S. Job Training Partnership Act experiment, this literature includes the analysis of quantile treatment effects (however interpreted) as in Bitler et al. (2006, unpublished manuscript, 2008), who look at programs for single parents in the US and Canada. Biddle et al. (unpublished manuscript) investigate heterogeneous effects of workplace injuries on earnings using a variant of the random coefficient model that we apply below to PROGRESA. In a developing country context, Dammert (unpublished manuscript) examines quantile treatment effects in a conditional cash transfer program in Nicaragua and, in work conducted independently of our own, Chavez-Martin del Campo (unpublished manuscript) does the same for PROGRESA.

The remainder of the paper has the following structure. Section 2 describes the PROGRESA program, its experimental evaluation and the data used in our analysis. Section 3 lays out our econometric framework and Section 4 discusses identification of the joint distribution of outcomes. Section 5 presents bounding estimates and the related perfect positive dependence case, while Section 6 investigates subgroup variation in mean impacts. Section 7 considers the random coefficient model. Section 8 discusses the policy implications of our findings and of heterogeneous treatment effects more generally while Section 9 concludes.

2. Institutions and data

2.1. The PROGRESA Program

The PROGRESA program targeted Mexico's rural poor; under the name Opportunidades, it now serves the urban poor as well. As documented in Skoufias et al. (2001), at the end of 1999 it covered 2.6 million families or about 40% of rural households and one-ninth of all Mexican families and boasted a budget of about $ 777 million or roughly 0.2% of Mexican GDP. As the program and its experimental evaluation (described below) have generated a large literature in economics, we focus here only on the features directly relevant to our analysis: the targeting/eligibility scheme and the PROGRESA treatment. Skoufias (2005) provides more details about PROGRESA.

[1] Though we explore a number of alternative approaches to pinning down the joint distribution of treated and untreated outcomes, we do not explore all of the interesting alternatives available in the literature. For example, we do not explore the use of copulas, as described in Nelsen (1999) and applied in Brendstrup and Paarsch (2007). Nor do we pursue the Bayesian approach outlined in Tobias (2006).

The targeting of PROGRESA involved two stages: (1) the selection of the localities where PROGRESA operates, (2) the selection of eligible households within the selected localities. PROGRESA operated only in remote localities that met two criteria. They had to have at least one primary school and a health center and they had to have a high score on a "village marginality index" based on pre-program village level data from 1997 on the illiteracy rate of household heads, access to basic infrastructure (running water, a drainage system, electricity), average housing characteristics (ratio of household members to rooms, dirt floor) and the importance of agricultural activities. Within selected villages, household eligibility depended on a "poverty index" based on 1997 data on household income and assets. A cutoff value of the poverty index defined poor families eligible for PROGRESA. As discussed in detail in Chapter 4 of Skoufias (2005), PROGRESA had a second round of eligibility determination in July 1999 that expanded the set of eligible households, mainly by increasing the number of elderly poor with no resident children. We define eligibility based on the later, more comprehensive definition throughout the paper. Doing so makes our results comparable over time in terms of population covered but also means that roughly one-third of the sample in the first time period we examine (see below) did not yet have the PROGRESA budget set.

The PROGRESA treatment has two components; taken together they imply quite different budget sets for eligible and ineligible households. The first component consists of grants given to families whose children regularly attend grades 3–9.[2] The grants vary by sex and grade, with higher payments to girls (who traditionally have lower rates of school attendance) and grade (to reflect the higher opportunity cost of market work for older children). A cap on the total grant binds for large households and affects their incentives at the margin.

The second component consists of payments, called the "food cash transfer" to eligible households that make regular visits to health centers and participate in health talks. Payment receipt requires one health center visit per year for adults, two to five visits per year for pregnant and breast-feeding women and two to seven visits per year for infants and children. Though named and marketed as a means of encouraging improved nutrition, households may spend the payments as they like. In addition to the cash payments, the program provides nutritional supplements to under-nourished children and infants and to pregnant and breast-feeding women.

PROGRESA distributes its grants bimonthly (at least in principle). Payments began in May 1998 in the treatment villages, though there may have been delays in some cases due to implementation difficulties. PROGRESA requires beneficiary households to withdraw from pre-existing social programs with similar goals and target populations, such as Niños de Solidaridad, DICONSA, LICONSA and INI, which continued to serve everyone in the control villages.[3] These programs comprise an important component of the control state, which thus consists of a different (and less generous) system of transfers rather than no transfers.

2.2. The experimental evaluation

The PROGRESA evaluation took advantage of the institutional necessity of staggered implementation of the program to randomly assign around 500 villages to either the first or the final group to receive the program. As a result, the control villages receive the program about two years after the treatment villages.[4] Village level random assignment has two key advantages: first, it avoids social unrest that might occur in small villages if some received generous transfers and others did not. Second, it avoids contamination of the experimental control group due to within-village spillovers. Unrest and spillovers certainly have the potential to compromise both internal and external validity.

Within treatment localities, eligible households could receive the PROGRESA transfers provided they met the conditions for doing so in terms of child school attendance or family attendance at health clinics. The PROGRESA treatment consists of the budget set (in terms of time and money) that embodies the conditional transfers. Thus, even treatment group members who do not behave in ways that imply receipt of transfers have received the PROGRESA treatment. Because the treatment consists of a budget set, PROGRESA does not raise any issues of treatment group dropout (sometimes called partial compliance) as in e.g. Heckman et al. (1998). Over the period October 1998 to November 1999 transfers to experimental treatment group households averaged about 197 pesos per household per month,[5] or about 20% of average consumption in eligible households.

2.3. Data

The data from the PROGRESA evaluation consists of repeated observations (panel data) for 24,000 households from 506 villages (320 in the treatment group and 186 in the control group) from seven states over five rounds of surveys (baseline surveys in October 1997 and March 1998 and follow-ups in November 1998, June 1999 and November 1999). In this paper we use data on consumption from two follow-up rounds, namely ENCEL (*Encuesta de Evaluacion de los Hogares*) of November 1998 and November 1999 as well as background data (collected prior to the program) from the October 1997 baseline survey.[6] We treat each round as a cross section and restrict our attention to eligible households. We also use data from a time allocation module administered only in June 1999. Appendix A (which includes Tables A.1 and A.2) provides additional details about our analysis sample; see Behrman and Todd (1999) for evidence on the quality of random assignment.

We use data at the household level, and focus primarily on per capita consumption as an outcome measure. We also examine children's participation in school, in market work and in domestic work using the time allocation data. Appendix A describes the construction of our consumption measure and the coding of the child time use indicators in greater detail.

Per capita consumption has the virtue that it neatly summarizes household well-being in the current period. Indeed, Deaton (1997) argues for consumption as preferable to income as a measure of well-being among poor households because, unlike agricultural output, it should remain relatively smooth over time. Our

[2] A child must attend class at least 85% of the time for the household to qualify for the educational grant, but failure to do so does not affect eligibility for other program benefits. A complex system of verification based on forms completed and signed by teachers and school directors governs the payment of the educational grants.

[3] *Niños de Solidaridad* provides educational grants, DICONSA maintains subsidized prices for basic food items, and LICONSA provides poor families with free tortillas and subsidizes the price of milk. INI targets indigenous people and provides lodging and food or educational grants to students. Figure 4.1 of Skoufias (2005) shows very low benefit receipt rates for these other programs among PROGRESA beneficiaries in the treatment villages.

[4] This design raises concerns about anticipation effects in the control villages. We do not know of any evidence on this question.

[5] These figures represent 1998 pesos; 197 pesos equal roughly US$ 20.

[6] We restrict ourselves to two follow-up rounds because no reliable consumption data (our main outcome of interest) is collected before November 1998. For the sake of comparison, we focus on the two rounds of data collected at the same time of the year, and thus do not present results for the June 1999 round.

Table 1
Descriptive statistics

	November 1998	November 1999
Monthly average per capita consumption	204.47	186.07
	(149.73)	(127.91)
Assigned to treatment group	0.6172	0.6115
	(0.4860)	(0.4874)
1997 household poverty score	−696.42	−695.17
	(117.58)	(116.78)
1997 village marginality index	0.4785	0.4787
	(0.7562)	(0.7369)
Household size	5.89	5.93
	(2.69)	(2.68)
Number of children less than 2 years old	0.51	0.51
	(0.76)	(0.76)
Number of children 3–5 years old	0.55	0.55
	(0.72)	(0.72)
Number of children 6–10 years old	1.11	1.13
	(1.13)	(1.13)
Number of boys 11–14 years old	0.34	0.35
	(0.59)	(0.59)
Number of girls 11–14 years old	0.33	0.33
	(0.58)	(0.58)
Number of boys 15–19 years old	0.34	0.34
	(0.61)	(0.62)
Number of girls 15–19 years old	0.32	0.33
	(0.59)	(0.60)
Number of men 20–34 years old	0.50	0.49
	(0.62)	(0.61)
Number of women 20–34 years old	0.55	0.55
	(0.59)	(0.59)
Number of men 35–54 years old	0.44	0.45
	(0.51)	(0.51)
Number of women 35–54 years old	0.45	0.45
	(0.51)	(0.51)
Number of men at least 55 years old	0.27	0.27
	(0.45)	(0.45)
Number of women at least 55 years old	0.26	0.26
	(0.47)	(0.47)
Male head of household	0.89	0.89
	(0.31)	(0.30)
Head is an agricultural worker	0.6244	0.63
	(0.4842)	(0.48)
Head's education (in years)	2.69	2.70
	(2.68)	(2.66)
Head is indigenous	0.38	0.38
	(0.48)	(0.48)
Age of head	46.47	46.54
	(15.89)	(15.77)
Number of observations	16,464	14,430

Standard deviations appear in parentheses.

focus on per capita consumption may hide within-household differences in program impacts. We also have stronger expectations of heterogeneous impacts for non-transferable outcomes such as schooling than for consumption. Both of these factors make our analysis of per capita consumption more conservative in the sense that they work against finding heterogeneous treatment effects. At the same time, consumption does not have a clear normative interpretation; the program goal of getting older children to remain in school, which should increase the future consumption of the (dynastic) household by increasing human capital, may lower consumption in the current period. This will occur in households where the PROGRESA transfers do not make up for losses in income from market work (that also do not get compensated by changes in adult behavior). In such households, expected discounted utility has presumably increased, but current consumption does not reflect this. As a result, current consumption represents an imperfect vehicle for a full-scale welfare analysis.

The top row of Table 1 presents the mean of per capita consumption in the two surveys while the top rows of each panel of Table 2 present means of the child time use indicators.[7] The remainder of Table 1 presents means of the background variables that we use as conditioning variables and for the analysis of systematic variation in average treatment effects.

3. Econometric framework

We use the notation of the potential outcomes framework for studying treatment effects.[8] In the case of a binary treatment, each individual "i" has two potential outcomes, denoted Y_{1i} for the

[7] We lack a good explanation for the fall in mean consumption from the 1998 survey to the 1999 survey. It occurs in both the treatment group and the control group (see the top row of Table 6 for the control group means) and so does not result from the program. It also does not result from selective attrition; we find the same pattern if we look only at households present in both the 1998 and 1999 samples (not shown).

[8] The statistics literature calls this the Rubin Causal Model and talks about causal effects rather than treatment effects. A debate regarding proper citations for the potential outcomes framework enlivens the footnotes of many papers; popular choices include Neyman (1923), Fisher (1935), Roy (1951), Quandt (1972) and Rubin (1974). We humbly add Frost (1920).

Table 2
Frechét–Höffding bounds on binary outcomes

	Boys			Girls		
	8–11	12–15	16–18	8–11	12–15	16–18
School						
Fraction attending school in the treatment group	0.61 (0.01)	0.46 (0.01)	0.12 (0.01)	0.61 (0.01)	0.44 (0.01)	0.09 (0.01)
Fraction attending school in the control group	0.59 (0.01)	0.44 (0.01)	0.11 (0.01)	0.61 (0.01)	0.38 (0.01)	0.07 (0.01)
Bounds on P_{SN}	[0.02, 0.40] (0.01), (0.01)	[0.014, 0.45] (0.01), (0.01)	[0.014, 0.12] (0.01), (0.01)	[0.007, 0.38] (0.01), (0.01)	[0.05, 0.44] (0.01), (0.01)	[0.02, 0.09] (0.01), (0.006)
Bounds on P_{NS}	[0.0002, 0.38] (0.001), (0.01)	[0.0007, 0.44] (0.002), (0.01)	[0.0005, 0.11] (0.002), (0.01)	[0.002, 0.38] (0.005), (0.01)	[0.00, 0.38] (0.00), (0.01)	[0.0001, 0.07] (0.001), (0.006)
Labor market						
Fraction working in the treatment group	0.05 (0.01)	0.18 (0.01)	0.43 (0.01)	0.04 (0.003)	0.06 (0.004)	0.09 (0.007)
Fraction working in the control group	0.06 (0.01)	0.21 (0.01)	0.42 (0.01)	0.03 (0.003)	0.06 (0.004)	0.09 (0.006)
Bounds on P_{WN}	[0.00008, 0.05] (0.0005), (0.003)	[0.00, 0.19] (0.00), (0.01)	[0.01, 0.43] (0.01), (0.01)	[0.007, 0.04] (0.005), (0.003)	[0.001, 0.06] (0.003), (0.004)	[0.002, 0.09] (0.005), (0.006)
Bounds on P_{NW}	[0.01, 0.06] (0.006), (0.005)	[0.02, 0.21] (0.01), (0.01)	[0.002, 0.42] (0.005), (0.01)	[0.0001, 0.03] (0.0007), (0.003)	[0.004, 0.06] (0.004), (0.005)	[0.005, 0.09] (0.006), (0.007)
Domestic activities						
Fraction working in the treatment group	0.34 (0.01)	0.40 (0.01)	0.33 (0.01)	0.36 (0.01)	0.52 (0.01)	0.51 (0.01)
Fraction working in the control group	0.35 (0.01)	0.40 (0.01)	0.34 (0.01)	0.37 (0.01)	0.56 (0.01)	0.55 (0.01)
Bounds on P_{DN}	[0.002, 0.34] (0.005), (0.01)	[0.004, 0.40] (0.007), (0.01)	[0.003, 0.33] (0.007), (0.01)	[0.001, 0.36] (0.003), (0.007)	[0, 0.43] (0), (0.01)	[0.00, 0.44] (0.00), (0.01)
Bounds on P_{ND}	[0.007, 0.35] (0.01), (0.01)	[0.005, 0.40] (0.008), (0.01)	[0.01, 0.34] (0.01), (0.01)	[0.01, 0.37] (0.01), (0.01)	[0.03, 0.47] (0.012), (0.01)	[0.04, 0.48] (0.01), (0.01)

S indicates school, W indicates work in the labor market and D indicates domestic activities.
$P_{.N}$ = probability of participating in the activity in the treated state, and not participating in the activity in the untreated state.
$P_{N.}$ = probability of not participating in the activity in the treated state, and participating in the activity in the untreated state.
Standard errors from 250 bootstrap replications appear in parentheses.

treated outcome and Y_{0i} for the untreated outcome. Each individual experiences only one of these two potential outcomes; which one depends on his or her treatment choice. This missing data problem, that we cannot observe both the treated and untreated outcomes for any single unit, constitutes the evaluation problem.

Defining a treatment indicator $D_i \in \{0, 1\}$ allows us to represent the observed outcome as $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$. In this notation, we may further define the impact of treatment on individual "i" as $\beta_{Di} = Y_{1i} - Y_{0i}$. The literature focuses on two parameters, the Average Treatment Effect (ATE) in the population, $E(Y_{1i} - Y_{0i}) = E(\beta_{Di})$ and the Average Treatment Effect on the Treated (ATET), $E(Y_{1i} - Y_{0i} \mid D_i = 1) = E(\beta_{Di} \mid D_i = 1)$. We assume throughout this paper that the assumptions required for data from a randomized experiment to identify the ATET parameter in the population hold in this context. This includes the assumptions of no equilibrium effects – called the Stable Unit Treatment Value Assumption (SUTVA) in the statistics literature – and no randomization bias (individuals do not act differently due to the presence of random assignment).[9]

Writing the potential outcomes in linear regression form we have

$$Y_i = \beta_0 + \beta_{Di} D_i + \varepsilon_i = \beta_0 + \beta_D D_i + [(\beta_{Di} - \beta_D) D_i + \varepsilon_i], \qquad (1)$$

where $\beta_0 = E(Y_0)$ and $\beta_D = E(\beta_{Di} \mid D_i = 1)$. The composite error term in square brackets in (1) includes the idiosyncratic component of the impact (for treated individuals) and the idiosyncratic component of the untreated outcome. Adding exogenous covariates to the model, and allowing both the untreated outcome and the impact to vary with those covariates, yields

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_X X_i + (\beta_D + \beta_{DX} X_i) D_i \\ & + [(\beta_{Di} - \beta_D - \beta_{DX} X_i) D_i + \varepsilon_i], \end{aligned} \qquad (2)$$

where we leave implicit the distinction between vectors and scalars. This formulation divides the impact for individual "i" into two components. We call the first $(\beta_D + \beta_{DX} X_i)$ the "systematic" component of the impact and the other$\tilde{\beta}_{Di} \equiv (\beta_{Di} - \beta_D - \beta_{DX} X_i)$ the "unobserved" or "idiosyncratic" component of the impact. A separate division into systematic and idiosyncractic components arises for each vector X.

Until very recently, much of the applied literature assumed either a strict common effect world, in which $\beta_{Di} = \beta_D$ for all "i"or a common effects within subgroups world, in which $\beta_{Di} = \beta_D + \beta_{DX} X_i$. In the first case, the regression model in Eq. (1) simplifies to

$$Y_i = \beta_0 + \beta_D D_i + \varepsilon_i. \qquad (1')$$

In the second case, the regression model in Eq. (2) simplifies to

$$Y_i = \beta_0 + \beta_X X_i + (\beta_D + \beta_{DX} X_i) D_i + \nu_i. \qquad (2')$$

Heterogeneity in the treatment and heterogeneity in individual circumstances relevant for treatment response render the common effect model implausible in many contexts. In the course of our analysis we test the common effect assumption in both of its forms.

4. Identification

In the usual non-experimental or observational context, the main econometric concern is correlation between the unobserved

[9] See e.g. Burtless (1995), Heckman and Smith (1995), Heckman et al. (1999) and Bloom (2005, 2006) for general methodological discussions of field experiments in labor economics and Duflo et al. (2006) for a similar discussion in the context of development economics. Angelucci and De Giorgi (forthcoming) analyze the equilibrium effects of PROGRESA.

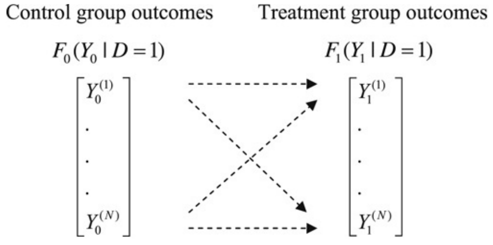

Fig. 1. Illustration of Frechét–Höffding bounding distributions.

component of the untreated outcome and the treatment indicator, conditional on included covariates. This is the classical selection bias problem that arises when individuals select into a treatment in part on the basis of the idiosyncratic component of their untreated outcome. In the PROGRESA context, random assignment suffices to eliminate this selection problem as, by construction, it makes the treatment indicator independent of all other variables, whether observed or unobserved. Moreover, random assignment makes the individual treatment effect unrelated to treatment status among the population at risk of random assignment. Thus, even though we can follow Heckman (1996) and think of randomly assigned treatment status as an instrument, we do not need to worry about the issues engendered by correlation between the instrument and the idiosyncratic component of the impact, as discussed in, e.g. Imbens and Angrist (1994), Angrist et al. (1996) and Heckman and Vytlacil (1998). Under the assumptions noted above, the data from the PROGRESA experiment suffice to identify the ATET.

More formally, experimental data identify the two marginal outcome distributions $F(Y_1 \mid D = 1)$ and $F(Y_0 \mid D = 1)$. These marginal distributions in turn identify the ATET as each marginal distribution pins down one of the relevant means. However, as highlighted by Heckman et al. (1997), experimental data do not suffice to identify the joint distribution of outcomes $F(Y_1, Y_0 \mid D = 1)$ or any parameters that depend on that joint distribution and not just the two marginals. Examples of such parameters include the percentiles of the distribution of impacts, the impact variance and the fraction with a positive impact.

The intuition here is that the marginals do not fully constrain the joint distribution. To see this, consider Fig. 1, which shows the treated observations from an experiment ranked in order on the left hand side, with $Y_i^{(j)}$ denoting the jth ordered observation and the untreated observations from the same experiment ranked in order on the right hand side, using similar notation, under the assumption of a common sample size N in the two groups. Different joint distributions imply a different mapping between the ranks of the treated distribution and the ranks of the untreated distribution. Heckman et al. (1997) formalize this intuition.

Fig. 1 illustrates two extreme cases that will turn out to have important properties in the next section. The first case, which we call Perfect Positive Dependence (PPD), and which the literature sometimes calls rank preservation or rank invariance, embodies a rank correlation of one. In this case, the ranks in the two marginal distributions correspond so that, for example, the median outcome in the treated distribution has as its counterfactual the median outcome in the untreated distribution, and so on for all other quantiles. The horizontal arrows embody this case. This case has some economic plausibility, especially for programs expected to have only modest effects on most participants. For example, if we take (very) seriously a "one factor" model, in which individuals have one skill and outcomes in both the treated and untreated states perfectly reflect that skill, then we obtain the PPD case. More generally, this case may represent a useful approximation in contexts where the one factor model basically holds but with some variation added on due to other factors including, but not limited to, the program under consideration.

The second case, which we call Perfect Negative Dependence (PND), embodies a rank correlation of minus one. In this case, for example, the qth percentile of the treated outcome distribution has as its counterfactual the $(100 - q$th) percentile of the untreated distribution. The crossing arrows embody this case. This case lacks surface plausibility as it posits that those who do the best when treated do the worst when not treated and vice versa. This represents an extremely strong form of comparative advantage, and also assumes that individuals cannot undo the effects of the treatment; for example, if an athlete becomes a clerk and finds out that she has little facility at clerking, it presumes that she cannot then go back to athletics. In our view, this assumption about the joint distribution has little relevance to the real world.

In addition to the two extreme joint distributions just considered, we also devote a section to random coefficient models, which identify the joint distribution of outcomes by assuming independence between the idiosyncratic components of the program impact and of the untreated outcome level. In the world of Eq. (1) this amounts to assuming independence of the two components of the composite error term, with the implication that the marginal density of Y_1 equals the product of the marginal density of the impacts and the marginal density of Y_0. We consider the empirical plausibility of this assumption in the PROGRESA context (and in general) in the relevant section.

5. Non-parametric bounds

5.1. Statistical theory

We begin our investigation of heterogeneous treatment effects in PROGRESA by using the data to estimate non-parametric bounds on the variance of the treatment effects. In particular, we estimate the bounds due to Fréchet (1951) and Höffding (1940), which we henceforth call the "FH" bounds. The FH bounds represent the solution to a classic question in statistics: what information do the marginal distributions of two variables contain regarding their joint distribution? The bounds do not rely on any assumptions beyond those required for experimental data to identify the ATET.

In the case of continuous random variables, the FH bounds are given by

$$\begin{aligned}&\max[F_1(Y_1 \mid D=1) + F_0(Y_0 \mid D=1) - 1, 0]\\ &\quad \leq F(Y_1, Y_0 \mid D=1)\\ &\quad \leq \min[F_1(Y_1 \mid D=1), F_0(Y_0 \mid D=1)]. \qquad (3)\end{aligned}$$

The FH upper-bound distribution corresponds to the PPD case discussed in the preceding section while the FH lower-bound distribution corresponds to the PND case. As discussed in Section 3(c) of Heckman et al. (1997), for binary outcomes, the two bounding distributions put as much or as little probability mass as possible in the diagonal elements of the corresponding 2 × 2 table. Rüschendorf (1981) prove tightness of the bounds and Mardia (1970) proves that the bounding distributions represent valid probability distributions.

Cambanis et al. (1976) show that if $k(Y_1, Y_0)$ is superadditive (or subadditive) then the extreme values of $E(k(Y_1, Y_0 \mid D = 1))$ occur at the two bounding distributions. As $\rho_{Y_1,Y_0} = \text{corr}\,(Y_1, Y_0)$ is superadditive, we can bound it using the FH bounding distributions. In addition, because we can write the variance of the impacts as a function of ρ_{Y_1,Y_0} and the two marginal distributions, as in

$$\text{var}(\beta_{Di}) = \text{var}(Y_1) + \text{var}(Y_0) - 2\rho_{Y_0,Y_1}\sqrt{\text{var}(Y_1)\text{var}(Y_0)}, \qquad (4)$$

we can also obtain bounds on $\text{var}(\beta_{Di})$ from the two bounding distributions. In particular, the FH upper bound distribution provides the lower bound on $\mathrm{var}(\beta_{Di})$ while the FH lower bound distribution provides the upper bound.

The impacts associated with the PPD correspond exactly to those considered in the recent literature on quantile treatment effects.[10] However, the two ways of looking at the same set of estimates imply very different interpretations. The quantile treatment effect framework makes no assumptions about the joint distribution and simply interprets the estimated difference in outcomes as informative about the nature of differences between the marginal distributions of outcomes in the treated and untreated states. Such differences do not necessarily represent the impact for any particular individual or group of individuals.

In contrast, the PPD assumption associated with the FH lower bound means that the quantile differences represent impacts for the individuals at the given quantile of the untreated outcome distribution. Put another way, without an assumption about the joint distribution, quantile treatment effects estimate impacts *on* quantiles of the outcome distribution; with the PPD assumption, the same procedure identifies impacts *at* quantiles of the outcome distribution and, thereby, the full distribution of impacts.

5.2. Implementation

We cannot obtain estimates of the parameters of interest that depend on the joint distribution of outcomes using the simple procedure implicit in Fig. 1 because our treatment and control groups differ in size. Instead, we collapse the marginal outcome distributions into percentiles. For the FH lower bound case, we then simply difference the corresponding percentiles to obtain percentile-specific impacts. In the FH upper bound case, we reorder the untreated outcome percentiles from lowest to highest and then take differences to obtain percentile-specific impacts. We obtain the outcome correlation directly from the percentiles of the marginal distributions and then use Eq. (4), along with estimates of $\mathrm{var}(Y_0)$ and $\mathrm{var}(Y_1)$ obtained from the individual data, to obtain the FH bounds on the impact variance. Taking square roots yields the bounds on the impact standard deviation. We obtain the fraction with a positive impact directly using the percentile-specific impacts associated with each FH bounding distribution.

We present estimated standard errors based on a standard bootstrap of the entire procedure. That is, we draw bootstrap samples of observations of the same size as our analysis sample, and then repeat the entire procedure just outlined. As discussed in detail in Appendix E(ii) of Heckman et al. (1997), the actual coverage probabilities of bootstrap confidence intervals, constructed based on either the bootstrap standard error or the percentiles of the distribution of bootstrap estimates, differ strongly from the nominal coverage probabilities when the population value of the variance equals zero. Our problem represents a special case of a more general problem with bootstrap confidence intervals and the related statistical tests of nulls corresponding to values at the boundary of the parameter space; see e.g. Andrews (2000) for a technical discussion.

In our context, even when the null holds, each percentile-specific impact equals exactly zero with probability zero. As a result, the 95 percent bootstrap confidence interval constructed using the percentiles of the bootstrap estimates contains zero with probability zero. Thus, a test based on inverting a confidence interval constructed using percentiles of the bootstrap distribution will essentially always reject the null, even when it holds in the population. In contrast, confidence intervals based on the standard deviation of the bootstrap standard deviation estimates do sometimes contain zero, but far too infrequently. As a result, to obtain a reliable *p*-value for a test of the null of a zero impact variance, we simulate the distribution of the impact standard deviation under the null by drawing both treated and control samples from the control group data; Appendix B provides the details.

5.3. Estimates

Table 2 presents bounds on the probabilities of being in school, of working in the paid labor market, and of performing domestic labor in the home, where we code each of these as a binary (does it/does not do it) outcome. The table has three panels, one for each of these three outcomes, and six columns, three for boys in each of three age groups and three for girls in the same three age groups. We expect the PROGRESA treatment to differentially affect boys and girls and children of different ages given the differential payments by grade level and sex and differences in the opportunity costs of schooling.

Within each panel, the first row presents the fraction of the treatment group doing the activity in that panel and the second row presents the corresponding fraction for the control group. The third row presents the FH bounds on the off-diagonal cells of the 2 × 2 table corresponding to doing the activity in the treated state and not doing it in the control state, while the fourth row presents the FH bounds for the other off-diagonal cell. For example, the FH bounds indicate that between 0.02 and 0.40 of boys ages 8–11 would attend school when treated but not attend school when not treated, while between 0.0002 and 0.38 would attend school when not treated but not attend school when treated. These numbers in turn imply that at least $0.21 (= 0.61 - 0.40)$ and at most $0.59 (= 0.61 - 0.02)$ of this group of boys would attend school regardless of treatment status. In contrast, for secondary school age boys, the lower bound on the fraction who would attend school in either state equals $0.00 = (0.12 - 0.12)$. A similar pattern holds for girls. This finding is consistent with Todd and Wolpin's (2006) argument that PROGRESA should target its financial incentives relatively more strongly at older children rather than "buying the base" of younger children, many of whom would attend school anyway.

The mean outcomes and the impacts they imply replicate patterns produced elsewhere: for boys, school attendance falls dramatically with age while market work rises equally dramatically. For girls, school attendance again falls dramatically with age, while work in the home rises. PROGRESA has a modest but not trivial mean effect on these outcomes in the intended directions. The loudest lesson from Table 2 is that the bounds are not very informative in this context. In particular, the lower bounds on the off-diagonal cells consistently just barely exceed zero, indicating that the marginal distributions provide little information beyond that implicit in a binary outcome.

Table 3 presents the FH bounds on various parameters related to the distribution of impacts for per capita consumption (hereafter just "consumption") in November 1998 (the first one or two columns) and November 1999 (the second one or two columns). The first row in the table shows average consumption in the control group, which equals 201.55 pesos in 1998 and 176.44 pesos in 1999. The second row shows the mean impact estimates, which equal 4.83 in 1998 and 15.79 in 1999. The next eight rows display the 5th, 25th, 50th, 75th and 95th percentiles of the estimated distribution of impacts followed by the fraction with a positive impact, the impact standard deviation and the outcome correlation.

[10] Lehman and Erich (1974) and Doksum (1974) introduced quantile treatment effects into the statistics literature. Koenker and Basset (1978) provide (to our knowledge) their first appearance in the economics literature. Recent contributions include Heckman et al. (1997), Koenker and Billias (2001), Abadie et al. (2002) and Bitler et al. (2006).

Table 3
Parameter estimates for the Fréchet–Höffding bounding distributions

	Nov. 98 per capita consumption		Nov. 99 per capita consumption	
Average untreated outcome	201.55		176.44	
(standard deviation)	(144.17)		(125.82)	
Average impact	4.83		15.79	
(standard error)	(2.39)		(2.17)	
Statistic (Bootstrap SE)	Perfect positive dependence	Perfect negative dependence	Perfect positive dependence	Perfect negative dependence
5th percentile	2.27	−413.64	10.09	−342.79
	(1.29)	(7.77)	(1.23)	(6.95)
25th percentile	4.67	−131.02	15.98	−98.76
	(1.35)	(2.48)	(1.27)	(2.16)
50th percentile	5.46	3.56	16.38	14.47
	(1.90)	(2.00)	(1.84)	(1.86)
75th percentile	1.83	131.07	17.11	126.01
	(3.79)	(1.93)	(3.03)	(1.58)
95th percentile	−7.07	377.79	22.79	343.97
	(8.88)	(3.62)	(10.92)	(4.35)
Fraction positive	0.84	0.51	1.00	0.54
	(0.001)	(0.284)	(0.001)	(0.286)
Impact standard deviation	12.21	260.90	3.01	224.02
	(4.28)	(1.02)	(3.41)	(0.86)
Outcome correlation	0.9973	−0.6031	0.9997	−0.6156
	(0.003)	(0.004)	(0.003)	(0.005)
Cutoff value for $p = 0.50$	5.58		5.62	
Cutoff value for $p = 0.40$	6.38		6.46	
Cutoff value for $p = 0.30$	7.35		7.51	
Cutoff value for $p = 0.20$	8.67		8.86	
Cutoff value for $p = 0.10$	10.77		10.93	
Cutoff value for $p = 0.05$	12.67		12.89	
Cutoff value for $p = 0.01$	16.46		16.59	

Cut-off values are from the distribution of the impact standard deviation statistic under the null of no variance in impacts. Details on the bootstrap and Monte Carlo simulations are given in Appendix B.

Ignoring sampling variation, the FH bounds indicate that the impact standard deviation lies between 12.21 and 260.90 pesos in 1998 and between 3.01 and 224.02 pesos in 1999. The FH bounds show that under the PPD assumption, about 84 percent of households in 1998 and 100 percent in 1999 had increased consumption due to PROGRESA, compared to only 51 and 54 percent under PND. Under PPD, in 1999 the impact increases with the percentile of the outcome distribution, indicating that those with the highest consumption levels without PROGRESA had the largest impacts.

The final seven rows of Table 3 present selected quantiles of the simulated distribution of the impact standard deviation under the null of a zero impact standard deviation. Comparing the lower bound on the impact standard deviation to these values, we find that we can reject this null at the 10% level in the 1998 data. This finding represents very valuable information, which we obtain without imposing any assumptions on the data beyond those generally invoked in experimental analyses. In contrast, we cannot reject the null for any reasonable significance level in the 1999 data. It does not follow, of course, that the data do not embody heterogeneous impacts; rather, it means that the marginal distributions alone do not imply that they do. The difference between the two years likely results in part from incomplete program implementation in 1998.

5.4. Testing the PPD assumption

We argued earlier in the paper that, while the PND case has little empirical plausibility, the PPD does, at least for treatments expected to affect outcomes only modestly. In this section, we test an implication of the PPD assumption using a test developed in Bitler et al. (2005).[11] Their test formalizes the intuition that if quantiles of the untreated outcome represent the counterfactual outcomes for the same quantiles of the treated outcome distribution, then exogenous covariates should have the same distributions in each quantile of the two outcome distributions. If we reject the null of equal distributions, this represents strong evidence against the PPD assumption; in contrast, if we fail to reject the null, it represents only weak evidence in favor of that assumption, as equal covariate distributions could also hold for many other joint distributions of outcomes.

We face two main choices in implementing the test. First, how wide should we make the regions over which we test equality of the covariate distributions? We could do a separate test for every percentile, or we could do a smaller number of tests using larger intervals such as quartiles. Doing many tests means low power (due to small sample sizes) for each test and also raises important multiple comparisons issues. On the other hand, relying on a small number of tests runs the risk of missing departures from equality of distributions within small sub-intervals of the outcome distributions. We opt to follow Bitler et al. (2005) and test equality of means by quartile; as it turns out, this suffices to make the point in our data.

The results of our tests appear in Table 4. Each of the four columns of results corresponds to one quartile of the marginal outcome distributions from November 1999. Each row corresponds to a particular exogenous background variable. The top value in each entry represents the mean difference in the row variable between the treatment and control observations in the column quartile. A 90% bootstrap confidence interval appears below each estimate. The bootstrapping takes into account the variance component that results from the initial assignment to quartiles.

The variables in the table consist of the poverty score and village marginality index, characteristics of the household head, and measures of the age and gender composition of the children and adults in the household for a total of 21 variables and, thus, $84(= 4 \times 21)$ tests. Under the assumption of independence of the

[11] We cite the working paper version here because the test does not appear in the published version, Bitler et al. (2008).

Table 4
Treatment-control differences at quantiles of the outcome distribution

	0–25th percentile	25–50th percentile	50–75th percentile	75–100th percentile
Household poverty score	−16.038* [−5.943; 6.510]	11.311* [−6.492; 5.835]	18.749* [−6.077; 5.967]	28.176* [−6.842; 6.543]
Village marginality index	−0.262* [−0.040; 0.043]	−0.030 [−0.039; 0.042]	0.027 [−0.043; 0.038]	0.017 [−0.040; 0.037]
Head is an agricultural worker	−0.011 [−0.025; 0.026]	−0.007 [−0.028; 0.026]	0.015 [−0.028; 0.026]	−0.008 [−0.029; 0.026]
Male head of household	−0.003 [−0.015; 0.015]	−0.009 [−0.015; 0.016]	0.019* [−0.017; 0.016]	0.008 [−0.022; 0.020]
Head’s education (in years)	0.282* [−0.137; 0.143]	−0.023 [−0.156; 0.150]	0.127 [−0.159; 0.158]	−0.074 [−0.155; 0.156]
Indigenous head	−0.099* [−0.029; 0.029]	0.024 [−0.028; 0.030]	0.044* [−0.027; 0.027]	0.024 [−0.026; 0.025]
Household size	−0.092 [−0.145; 0.142]	−0.025 [−0.149; 0.142]	0.190* [−0.137; 0.138]	0.111 [−0.133; 0.139]
Head’s age	0.207 [−0.803; 0.713]	−0.935* [−0.855; 0.852]	−1.197* [−0.904; 0.890]	−1.318* [−1.015; 1.015]
Children less than 2 years old	0.040 [−0.053; 0.050]	0.045* [−0.047; 0.043]	0.049* [−0.043; 0.039]	0.014 [−0.030; 0.031]
Children 3–5 years old	−0.074* [−0.049; 0.045]	0.050* [−0.040; 0.044]	0.029 [−0.038; 0.040]	0.030 [−0.034; 0.034]
Children 6–10 years old	−0.046 [−0.067; 0.066]	0.065* [−0.063; 0.061]	0.063* [−0.062; 0.061]	0.081* [−0.051; 0.052]
Boys 11–14 years old	−0.002 [−0.036; 0.038]	−0.020 [−0.034; 0.035]	0.013 [−0.031; 0.031]	0.034* [−0.027; 0.027]
Girls 11–14 years old	−0.053* [−0.038; 0.036]	−0.007 [−0.037; 0.035]	0.007 [−0.036; 0.033]	0.021 [−0.028; 0.025]
Boys 15–19 years old	0.044* [−0.041; 0.037]	−0.005 [−0.037; 0.038]	0.019 [−0.036; 0.033]	0.002 [−0.027; 0.027]
Girls 15–19 years old	−0.022 [−0.036; 0.035]	−0.035 [−0.038; 0.037]	0.005 [−0.033; 0.033]	−0.016 [−0.027; 0.029]
Men 20–34 years old	0.026 [−0.037; 0.037]	0.008 [−0.036; 0.034]	0.022 [−0.036; 0.032]	0.019 [−0.029; 0.030]
Women 20–34 years old	0.006 [−0.038; 0.033]	−0.017 [−0.037; 0.036]	−0.019 [−0.034; 0.031]	0.011 [−0.030; 0.030]
Men 35–54 years old	−0.053* [−0.030; 0.030]	−0.020 [−0.029; 0.030]	0.025 [−0.030; 0.029]	0.010 [−0.028; 0.026]
Women 35–54 years old	−0.013 [−0.033; 0.030]	−0.015 [−0.032; 0.032]	0.021 [−0.028; 0.030]	−0.012 [−0.027; 0.027]
Men more than 55 years old	0.027* [−0.024; 0.025]	−0.017 [−0.025; 0.026]	−0.026* [−0.025; 0.026]	−0.028* [−0.025; 0.027]
Women more than 55 years old	0.005 [−0.026; 0.028]	−0.016 [−0.026; 0.026]	−0.017 [−0.026; 0.026]	−0.040* [−0.028; 0.030]

Bootstrap CIs in square brackets.
* Denotes significance at the 10% level.

different tests, we would expect about eight or nine rejections (at the ten percent level) in the table under the null of no differences in the distributions of covariates between treatment and control observations in the four quartiles. In fact, we obtain 28 rejections, though of course the tests are not independent. The strongest departures from equality occur for the poverty score and for the variables related to the age and gender composition of the children in the household. Our estimates suggest that households with high initial poverty scores migrated up the outcome distribution when treated, as did households with a larger family size, a larger number of infants and primary school age children and a smaller number of older adults.

In summary, Table 4 provides some evidence inconsistent with PPD in 1999. This result has two consequences for the analysis: (1) the lack of heterogeneity in program impacts based on the Fréchet–Hoeffding lower bound cannot be interpreted as lack of heterogeneity in actual impacts; and (2) we cannot interpret the quantile treatment effects in Table 3 as impacts at particular quantiles of the untreated outcome distribution. These conclusions come with one further caveat: as documented in Behrman and Todd (1999), due to village level rather than individual randomization, some differences in mean characteristics remain between the treatment and control groups. These lingering differences could contribute to our rejection of the PPD in the 1999 data.

6. Systematic impact heterogeneity

Most program evaluations include impact estimates for specific subgroups, such as men and women and individuals with more or less education. At the same time, some readers criticized the Heckman et al. (1997) study (correctly) for failing to distinguish between systematic subgroup variation in treatment effects and idiosyncratic variation in the effect of treatment that remains after removing subgroup variation. If subgroup variation in impacts represents most of the overall variation, then the methods for examining idiosyncratic impact heterogeneity applied in this paper have little practical value.

In this section, we respond to these concerns by first examining the importance of subgroup impacts in PROGRESA.[12] These subgroup impacts have important implications for program design and data collection, as discussed in detail in the policy section below. We then repeat the bounding exercise on the idiosyncratic component of the variation in treatment effects that remains after removing the systematic component to see if, in the PROGRESA

[12] Skoufias (2005) and Behrman et al. (2005) also provide some (more limited) subgroup impact estimates for PROGRESA. See Horwitz et al. (1996) for an example of the large literature in epidemiology and related fields on the value and interpretation of subgroup impacts.

Table 5
Systematic variation in impacts

Point estimate (Standard error)	Nov. 98 per capita consumption	Nov. 99 per capita consumption
Treatment	76.742[b]	72.361[b]
	(−26.225)	(24.772)
Treatment ×Poverty score	0.062[b]	0.078[c]
	(−0.023)	(0.021)
Treatment ×Village marginality index	91.780[c]	34.394[c]
	(−10.581)	(9.761)
Treatment ×Village marginality index ×Poverty score	0.126[c]	0.048[b]
	(−0.016)	(0.015)
Treatment ×Children less than 2 years old	2.181	−5.675
	(−3.403)	(3.160)
Treatment ×Children 3–5 years old	0.831	0.124
	(−2.909)	(2.736)
Treatment ×Children 6–10 years old	5.949	−0.853
	(−3.183)	(2.974)
Treatment×Boys 11–14 years old	8.715[a]	1.336
	(−4.066)	(3.863)
Treatment ×Girls 11–14 years old	6.73	4.139
	(−4.077)	(3.780)
Treatment ×Boys 15–19 years old	1.707	−5.309
	(−3.787)	(3.545)
Treatment ×Girls 15–19 years old	8.620[a]	1.404
	(−3.749)	(3.509)
Treatment ×Men 20–34 years old	14.169[c]	1.398
	(−4.059)	(3.888)
Treatment ×Women 20–34 years old	8.143	1.213
	(−4.358)	(3.971)
Treatment ×Men 35–54 years old	17.061[b]	−0.681
	(−5.778)	(5.195)
Treatment ×Men 35–54 years old	2.835	−2.151
	(−5.273)	(5.173)
Treatment ×Men more than 55 years old	21.381[b]	−7.142
	(−8.046)	(7.428)
Treatment ×Women more than 55 years old	−13.012[a]	−1.165
	(−5.881)	(5.678)
Treatment ×Log(Family size)	−47.394[a]	−7.392
	(−18.857)	(17.549)
Treatment ×Male head of household	−3.526	11.583
	(−8.695)	(8.467)
Treatment ×Indigenous head of household	9.150[a]	8.119*
	(−4.253)	(3.986)
Treatment ×Head's age	0.007	0.068
	(−0.258)	(0.247)
Treatment ×Head's education (in years)	1.151	−0.700
	(−0.874)	(0.773)
Treatment ×Head is an agricultural worker	4.048	−0.954
	(−4.21)	(4.058)
R-squared	0.336	0.328
F-statistic for the null that all interactions = 0	9.27	4.66
(*p*-value)	(0.001)	(0.001)
F-statistic for the null that all interactions except those involving the poverty score and the village marginality index = 0	2.34	1.94
(*p*-value)	(0.001)	(0.008)
N	16,464	14,430

Treatment = 1 for treatment group observations, 0 otherwise. The specification also includes direct effects for all of the covariates that we interact with the treatment indicator.

[a] Statistical significance at the ten percent level.

[b] Statistical significance at the five percent level.

[c] Statistical significance at the one percent level.

case, we need to worry about idiosyncratic heterogeneity in program impacts. More formally, in terms of our notation we examine whether we can statistically distinguish the lower bound on $\mathrm{var}(\tilde{\beta}_{Di})$ from zero.

6.1. Systematic variation in impacts by subgroups

Table 5 presents our estimated subgroup impacts. Columns 1 and 2 of Table 5 present estimates for consumption in November of 1998 and 1999, respectively. In contrast to the common practice in government evaluation reports of examining the subgroups one at a time, we estimate Eq. (2) including interactions between the treatment indicator and the household poverty score and village marginality index, variables capturing the number and sex of children in the household, (log) household size and characteristics of the household head such as sex, age, education and working in agriculture. The final three rows present *p*-values from F-tests of two null hypotheses: that all of the coefficients on the interaction terms equal zero and that all of the coefficients on interaction terms other than those for the poverty score and the village marginality score and their interaction equal zero. Although we did not choose our subgroups in advance, they all represent obvious choices given the design of the PROGRESA treatment, the program's eligibility rules, the evaluation design, and the population at issue.

We highlight three findings from Table 5. First, in both 1998 and 1999 the poverty score and village marginality index, and their

Table 6
Estimates including systematic impact heterogeneity

	Nov. 98 per capita consumption	Nov. 99 per capita consumption
Average untreated outcome	201.55	176.44
	(144.17)	(125.82)
Average systematic impact	7.11	16.79
	(1.86)	(1.60)
Systematic impact standard deviation	21.73	14.27
	(2.12)	(1.67)
Fraction positive given systematic variation in impacts	0.6075	0.8998
	(0.0436)	(0.0368)
Lower-bound on the idiosyncratic impact standard deviation	11.47	4.74
	(5.26)	(2.75)
Upper-bound on the idiosyncratic impact standard deviation	222.62	190.66
	(4.17)	(3.25)
Cutoff value for $p = 0.50$	4.68	4.81
Cutoff value for $p = 0.40$	5.36	5.58
Cutoff value for $p = 0.30$	6.11	6.46
Cutoff value for $p = 0.20$	7.20	7.65
Cutoff value for $p = 0.10$	8.85	9.46
Cutoff value for $p = 0.05$	10.35	11.08
Cutoff value for $p = 0.01$	13.34	14.14

Cut-off values for the distribution of the idiosyncratic impact standard deviation statistic under the null of no idiosyncratic variance in impacts are constructed as described in the text and in Appendix B.

product, strongly predict differences in the mean treatment effect on consumption. In both cases, all three variables have positive coefficient estimates, indicating larger impacts on consumption for poorer households, for households in more marginal villages and particularly for poorer households in more marginal villages. This pattern has important implications for program targeting, as we discuss later on. Second, we find much more subgroup heterogeneity in treatment effects in the 1998 data than in the 1999 data; we again interpret this, at least in part, as due to the effects of partial program implementation. Finally, we can strongly reject the null of zero coefficients on all of the interaction terms and, perhaps more surprisingly in 1999, all of the interactions other than those involving the household poverty score and village marginality index.

6.2. Bounds conditional on subgroup impacts: econometrics and implementation

We now present estimates of the FH bounds on the impact variance after removing subgroup variation in mean impacts. We obtain our estimates of the FH bounds on $\mathrm{var}(\tilde{\beta}_{Di})$ using the residuals from estimating Eq. (2) in the preceding section. The use of residuals from (2) rather than the observed outcomes constitutes the primary difference relative to our procedure in the preceding section. We also include main effects in the variables we interact with the treatment indicator. Doing so yields a more powerful test via a reduction in the residual variance.

We approximate the distribution of the FH bounds under the null using simulations. In each simulation, we draw a sample equal in size to the original data set from the control group with replacement. We then estimate (2), obtain the residuals, randomly divide the residuals into treatment and control groups in proportion to the original data, convert the data into percentiles, and then use the percentiles to estimate the FH bounds as described above. We test the null of a zero idiosyncratic impact variance by comparing the estimated FH lower bound on the variance from the real data to the distribution of estimates from 100,000 simulations.

Our procedure differs from the related test of the same null hypothesis (but with different data) presented in Bitler et al. (unpublished manuscript). Their test begins, like ours, with estimation of (2). They then use their control group data combined with estimated subgroup effects to estimate the outcome distribution under the null of no idiosyncratic impact heterogeneity. Their test statistic consists of the maximum difference in percentiles between the treated and untreated outcome distributions. They obtain the distribution of their test statistic using sub-sampling methods, as discussed in Chernozhukov and Fernandez-Val (2005). Analytic and/or Monte Carlo comparison of the properties of these two testing schemes (and refinements thereof) represents a useful avenue for future research.

Bitler et al. (unpublished manuscript) extend their analysis to take account of point masses at zero in the earnings outcome variables they examine; without such an extension, rejection of the null follows almost automatically from the fact of point masses at zero in both the treated and untreated states combined with non-zero mean impacts. As our consumption data includes no zeros (or other substantial point masses) we do not require this complication.

6.3. Bounds conditional on subgroup impacts: results

Table 6 presents estimates of the FH bounds on the idiosyncratic impact variance after removing the systematic variation in mean impacts. The two columns present estimates for November of 1998 and 1999, respectively. The first row presents the mean untreated outcome, the second the mean systematic impact, the third the standard deviation of the systematic impacts and the fourth the percentage with a positive impact based only on the systematic impacts. The mean systematic impact rises from about $0.035(= 7.11/201.55)$ of the untreated outcome in 1998 to about $0.095(= 16.79/176.44)$ of the untreated outcome in 1999. Based on the systematic impacts alone, 60.75% experience an increase in consumption in 1998 and 89.98% do so in 1999.

Rows 5 and 6 of Table 6 give the estimated FH bounds on the idiosyncratic impact standard deviation, while the remaining rows give various percentiles of the distribution of estimates of the bounds under the null of a zero idiosyncratic variance obtained from the simulations. We find four patterns worth noting. First, removing the systematic variation in impacts leads to surprisingly little change in the estimated FH lower bound on the idiosyncratic impact standard deviation, which changes from 12.21 to 11.47 for 1998 and from 3.01 to 4.74 for 1999. Second, at the FH lower bound, the ratio of the idiosyncratic and systematic impact standard deviations equals $0.068(= \ 1.47/21.73)$ in 1998 and $0.332(= 4.74/14.27)$ in 1999. Thus, even in that extreme case, important idiosyncratic variance remains. Third, as in Table 3, we cannot reject the null of a zero idiosyncratic impact variance in 1999 at conventional levels. Finally, as before, we attribute the reduction in the impact variance, both systematic and idiosyncratic, from 1998 to 1999, as well as the increase in the mean impact, to partial implementation in 1998.

7. Random coefficient models

7.1. Motivation

The broader social science literature commonly assumes independence of the impact and the untreated outcome level. This assumption underlies the random coefficient models sometimes used in applied work in economics. The random effects panel data model represents a special case of the random coefficient model, in which only the intercept has a random coefficient. Outside economics, this assumption plays an important role in the multi-level models widely used in educational statistics and elsewhere.

How plausible is the assumption of independence of impacts and untreated outcome levels in the PROGRESA context? First of all, we have random assignment of a budget constraint based on fixed eligibility criteria. Thus, conditional on eligibility, selection into treatment based on expected impacts plays no role here. On the other hand, while all treated units face the PROGRESA budget constraint, not all of them choose to receive the subsidy in all periods or for all of their children. This choice should correlate with both the market wages and household productivity of children and thus with levels of schooling and consumption in the absence of PROGRESA (and with variables correlated with those levels). In some contexts, arguments that individuals have little information on which to base their decisions and so participate more or less at random conditional on eligibility might justify the random coefficient model, but such arguments lack plausibility here in regard to take-up of the subsidy as individuals will already have a sense of their children's payoff from schooling and their opportunity cost in domestic production or market work.

Thus, overall, we view the a priori case for the random coefficient model in the PROGRESA context as mixed at best, though it becomes more plausible once we remove the systematic variation in impacts. Thus, we analyze it in part because we think it has some plausibility in the latter case and in part because of our desire to provide a template for future work in other contexts, some of which may provide greater a priori support for the random coefficient model.

7.2. Econometrics

The literature provides a variety of estimators for the random coefficient model; they differ mainly in the amount of structure imposed on the impact distribution. We proceed in three steps. In the first step, we impose only the independence assumption. This allows us to estimate the impact variance and perform tests on the impact variance, but does not suffice to identify the full distribution of impacts. Under this assumption, in model (1) we can estimate the impact variance as

$$\mathrm{var}(\beta_{Di}) = \mathrm{var}(Y_1) - \mathrm{var}(Y_0). \tag{5}$$

Nothing in the data constrains the sign of this difference; which therefore provides a crude test of the random coefficient model. Assuming a normal distribution for the impacts, as in the Hildreth and Houck (1968) estimator but without, in our case, doing the ML estimation, then allows estimation of the fraction with a positive impact. We make the same point in two additional ways by performing Breusch–Pagan and likelihood ratio tests for groupwise heteroscedasticity (which in this case means different residual variances in the treated and untreated samples); Wooldridge (2002) and Judge et al. (1985) (and many others) describe these tests.

In the second step, we again decompose the impacts into systematic and idiosyncratic components by estimating (2) as in the preceding section. Statistically and substantively significant coefficients on interaction terms for variables correlated with the untreated outcome (which means all of them in our case) represent a rejection of the independence assumption for the model without covariates. This second scheme relies on the weaker (and more plausible) assumption of independence between the idiosyncratic component of the impacts and the untreated outcomes.

We then use the residuals from this regression to calculate the impact variance under independence as the difference between the variance of the treatment group residuals and the variance of the control group residuals. We calculate the variance (across persons) of the systematic component of the impact by constructing the estimated systematic component of the impact for each observation and then taking their variance. This procedure will somewhat overstate the population variance in the systematic component of the impacts due to the presence of estimation error in the estimated systematic component for each observation.

In the third step, we apply a recent estimator developed by Biddle et al. (unpublished manuscript) that restricts the distribution of random coefficients to come from the Pearson family of distributions. The Pearson family includes the normal, chi-square, beta and gamma distributions as special cases. Though it allows for only one mode it includes bell-shaped curves as well as J-shaped or U-shaped curves; see e.g. Kendall and Stuart (1963) for details. Though still a non-trivial assumption, this model substantially relaxes the normality assumption of the standard model. The Biddle et al. (unpublished manuscript) estimator relies on the first four moments of the treated and untreated outcome distributions to identify the particular member of the Pearson family relevant to a given data set under the random coefficient assumption. Their paper considers the more general case of observational data; we implement a simpler version (with no first stage nearest neighbor matching of treated and untreated units) given our access to experimental data. The Appendix in Biddle et al. (unpublished manuscript) derives the first four moments of β_{Di} from the first four moments of Y_1 and Y_0 under the independence assumption. The first four moments of β_{Di} uniquely identify a particular Pearson family member.

Finally, we do not undertake deconvolution as in Heckman et al. (1997) or Wu and Perloff (unpublished manuscript) because its complexity and related absence from standard software packages makes it unlikely to see much use in applied work.

7.3. Results and discussion

The first row of Table 7 presents estimates of the impact standard deviation obtained by taking the square root of the difference in the variance of the treated and untreated outcomes. Based on the bootstrap standard errors, we can easily reject the composite null of a non-positive variance in both 1998 and 1999. The estimate equals 53.17 in 1998 and 34.12 in 1999. To put this in context, note that transfer payments under the program averaged about 200 pesos per month, and recall from Table 3 that the mean impacts equal 4.83 and 15.79 pesos per month. The estimated impact standard deviations from the simple random coefficient model lie well within the FH bounds (another test) and substantially exceed the lower bound values.

The next two rows in Table 7 display results from Breusch–Pagan and LR tests of heteroskedasticity as a function of treatment status. These tests support the inference based on the estimated

Table 7
Estimates from random coefficient models

	Nov. 1998 per capita consumption	Nov. 1999 per capita consumption
Impact standard deviation based on difference in variances (no interactions)	53.17	34.12
	(12.89)	(12.47)
	[0.001]	[0.006]
Breusch–Pagan LM statistic	69.13	19.16
	[0.001]	[0.001]
LR test statistic	75.01	20.92
	[0.001]	[0.001]
Fraction with positive impact under normality	0.5571	0.6998
	(0.019)	(0.0741)
Systematic impact standard deviation	21.73	14.27
	(2.12)	(1.67)
	[0.001]	[0.001]
Idiosyncratic impact standard variation	52.83	34.47
	(12.50)	(12.73)
	[0.001]	[0.007]
Total impact standard deviation (with interactions)	57.73	38.14
	(12.64)	(12.84)
	[0.001]	[0.003]

Bootstrap standard errors appear in parentheses, the corresponding *p*-values appear in square brackets.

impact standard deviation as we can strongly reject the null of homoskedasticity and thus the null of a zero impact variance for both tests in both years.

The fourth row of Table 7 presents the estimated fraction with a positive impact under the assumption that the impacts have a normal distribution, which equals 0.55 in 1998 and 0.68 in 1999. As in the PPD case, this fraction turns out higher in 1999 than 1998, which again likely reflects the delayed implementation.

The fifth and sixth rows of Table 7 present the estimated standard deviations of the systematic and idiosyncratic impacts from model (2) under the independence assumption, while the seventh row gives the standard deviation of the overall impacts in this model. Because the presence of systematic impacts leads us to reject the simple random coefficient model, the overall impacts from the random coefficient model based on (2) need not have the same variance as those from the model without covariates. We find substantial systematic and idiosyncratic impact variance components, with the idiosyncratic component the larger of the two empirically, as well as a larger overall impact variance here than in the unconditional case.

Application of the estimator in Biddle et al. (unpublished manuscript), which imposes independence but relaxes the normality assumption in favor of membership in the broad Pearson family of distributions, leads to a Pearson Type I (Beta) distribution in 1998 and a Pearson Type IV distribution in 1999. Fig. 2 displays the estimated distribution of impacts for November 1998 and Fig. 3 presents the corresponding figure for 1999. We highlight two findings from this analysis. First, both distributions imply no negative impacts on consumption, unlike what we found under the PPD assumption or the random coefficient model with normality. Instead, they feature small impacts for many along with a long tail of larger positive impacts. Second, and perhaps most importantly, the distributions obtained from the Biddle et al. (unpublished manuscript) procedure differ quite strongly from the normal distribution commonly assumed in applications of the random coefficient model. Our results suggest the value of frameworks that relax the normality assumption.

8. Policy

We discuss the policy implications of our analyses under four headings: what policies to undertake (e.g. should there be a program like PROGRESA), how best to design a policy (e.g. how should the conditional transfers in PROGRESA vary with age), who should policies target (e.g. who should be eligible for PROGRESA) and how best to evaluate policies. We consider each heading in turn.

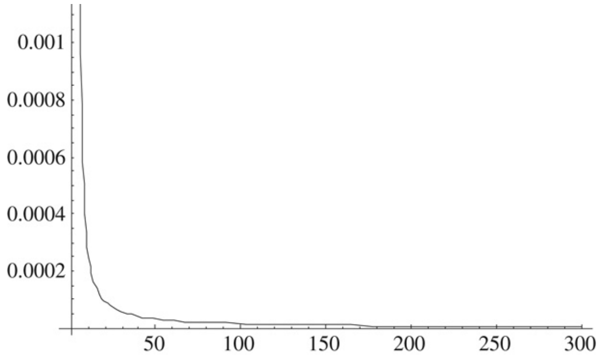

Fig. 2. Distribution of impacts from Biddle et al. (unpublished manuscript) Pearson family procedure. November 1998 per capita consumption.

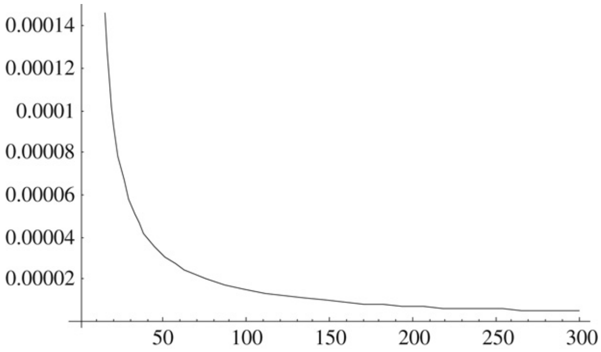

Fig. 3. Distribution of impacts from Biddle et al. (unpublished manuscript) Pearson family procedure. November 1999 per capita consumption.

The standard model of policy evaluation focuses, essentially, on changes in GDP between a world with and without a particular program. As discussed in Heckman and Smith (1998) this approach presumes that institutions exist such that policy winners compensate policy losers so as to make any policy that increases GDP Pareto improving. Of course, in practice such transfers often do not take place, or take place only in part or via inefficient in-kind transfers, as under the U.S. Trade Adjustment Act designed to compensate those who lose their jobs due to foreign competition.

In a world without complete compensation, a program with negative impacts for some individuals may increase GDP but not yield a Pareto improvement. As a result policymakers may want to take into account the number of policy losers and the size of their losses in addition to overall effects on GDP in choosing among alternative polices. More generally, policymakers may care about the entire distribution of impacts, whether positive or negative, due to concern about equity between and within groups or for crude political reasons. Our paper outlines and applies a variety of different methods for estimating the extent of impact heterogeneity and, more narrowly, the number made worse off by a program and the extent of their losses and so provides a template for similar analyses in other policy contexts.

In the narrower context of PROGRESA our analyses add value to the existing literature in (at least) three ways, but fall short of a full social welfare analysis incorporating heterogeneous impacts on all outcome variables of interest. First, our estimate of the FH bounds on the impact variance, combined with our rejection of the PPD assumption, together suggest the importance of heterogeneous impacts on consumption. These impacts embody a direct effect that increases consumption via the transfer payments and an indirect effect that reduces consumption via withdrawal of children from the paid labor force in favor of school attendance. Combined with heterogeneous impacts on schooling (implied by the fact that not all eligible families receive the transfer but school enrollment increases), our findings strongly imply heterogeneous welfare effects at the household level. Second, all of the models we estimate indicate at least half of the treatment group saw increased consumption, and most yield much higher fractions. These results suggest that only a modest minority of households experience reduced consumption as a result of responding to PROGRESA's incentives to remove children from the labor market and send them to school. Third, we present new subgroup impacts that reveal important variation in impacts by household poverty level, village marginality index and their interaction. Our estimates suggest that expanding the program to richer villages and/or better off households will yield smaller impacts on consumption at the margin.

Our work has little to say about the design of conditional cash transfer programs such as PROGRESA. The nature of the experiment, which included only one treatment arm, means that doing so requires a structural analysis along the lines of Todd and Wolpin (2006). Such an analysis lies beyond the scope of this paper.

Our findings regarding heterogeneity in program impacts have much to say about how to target programs in general and also about targeting in PROGRESA in particular. By targeting, we have in mind both the use of simple eligibility rules, e.g. household income less than some value, and more complex statistical treatment rules, as in the U.S. Worker Profiling and Reemployment Services System analyzed in Black et al. (2003). Targeting of social programs has occasioned much discussion in recent years in the policy literature.[13]

From an efficiency standpoint, optimal program allocation means assigning program eligibility to those with the largest impacts of participation net of costs. Although we have not combined our impacts on schooling, market work and consumption with one another or with impacts on other outcomes such as nutrition[14] into a single net impact estimate, and also do not take account of program costs, our finding of substantial heterogeneity in impacts based on observable characteristics suggests the potential for substantial efficiency gains from more finely tuned targeting of PROGRESA eligibility. Moreover, our evidence on idiosyncratic impact heterogeneity that remains after removing a substantial amount of systematic heterogeneity hints at the possibility for efficiency gains from targeting based on variables not included in our analysis.

Finally, many governments have no formal policy on when and how they will evaluate programs, an important omission in our view. Our analysis provides one additional piece of evidence regarding the high value of experimental evaluations of important policies that combine thoughtful design, careful implementation, high quality data collection and large sample sizes. Not only do such evaluations provide convincing mean impact estimates overall and for subgroups of interest, they also provide the foundation for valuable additional research on program design, impacts (as in our work) and on the population served by the program. In the particular context of PROGRESA, our finding of remaining heterogeneity in impacts even after removing systematic variation suggests the value of digging deeper into the economics and institutions when evaluating similar programs so as to collect variables that will capture this variation. Our results also highlight the value of longer term data collection, as our analysis of the 1998 data suffers from a lack of external validity due to unanticipated delays in program implementation.

9. Concluding remarks

Our paper lays out a variety of methods for examining heterogeneity in program impacts by estimating bounds on, or making additional assumptions about, the joint distribution of treated and untreated outcomes. We also present and apply methods for testing some of these assumptions. Our analysis includes an alternative to the Bitler et al. (unpublished manuscript) method for examining the importance of idiosyncratic impact heterogeneity after taking account of systematic heterogeneity based on observables. We apply these methods to the data from the recent PROGRESA evaluation.

The preceding section highlighted the policy relevance of our analysis and empirical findings. More generally, we have tried to integrate and augment the recent literature on heterogeneous treatment impacts as a way of providing a foundation for future analyses along these lines. We have shown the value of going beyond the widely used random coefficient model with normality of the impact distribution, to consider both other assumptions and more flexible versions of the random coefficient model. Careful thinking about heterogeneous treatment effects, and their formal integration into program evaluation, represents a profound (and as yet incomplete) departure from earlier practice. We intend our paper to speed that departure and thereby hasten the arrival of deeper, more relevant and more useful program evaluations.

Acknowledgements

We thank participants in the conference associated with this special issue, as well as seminar participants at the University of Guelph, the 2005 meetings of the European Economics Association, the World Bank Research Department, the 2004 Journée du CIRPÉE conference, the 2004 Northeast Universities Development Consortium Conference and the 2003 European Development Research Network Conference. We are especially grateful to our discussant, our editors, and two anonymous referees for helpful comments. We also thank the editors for their patience.

[13] See e.g. Berger et al. (2001), Eberts et al. (2002) and Schuck and Zeckhauser (2006) for general discussions of targeting and Behnke et al. (2007) for an evaluation of targeting in the context of Swiss active labor market policy. Manski (2005) and his related papers discuss statistical treatment rules at a higher conceptual and technical level.

[14] See Djebbari (unpublished manuscript) for an analysis of PROGRESA's impact on household nutrition.

Table A.1
Sample information — consumption analysis

	November 1998	November 1999
Number of households randomly assigned	24,073	22,116
Number of eligible households randomly assigned	18,743	17,293
Number of eligible treatment group households	11,585	10,475
Number of eligible control group households	7158	6818
Number of survey completers lost due to item non-response on items used in our analysis	2279	2863
Treatment group analysis sample representation rate[a]	87.72%	84.23%
Control group analysis sample representation rate[b]	88.02%	82.22%

[a] Equals the number of eligible treatment group households in our analysis sample divided by number of eligible treatment group households.
[b] Equals the number of eligible control group households in our analysis sample divided by number of eligible control group households.

Table A.2
Sample information — time use analysis

	June 1999
Number of children randomly assigned	37,977
Number of eligible children randomly assigned	31,438
Number of eligible treatment group children	19,327
Number of eligible control group children	12,111
Number of survey completers lost due to item non-response on items used in our analysis	3345
Treatment group analysis sample representation rate[a]	89.01%
Control group analysis sample representation rate[b]	89.10%

[a] Equal to number of eligible treatment group children in our analysis sample divided by number of eligible treatment group children.
[b] Equal to number of eligible control group children in our analysis sample divided by number of eligible control group children.

Appendix A

A.1. Details on the construction of the per capita consumption outcome

Per capita consumption is the average consumption in the household over all its members. It is comprised of food and non-food consumption. Food expenditures include household level data on food outlays made in the seven days preceding the interview for 36 food items. The value of food consumed from own production in that same period of time is added to food outlays to obtain the value of food consumption. Food consumed from own production is valued by imputing a locality level price (base on interviews with local leaders in each village). Non-food expenditures are expenses reported on a weekly, monthly and semi-annual basis. Non-food expenses reported on a weekly basis include transportation and tobacco. Monthly outlays include school tuition, health-related expenses, home cleaning, electricity and home fuel expenditures. Expenditures reported on a semi-annual basis include home and school supplies, clothes, shoes, toys and payments for special events. The value of consumption is computed as the sum of non-food expenditures and the value of food consumption.

A.2. Details on the construction of the time use indicators

We classify children's activities in three categories, namely schooling activities, income-generating ("work" in Table 2) activities and domestic activities, using a detailed module on time allocation. Income-generating activities are all activities involving work outside the house, including wage labor (for an employer, on one's own firm/farm with salary or other paid casual work) and non-wage labor (as an aide, on one's own firm/farm and other non-paid casual work). Domestic activities include all activities that take place at home and that could have been performed by someone hired by the household, such as cleaning the house, washing, sewing and ironing clothes, shopping for the household, preparing meals and washing dishes, fetching water or wood, disposing of garbage, taking care of animals and fields, looking after children (including taking them to school), or looking after elderly or sick people. Schooling activities consist of attendance at school and time spent studying outside the classroom.

Appendix B. Details of the Algorithms Used in the Paper

B.1. Table 3: Parameter estimates for the Frechét–Höffding bounding distributions

Bootstrap standard errors are obtained in the following manner:

1. Sample with replacement from the actual data (including both the treated and untreated observations) 1000 times, indexing each sample by $b = 1, \ldots, 1000$.
2. For the bth bootstrap sample:
   a. Collapse the $D = 0$ and $D = 1$ distributions into percentiles.
   b. Match the percentiles in ascending order, compute differences across percentiles to obtain the impact distribution in the perfect positive dependence case then compute the percent positive, the impact standard deviation and the outcome correlation.
   c. Match percentiles in descending order, compute differences across percentiles to obtain the impact distribution in the perfect negative dependence case, then compute the percent positive, the impact standard deviation and the outcome correlation.
3. The bootstrap standard error is the standard deviation of the statistics from steps 2b and 2c.

The Monte Carlo simulation procedure for testing the null of a zero impact variance proceeds as follows:

1. Sample with replacement from the control group sample in order to obtain a sample of size identical to the total (treatment and control) sample size for each survey round.
2. For each Monte Carlo sample:
   a. Randomly sort the data
   b. Let n denote the treatment group sample size in the actual data. Assign n observations to a synthetic treatment group $D^* = 1$ and N–n observations to a synthetic control group $D^* = 0$.
   c. Add the average impact value to each outcome of the $D^* = 1$ group.
   d. Collapse the $D^* = 1$ and $D^* = 0$ outcome distributions into percentiles.

e. Take the differences across percentiles and compute the standard deviation of these differences.

3. For all Monte Carlo samples:
   a. Sort the 100,000 samples in ascending order based on the standard deviation from step 1e.
   b. Use the relevant percentiles of the simulated distribution of impact standard deviations under the null to calculate the p-value for the test.

B.2. Table 4*: Treatment-control differences at quantiles of the outcome distribution*

We obtain confidence intervals for the treatment-control differences in mean outcomes at particular quantiles of the outcome distributions under the null of no population differences as follows:

1. Obtain a bootstrap treatment group sample by sampling with replacement from the treated observations. Obtain a bootstrap control group sample by sampling with replacement from the untreated observations. Combine the bootstrap treatment group and bootstrap control group into a full bootstrap sample. Repeat 1000 times, indexing the bootstrap samples by $b = 1, \ldots, 1000$.
2. For each bootstrap sample b:
   a. Randomly sort the data.
   b. Let n denote the treatment group sample size in the actual data. Assign n observations to a synthetic treatment group $D^* = 1$ and $N - n$ observations to a synthetic control group $D^* = 0$ group.
   c. Compute the mean of each observable characteristic for each quantile group of the synthetic treatment group outcome. Do the same for each quantile group of the synthetic control group outcome. Take the difference in these means.
3. For each characteristic and quantile group:
   a. Sort the 1000 samples in ascending order based on the differences in means for each characteristic and quantile group.
   b. Construct a confidence interval under the null that the difference in means for the quantile group is equal to zero by taking the relevant percentiles of the sorted values.
4. Using the actual treatment and control groups, compute the difference in means of observable characteristics for each quantile group as in step 2c. Compare the differences from the actual data to the confidence intervals to determine statistical significance.

B.3. Table 6*: Estimates including systematic impact heterogeneity*

The bootstrap standard errors are obtained as follows:

1. Sample with replacement from the actual data (including both experimental groups) 1000 times, indexing each sample by $b = 1, \ldots, 1000$.
2. For the bth bootstrap sample:
   a. Regress the outcome on a treatment indicator, the observable characteristics and interactions between the treatment indicator and each of the observable characteristics, including a triple interaction term between the household poverty score, the marginality index and the treatment indicator. Compute the average systematic impact and systematic impact standard deviation.
   b. Obtain the residuals from this regression.
   c. Collapse the $D = 0$ and $D = 1$ distributions of residuals into percentiles.
   d. Match percentiles in ascending order, compute differences across percentiles to obtain the impact distribution in the perfect positive dependence case and then compute the impact standard deviation.
   e. Match percentiles in descending order, compute differences across percentiles to obtain the impact distribution in the perfect negative dependence case and then compute the impact standard deviation.
3. Compute bootstrap standard errors using the sample of statistics from steps 2a, 2d and 2e.

B.4. Table 7*: Estimates from random coefficient models*

The bootstrap standard errors are obtained using 500 bootstrap replications. Details of the procedure for each statistic follow:

1. Bootstrap standard errors of the impact standard deviation and the fraction with a positive impact in the case with no interactions:
   a. Sample with replacement from the actual data 500 times, indexing each sample by $b = 1, \ldots, 500$.
   b. For the bth sample:
      (i) Estimate the average impact on the outcome of interest by OLS by regressing the outcome on the treatment indicator and a set of covariates.
      (ii) Square the predicted residuals from the regression. Regress the squared residuals on the treatment variable. The square root of the estimated coefficient is an estimate of the impact standard deviation for the bth sample.
      (iii) Using the average impact from step (i), the impact standard deviation from step (ii) and standard normal random numbers, estimate the fraction with a positive impact for a normal distribution with a mean equal to the average impact and the corresponding standard deviation.
   c. The bootstrap standard errors of the impact standard deviation from step (ii) and the fraction with a positive impact under normality from step (iii) are given by the standard deviation of these values from the bootstrap samples.
2. Bootstrap standard error for the systematic impact standard deviation:
   a. Sample with replacement from the actual data 500 times, indexing each sample by $b = 1, \ldots, 500$.
   b. For the bth sample:
      (i) Regress the outcome on the treatment indicator, a set of covariates and interactions between the treatment indicator and the covariates and then predict the systematic impact for each sample member using the estimated coefficients.
      (ii) Compute the standard deviation of the systematic impacts.
   c. The bootstrap standard error is the standard deviation of the statistics from (ii).
3. Bootstrap standard error for the idiosyncratic impact standard deviation:
   a. Sample with replacement from the actual data 500 times, indexing each sample by $b = 1, \ldots, 500$.
   b. For the bth sample:
      (i) Regress the outcome on the treatment indicator, a set of covariates and interactions between the treatment indicator and the covariates.
      (ii) Square the predicted residuals and regress the squared residuals on the treatment indicator. The square root of the estimated coefficient is an estimate of the idiosyncratic impact standard deviation for the bth sample.
   c. The bootstrap standard error is the standard deviation of the statistics from (ii).

References

Abadie, A., Angrist, J., Imbens, G., 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70, 91–117.

Andrews, D., 2000. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. Econometrica 68, 399–405.

Angelucci, M., De Giorgi, G., 2007. Indirect effects of an aid program: How do cash injections affect non-eligibles' consumption? American Economic Review (forthcoming).

Angrist, J., Imbens, G., Rubin, D., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–472.

Behnke, S., Frölich, M., Lechner, M., 2007. Targeting labour market programmes: Results from a randomized experiment. IZA Discussion Paper No. 3085.

Behrman, J., Todd, P., 1999. Randomness in the experimental samples of PROGRESA. International Food Policy Research Institute, Washington, DC.

Behrman, J., Sengupta, P., Todd, P., 2005. Progressing through PROGRESA: An impact assessment of a school subsidy experiment in rural Mexico. Economic Development and Cultural Change 54, 237–275.

Berger, M., Black, D., Smith, J., 2001. Evaluating profiling as a means of allocating government services. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Active Labour Market Policies. Physica 59–84.

Black, D., Smith, J., Berger, M., Noel, B., 2003. Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. American Economic Review 93, 1313–1327.

Biddle, J., Boden, L., Reville, R., 2004. A method for estimating the full distribution of a treatment effect, with application to the impact of workplace injury on subsequent earnings. Rand Corporation (unpublished manuscript).

Bitler, M., Gelbach, J., Hoynes, H., 2005. Distributional impacts of the Self-Sufficiency Project. NBER Working Paper No.11626.

Bitler, M., Gelbach, J., Hoynes, H., 2006. What mean impacts miss: Distributional effects of welfare reform experiments. American Economic Review 96, 988–1012.

Bitler, M., Gelbach, J., Hoynes, H., 2007. Can subgroup-specific mean treatment effects explain heterogeneity in welfare reform effects? Evidence from Connecticut's Jobs First experiment. University of California, Irvine (unpublished manuscript).

Bitler, M., Gelbach, J., Hoynes, H., 2008. Distributional impacts of the Self-Sufficiency Project. Journal of Public Economics 92, 748–765.

Bloom, H., 2005. Learning More from Social Experiments. Russell Sage, New York.

Bloom, H., 2006. The core analytics of randomized experiments for social research. MDRC Working Papers in Research Methodology.

Brendstrup, B., Paarsch, H., 2007. Semiparametric identification and estimation in multi-object, English auctions. Journal of Econometrics 141, 84–108.

Burtless, G., 1995. The case for randomized field trials in economic and policy research. Journal of Economic Perspectives 9, 63–84.

Cambanis, S., Simons, G., Stout, W., 1976. Inequalities for $E(k(X, Y))$ when the marginals are fixed. Zeitschrift für Wahrscheinlichkeitstheorie und Verwaltung 36, 285–294.

Chavez-Martin del Campo, J.C., 2006. Does conditionality generate heterogeneity in program impacts? The Progresa experience. Universidad de Guanajuato (unpublished manuscript).

Chernozhukov, V., Fernandez-Val, I., 2005. Subsampling inference on quantile regression processes. Sankhya 67, 253–276.

Dammert, A., 2007. Heterogeneous impacts of conditional cash transfers: Evidence from Nicaragua. McMaster University (unpublished manuscript).

Deaton, A., 1997. The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Johns Hopkins University Press, Baltimore.

Doksum, K., 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. The Annals of Statistics 2, 267–277.

Djebbari, H., 2004. The impact on nutrition of the intrahousehold distribution of power. Université Laval (unpublished manuscript).

Duflo, E., Glennerster, R., Kremer, M., 2006. Using randomization in development economics research. NBER Technical Working Paper No. 333.

Eberts, R., O'Leary, C., Wandner, S., 2002. Targeting employment services. W.E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Fisher, R., 1935. The Design of Experiments. Oliver and Boyd, London.

Fréchet, M., 1951. Sur les Tableaux de Corrélation Dont les Marges Sont Données. Annales de l'Université de Lyon. Section A: Sciences, Mathématiques et Astronomie 14, 53–77.

Frost, R., 1920. The road not taken. In: Frost, R. (Ed.), Mountain Interval. Henry Holt, New York.

Heckman, J., 1996. Randomization as an instrumental variable. Review of Economics and Statistics 78, 336–341.

Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of active labor market programs. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics. vol. 3A. North-Holland, Amsterdam, pp. 1865–2097.

Heckman, J., Smith, J., 1995. Assessing the case for social experiments. Journal of Economic Perspectives 9, 85–110.

Heckman, J., Smith, J., 1998. Evaluating the welfare state. In: Strom, S. (Ed.), Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial. In: Monograph Series, Cambridge University Press for Econometric Society, New York, pp. 241–318.

Heckman, J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. Review of Economic Studies 64, 487–535.

Heckman, J., Smith, J., Taber, C., 1998. Accounting for dropouts in the evaluation of social programs. Review of Economics and Statistics 80, 1–11.

Heckman, J., Vytlacil, E., 1998. Instrumental variables methods for the correlated random coefficient model. Journal of Human Resources 33, 974–1002.

Hildreth, C., Houck, J., 1968. Some estimators for a linear model with random coefficients. Journal of the American Statistical Association 63, 584–595.

Höffding, W., 1940. Masstabinvariante korrelationsmasse für diskontinuierliche verteilungen. Arkiv für Matematischen Wirtschaften and Sozialforschung 7, 49–70.

Horwitz, R., Singer, B., Makuch, R., Viscoli, C., 1996. Can treatment that is useful on average be harmful for some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. Journal of Clinical Epidemiology 49, 395–400.

Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–475.

Judge, G., Griffiths, W., Hill, C., Lutkepohl, H., Lee, T., 1985. The Theory and Practice of Econometrics, 2nd ed.. Wiley, New York.

Kendall, M., Stuart, A., 1963. Volume I: Distribution Theory, 2nd ed. In: The Theory of Advanced Statistics, Griffin, London.

Koenker, R., Basset, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Koenker, R., Billias, Y., 2001. Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments. Empirical Economics 26, 199–220.

Lehman, Erich, 1974. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

Manski, C., 2005. Social Choice with Partial Knowledge of Treatment Response. Princeton University Press, Princeton.

Mardia, K., 1970. Families of Bivariate Distributions. Griffin, London.

Nelsen, R., 1999. An Introduction to Copulas. Springer, New York.

Neyman, J., 1923. Statistical problems in agricultural experiments. Journal of the Royal Statistical Society 2, 107–180.

Quandt, R., 1972. Methods of estimating switching regressions. Journal of the American Statistical Association 67, 306–310.

Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers 3, 135–146.

Rubin, D., 1974. Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology 66, 688–701.

Rüschendorf, L., 1981. Sharpness of Fréchet bounds. Zeitschrift für Wahrscheinlichskeitstheorie 41, 293–302.

Schuck, P., Zeckhauser, R., 2006. Targeting in Social Programs: Avoiding Bad Bets, Removing Bad Apples. Brookings Institution, Washington, DC.

Schultz, T.P., 2004. School subsidies for the poor: Evaluating the Mexican Progresa poverty program. Journal of Development Economics 74, 199–250.

Skoufias, E., 2005. Progresa and its Impacts on the Welfare of Rural Households in Mexico. International Food Policy Research Institute, Washington, DC.

Skoufias, E., Davis, B., de la Vega, S., 2001. Targeting the poor in Mexico: An evaluation of the selection of households for PROGRESA. World Development 29, 1769–1784.

Skoufias, E., Parker, S., 2001. Conditional cash transfers and their impact on child work and school enrollment: Evidence from the PROGRESA program in Mexico. Economia 2, 45–96.

Tobias, J., 2006. Estimation, learning and parameters of interest in a multiple outcome selection model. Econometric Reviews 25, 1–40.

Todd, P., Wolpin, K., 2006. Using experimental data to validate a dynamic behavioral model of child schooling: Assessing the impact of a school subsidy program in Mexico. American Economic Review 96, 1384–1417.

Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.

Wu, X., Perloff, J., 2006. Information-theoretic deconvolution approximation of treatment effect distribution. University of California at Berkeley (unpublished manuscript).