

Review of CCT Impact Evaluations

CCTS HAVE BEEN REMARKABLE IN A VARIETY OF WAYS. ONE OF those ways is that perhaps more than any intervention in developing countries, CCTs have been evaluated credibly for their impact on a variety of outcomes—consumption, labor market participation, poverty, nutritional status, and schooling to name but a few. Indeed, it would not have been possible to write this report, at least not in its current form, without these evaluations to draw upon. This appendix discusses the strengths and weaknesses of some of the evaluations of CCTs that have been conducted. It does not, however, attempt to be an exhaustive methodological discussion of all available evaluations of CCT programs.

Impact evaluations involve credibly estimating counterfactual outcomes—the value an outcome would have taken if a given individual who benefited from a program had not received the benefit. (The same logic obviously also applies to other units, such as households, schools, or municipalities.) However, a given individual is never observed having both received and not received an intervention at the same point in time. Impact evaluation therefore can be thought of as a problem of missing data.

Drawing on the medical literature, studies of impact evaluation often refer to comparisons between a “treatment” group (those who received an intervention) and a “comparison” or “control” group (those who did not receive it). The comparison or control group is constructed in such a way as to make it an appropriate counterfactual for the treatment group. The difficulty therefore involves making those two groups comparable, except for the presence or absence of an intervention. For example, an evaluation of the impact of a CCT program on schooling would attempt to ensure that treatment and control groups are truly comparable in

terms of both their “observable” characteristics (variables like parental education) and their “unobservable” characteristics (variables like the motivation or inherent ability of children). A failure to make the two groups comparable in terms of those and other characteristics could bias the results.

There are different ways of estimating counterfactual outcomes, including random assignment, “quasi-experimental” methods like instrumental variables and regression discontinuity (RDD), and nonexperimental methods like regression techniques, matching, and double (or higher-order) differencing. All of those methods have their strengths and weaknesses, and all will be more credible in some settings than in others. Indeed, one of the most important lessons from the rapidly growing literature on impact evaluation is that blindly applying a given method or technique is unlikely to be a sensible approach to the evaluation problem. Rather, what is needed is a careful and thoughtful analysis of the extent to which the assumptions made by each of those methods are likely to hold when attempting to answer a particular question with a given data set.

The evaluations of CCTs have used a variety of methods. A number of programs have been evaluated using random assignment. Randomization involves using a lottery to assign one group to treatment and another to control. If the sample is large enough, this method has the virtue of equating all characteristics, observable as well as unobservable, of the treatment and control groups. Differences in outcomes between the two groups after the intervention then can be interpreted credibly as causal estimates of program impact. Because randomization requires no further assumptions, it is often regarded as the gold standard for evaluations.

When Mexico’s Oportunidades program began its operations in rural areas in the late 1990s, it randomly assigned a subsample of eligible villages to treatment and control groups. The first group of villages began receiving the program in 1998, whereas the second group was held back for approximately one year. In addition, rather than conducting an in-house evaluation, Oportunidades administrators hired the International Food Policy Research Institute and a respected consortium of international researchers to conduct the evaluation. Also, the data from the evaluation were made available to the public on the Internet so that other researchers could replicate or challenge the findings. Considering that it was difficult to predict *ex ante* whether the program would work,

these decisions were very brave—and influential. But the decisions have been vindicated: the Oportunidades data have been used in dozens of studies, and were influential in causing the spread of CCTs beyond the countries where they first were implemented, Brazil and Mexico.

In this report, we draw heavily on the Oportunidades data, both using existing studies and in our own calculations. Some of the more influential papers using the Oportunidades data include Schultz (2004), Behrman, Sengupta, and Todd (2005), and de Janvry and Sadoulet (2006) on education outcomes; Gertler (2004), Rivera et al. (2004), and Behrman and Hoddinott (2005) on nutrition outcomes; and Hoddinott and Skoufias (2004) and Skoufias (2005) on consumption patterns and poverty. More recently, the random assignment of Oportunidades has been used to estimate longer-term effects on outcomes, including completed schooling and test scores (Behrman, Parker, and Todd 2005), and investment and savings behavior (Gertler, Martínez, and Rubio-Codina 2006)—and we draw heavily on those studies as well. Finally, a handful of recent reports make use of the randomized assignment in Oportunidades to estimate structural behavioral models (Attanasio, Meghir, and Santiago 2005; Todd and Wolpin 2006a).

Nevertheless, even the Oportunidades data have their limitations (see, in particular, thoughtful discussions by Parker and Teruel [2005] and Parker, Rubalcava, and Teruel [2008]). Despite the experimental design, there appear to have been some significant differences between individuals who received transfers and those who did not (Behrman and Todd 1999). As a result, many studies using the Oportunidades data have focused on differences in the growth rates of outcomes between treated and control communities or individuals—a so-called differences-in-differences approach—rather than on simple differences in outcomes at follow-up. This approach is sensible and will tend to remove the source of bias if it is time invariant and additive—probably a reasonable assumption.

Another shortcoming of the Oportunidades data is that merging the data across waves of the surveys, which is necessary to construct the panels needed for a differences-in-differences approach, appears to have been a serious problem, with large fractions that could not be merged, especially in the evaluations that have used the data on anthropometrics. This shortcoming leaves researchers analyzing the impact of Oportunidades transfers on nutrition outcomes with the difficult choice

between two options: (1) to work with a smaller panel of households or children that could be merged effectively (the approach taken by Behrman and Hoddinott [2005]), which could result in biases associated with large and possibly nonrandom attrition out of the sample; or (2) to ignore any baseline differences between the two groups (the approach taken by Gertler [2004]), which essentially assumes that the differences between the two groups are negligible. More generally, attrition across survey rounds in Oportunidades appears to be non-negligible and correlated with the likelihood of being in the program, which also can introduce biases (see the discussion in Parker, Rubalcava, and Teruel [2008]).

Following from the Oportunidades evaluation, a number of programs in other countries launched randomized evaluations. These included evaluations of the RPS and Atención a Crisis programs in Nicaragua, PRAF in Honduras, and the BDH program in Ecuador. We use those evaluations quite extensively here, although some have limitations that we discuss below.

The evaluations of the RPS and Atención a Crisis programs in Nicaragua seem to have worked well. In both cases, the randomized design was successful—there appear to be no significant differences between treated and control households at baseline. Attrition rates in the evaluation of RPS were reasonably low (approximately 15 percent over four years) and, in the case of the evaluation of Atención a Crisis, extremely low (only 1.3 percent of households were lost between baseline and follow-up, although the period between the two surveys was short—approximately nine months). Moreover, attrition appears to be uncorrelated with treatment status, and the characteristics of attrited and other households were very similar—again, limiting the potential for important biases. Finally, there was no contamination of the control group, and take-up among eligible households was high. For all of those reasons, reports based on those evaluations—including Maluccio and Flores (2005), Maluccio (2005, 2008), and Macours, Schady, and Vakis (2008)—are likely to provide robust evidence of the impact of CCT programs in Nicaragua, at least during a pilot phase.

The randomized evaluation of PRAF in Honduras also appears to have worked reasonably well, although it faced a number of challenges. On the health side, the evaluation design originally considered four groups: (1) one group of municipalities in which households would receive the CCT, (2) another group in which there would be a supply-

side intervention to improve health services, (3) a group of municipalities that would receive both interventions, and (4) a group that would serve as a control. In practice, however, the supply-side intervention was not implemented and so could not be evaluated (Morris, Flores et al. 2004). Moreover, because of the relatively small number of households involved, there were some important baseline differences. For example, at baseline, the fraction of households that had received five or more antenatal visits was 37.9 percent in the group randomly assigned to receive the CCT intervention only, and 48.9 percent in the control group. That finding raises the possibility that some of the impact that was estimated—an 18.7 percentage point impact on the probability that a woman had received five antenatal visits—could be a result of mean reversion, as treated households simply caught up with those in the control group. Also, the evaluation of the effects of PRAF on education faced a problem in that the baseline survey was collected first among households in the treatment group (between August and October 2000) and only then for households in the control group (between November and December 2000). As Glewwe and Olinto (2004) discuss, that complicates matters because November and December are important coffee-harvesting months in Honduras and therefore baseline levels of child labor were significantly higher in control than in treatment areas, and school attendance levels were lower. For most outcomes, the authors reasonably focus on single-difference estimates of program impact (differences between treatment and control groups at follow-up), rather than differences-in-differences (differences in the growth rates between treatment and control groups), because the latter could have been biased by the artificially high levels of child labor at baseline among the control group. That kind of unexpected complication underlines the challenges of running randomized evaluations in practice, although some of the same problems obviously can occur with nonexperimental evaluations.

In Ecuador, there have been numerous evaluations of the BDH program. Paxson and Schady (2008) use panel data to estimate program effects on measures of child health and cognitive development. Households were assigned randomly to treatment and control groups, and there were no differences in observables at baseline between the two groups. Take-up among the treated was reasonably high (approximately 75 percent) and contamination of the control group was low (less than 4 percent). Attrition in the survey was low as well—6 percent of the

sample at baseline could not be re-interviewed at follow-up—and is uncorrelated with treatment status.

Other evaluations of the BDH pose more serious identification challenges. The data used by Edmonds and Schady (2008), Schady and Araujo (2008), and Schady and Rosero (2008) to analyze the impact of the program on school enrollment, child work, and household consumption patterns also are based on a randomized experiment. A lottery was used to assign households with school-age children to treatment and control groups, and that lottery appears to have been successful (the authors document that there are no baseline differences between the two groups in observable characteristics). However, there was a substantial contamination of the control group, 48 percent of whom received transfers. The precise reasons for the contamination are unclear. It appears that the list of households randomly excluded from the program was not passed on immediately to operational staff activating households for transfers. That situation was corrected after a few weeks but, as the authors explain, withholding transfers from households that already had begun to receive them was no longer feasible. Moreover, the contamination of the control group clearly was nonrandom: Schady and Araujo (2008) document significant differences between households that actually received transfers and those that did not (as opposed to lottery winners and lottery losers), especially with regard to education levels.

The solution adopted by Edmonds and Schady, Schady and Araujo, and Schady and Rosero is the following. They first focus on differences in outcomes between households assigned to the treatment and control groups by the lottery, rather than on differences between those who received the transfers and those who did not receive them. These so-called intent-to-treat effects abstract from the contamination of the experiment, and provide a lower-bound on the estimated impact of the BDH. The authors also present estimates in which assignment by the lottery is used as an instrument for receiving BDH transfers. That approach—using “partial randomization” as an exogenous source of variation, as proposed by Imbens and Angrist (1994)—is convincing because lottery status is clearly random, and Schady and Araujo show that there is a strong first stage. Nevertheless, it is not without costs. The estimated coefficients are local average treatment effects (LATE) that apply to “compliers”—those whose probability of receiving transfers was affected by the lottery (see Angrist, Imbens, and Rubin 1996). Those compliers cannot be identified without additional assumptions. If there

is heterogeneity of treatment effects, the LATE coefficients (although unbiased for the group of compliers) may not be relevant for other households in the sample. The external validity of the results reported in these papers, as in any other instrumental-variables regression, therefore may be somewhat limited.

Another paper that uses the Ecuador data to estimate BDH program effects on school enrollment is by Oosterbeek, Ponce, and Schady (2008). The authors begin by reproducing results very similar to those in Schady and Araujo (2008). As they point out, however, the sample of households used by Schady and Araujo are all drawn from “around” the 20th percentile of the proxy means. The reason for this is that the BDH originally envisioned two tiers of transfers, corresponding to households in the first and second quintiles of the proxy means. The original evaluation design therefore was based on RDD, with two cut-offs—one at the threshold between the first and second quintiles, another at the threshold between the second and third quintiles. However, after the sample was drawn, but before any of the households started receiving transfers, President Lucio Gutiérrez announced that all households in the first and second quintiles of the proxy means would receive transfers of the same magnitude. That decision obviously invalidated the original evaluation design for households around the threshold between the first and second quintiles. As a solution to this problem, it was agreed that the sample of households around that lower threshold would be assigned randomly to treatment and control groups—regardless of whether they were “just above” or “just below” the original cut-off.

Oosterbeek, Ponce, and Schady (2008) compare BDH program effects “around” the 20th percentile of the proxy means—estimated by instrumental variables, as described above—with those “around” the 40th percentile of the proxy means. Those latter estimates use RDD. In practice, this is a case of “fuzzy” RDD—households below the cut-off established by the 40th percentile of the proxy means are much more likely to receive transfers than those above, but a small fraction of ineligibles (approximately 8 percent) nonetheless received BDH transfers. Oosterbeek, Ponce, and Schady therefore instrument receiving BDH transfers with a dummy variable that takes on the value of 1 for households below the cut-off given by the 40th percentile, after flexibly accounting for the relationship between school enrollment and the score on the proxy means test. On the basis of those estimates, they conclude that the BDH had an impact on the enrollment decisions

made by “very poor” households (those around the 20th percentile of the proxy means) but no effect for “less poor” households (those around the 40th percentile). That conclusion is plausible, given that there is a good deal of evidence suggesting that CCT program effects on human capital outcomes, including school enrollment, tend to be larger among poorer households (for example, Maluccio and Flores [2005] on the RPS in Nicaragua; Filmer and Schady [2008] on the JFPR program in Cambodia). Nevertheless, the fact that the LATE estimates around the 20th and 40th percentiles of the proxy means refer to different groups of “complier” households somewhat muddies the interpretation put forward by Oosterbeek, Ponce, and Schady.

Two other evaluations of CCT programs use instrumental variables. Morris, Olinto et al. (2004) and Braido, Olinto, and Perrone (2008) evaluate the impact of the Bolsa Alimentação program in Brazil. The identification in those reports is ingenious. Both papers describe a series of administrative errors whereby some potential beneficiaries inadvertently were excluded from program benefits. Entire batches of beneficiaries were lost when files were transferred from participating municipalities to a central data-processing unit in Brasília, and the data-processing software initially rejected applications with names having nonstandard characters, such as *é*, *ç*, or *ô*. Morris, Olinto et al. and Braido, Olinto, and Perrone argue that this source of variation is as good as random, and therefore is uncorrelated with potential outcomes. That argument seems convincing. But given that these are LATE, the external validity of the estimated effects is unclear.

During the Indonesian crisis of 1997–98, the government made children in poor households eligible for a “scholarship” program. Given the crisis context, it is not surprising that little attention was paid to a possible evaluation of the effect of the program. Sparrow (2007) runs ordinary least squares regressions that suggest the program increased enrollment for children aged 10–12 by about 8 percentage points. He also uses “mistargeting” resulting from outdated poverty data as an instrument for receipt of the scholarship program. On the basis of those calculations, he estimates a larger program effect on enrollment (about 10 percentage points) for children aged 10–12. However, the identifying assumption—in effect, that enrollment decisions respond to current but not lagged poverty levels—is open to question.

A reasonably large number of papers have used RDD to estimate CCT program effects. In addition to Oosterbeek, Ponce, and Schady

(2008), discussed above, these include evaluations of Chile Solidario (Galasso 2006), of the PATH program in Jamaica (Levy and Ohls 2007), of the CESSP program in Cambodia (Filmer and Schady 2009a, 2009b, 2009c), and of the Turkey Social Risk Mitigation Project (Ahmed, Adato et al. 2006; Ahmed, Gilligan et al. 2006; Ahmed et al. 2007).

Levy and Ohls (2007) report intent-to-treat estimates of the impact of PATH. Take-up was high among eligible households (those below the cut-off of the proxy means)—approximately 80 percent—and the fraction of ineligible households (those above the cut-off) was reasonably low—approximately 10 percent. The authors collected both baseline and follow-up data. They experiment with various control functions for the proxy means, and settle on a linear formulation. They also present the results from a variety of placebo experiments, all of which suggest that, controlling for a linear formulation, there are no jumps at the threshold of the proxy means at baseline in any of a large number of observables. That finding adds considerable credibility to the identification strategy. One potential source of concern is the fact that the group of households that received PATH transfers (the treatment group) appears to have applied to the program somewhat earlier than those who did not receive transfers (the comparison group). That fact raises the possibility that there was selection on some unobservable related to “eagerness” or “need.” However, the solution adopted by Levy and Ohls—to control for the date of application in all of the main regressions—seems reasonable.

Ahmed, Adato et al. (2006) and Ahmed, Gilligan et al. (2006) also use RDD to estimate the impact of the CCT program in Turkey. As with other CCTs, the score on the proxy means is a significant but imperfect predictor of treatment: about 9 percent of households do not “comply” with their assignment (either eligible households that do not receive transfers, or ineligible households that do receive them). A conservative and standard approach to the problem of imperfect compliance would have been to use the initial assignment by the proxy means to calculate intent-to-treat estimates of program effects, or to calculate LATE estimates by instrumenting program participation with the eligibility rule based on the proxy means. Instead, the authors simply drop those groups of “ineligible beneficiaries” and “eligible nonbeneficiaries” from their sample, despite the fact that, as they acknowledge, “dropping those households from the sample for estimation contributes potential bias to the impact estimates” (Ahmed et al. 2007, p. 123).

Other papers have used double- or triple-differencing techniques to estimate CCT program effects. Both Filmer and Schady (2008) and Chaudhury and Parajuli (2008) estimate the effect of a CCT program for which girls, but not boys, are eligible in Cambodia and in the Punjab area of Pakistan, respectively. Filmer and Schady first compare the growth rates of girls' enrollment in districts that were eligible for the JFPR scholarship program with those that were not eligible. However, they show that preprogram growth rates in girls' enrollment were already higher in eligible districts, which suggests that the common trends assumption underlying their double-differencing estimates is unlikely to hold. They therefore use triple-differencing techniques, comparing the growth rate of girls' enrollment, relative to boys' enrollment, in JFPR-eligible and in other districts. The authors show that this growth rate is higher in JFPR-eligible districts. A similar approach (using boys as an additional control in estimation), with similar conclusions, is followed by Chaudhury and Parajuli in their analysis for the Punjab area of Pakistan.

Triple-differencing of this sort can provide credible estimates of program effects under reasonable assumptions—essentially, that in the absence of the program, the enrollment of girls, relative to that of boys, would have grown by the same amounts in treated and control districts. Showing that preexisting trends in the relative enrollment growth rate are very similar, as is done in Filmer and Schady (2008), provides reassurance on the identification strategy. In addition, both Chaudhury and Parajuli and Filmer and Schady compare the results from this triple-differencing technique with other estimates, using different data sets (for example, household data rather than administrative data), and they show that the estimated effects are very similar.

Separately estimating program effects using household and administrative data is also the basis of Khandker, Pitt, and Fuwa's (2003) analysis of the Female Stipend Program in Bangladesh. The authors show that estimates of program effects on girls' enrollment are similar using both sources of data. They also present estimates of program effects for boys, who were ineligible for the program. Using household data, they find no effect on boys' school enrollment; but using the administrative data, they estimate a worryingly large, negative effect of the program—29 percentage points, or about three times the magnitude of the positive effect on girls' enrollment. The authors point out that the administrative data cover only Female Stipend Program

schools, whereas the household data cover enrollment at any school, regardless of whether it was included in the program. Khandker, Pitt, and Fuwa suggest that the difference in boys' effects in the administrative and household data is a result of the transfer of boys out of program schools.¹ Although that suggestion is plausible, the very large magnitude of the coefficient raises some concerns about the estimation strategy and results in the Bangladesh study.

Attanasio, Battistin et al. (2005), Attanasio, Gómez et al. (2005), and Attanasio et al. (2006) identify program effects on the basis of changes over time in treatment and a matched set of ineligible communities to estimate the impact of the Familias en Acción program in Colombia. The identifying assumption therefore is that outcomes would have followed the same trends in both groups of communities in the absence of the program. As with any evaluation that matches eligible and ineligible communities, there is a concern that the characteristics that define eligibility are themselves correlated with outcomes or changes in outcomes. This is untestable, but the authors provide some ancillary support for their identification strategy: They show that average per capita household labor income was higher in comparison communities than in treatment communities prior to the implementation of the Familias en Acción program, but the *trends* in income over three preprogram years are similar. Nevertheless, that evaluation faced other challenges, including the fact that participation in the Familias program made households ineligible to participate in a community-based child care program, Hogares Comunitarios.

Another complication that arose in the Familias en Acción evaluation resulted from the program already having been announced in the treatment areas at the time the baseline survey was collected. As a result, families in treatment areas may have anticipated the effect of the program by enrolling their children in school. Under those circumstances, differences-in-differences that focus on changes in enrollment $E_t - E_{t-1}$ would likely underestimate the true program effects. Foreseeing this problem, the evaluation team collected retrospective data on schooling at the time of the "baseline" survey, and constructed a "pre-baseline" measure of school enrollment, E_{t-2} . This, rather than the measure E_{t-1} , is used in the differences-in-differences estimation.

A similar estimation strategy—first differences combined with matching—is also the basis for a number of evaluations of the impact of the urban Oportunidades program in Mexico. However, the pattern

of program effects estimated with those data is somewhat surprising. For example, Todd et al. (2005) estimate that the largest impacts of urban Oportunidades on school enrollment are found among children aged 6–7 years at baseline—a finding that is puzzling on a number of grounds: baseline enrollment in the urban sample for this age group is high, and decreases with child age; children in this age group would have been enrolled in grades that were ineligible for subsidies; and finally, the results for the Oportunidades sample in rural areas suggest that program effects are largest for children in age groups close to the transition from primary to secondary school, rather than for the youngest children (Schultz 2004). Although it is conceivable that the pattern of program effects in rural and urban areas of Mexico are very different, it is also possible (and perhaps more likely) that the double-differencing, matching estimation strategy introduced some hard-to-correct-for biases into the urban estimates.

Given that all communities in the original (randomized) Oportunidades rural sample started receiving payments in December 1999, more recent evaluations of the impact of Oportunidades in rural areas also have had to rely on matching to create a set of comparison communities. However, that effort has faced a number of important difficulties. One hundred fifty-two comparison communities were matched from a pool of 14,000 potential communities that had not received the program. The matching was done on the basis of the locality-level information from the 2000 Mexican census.

There are a number of reasons why this comparison group—and estimates of program effects that use it—should be treated with caution.² First, the comparison communities were drawn from different geographic areas than was the treatment group, and therefore they may have had other local area effects that could affect the levels or changes in the outcomes of interest. Second, although there appear to be no differences between the matched sets of treatment and comparison communities (not surprising, given that community characteristics were used to create the matches), individuals in the two sets of communities differ significantly in virtually every characteristic analyzed, and the differences often are large. For example, mean schooling levels of the household head and his spouse are approximately 2.7 in the original Oportunidades communities, but 4.5 in the matched comparison communities. Clearly, this could introduce a number of important biases. Third, to construct a “pre-intervention baseline” for the comparison

communities, data were collected on households in those communities in 2003, asking them about their characteristics in 1997. It seems reasonable to assume that this would introduce a good deal of measurement error based on recall bias to those data. Indeed, some recent work on China suggests that such data collection can work quite poorly (Chen, Mu, and Ravallion 2006). Moreover, these retrospective, pre-intervention data were collected in the matched set of communities, but not in the original Oportunidades communities for which the data originally collected in 1997 were used. As a result, recall bias may affect the propensity score used for matching. Finally, migration may have introduced selection problems if the sample of people living in the comparison communities in 2003—those who were asked about their characteristics in 1997—was different from those people who actually lived there in 1997. For all of those reasons, and because of the abundance of studies on Mexico that use the original, randomized (and likely more credible) data collected in the first generation of Oportunidades evaluations, in this report we do not make extensive use of these “second-generation” data collected in recent rounds of the Oportunidades evaluations. Our choice obviously has some costs because it limits the extent to which we can discuss program impacts on outcomes that only recently have been collected in the Oportunidades surveys (for example, adult obesity, hypertension, diabetes, or child cognitive development).

Similarly, we do not make extensive use of a handful of other evaluations, including two that are available for CCT programs in Latin America. In Brazil, Cardoso and Portela Souza (2004) use data from the 2000 population census to evaluate the impact of the Bolsa Escola program. They conclude that children in households that received cash transfers were 3–4 percentage points more likely to attend school than were matched children in the control group. However, the set of covariates used to construct the propensity score is small, and it is not immediately clear why “comparable” households received transfers in some cases but not in others. Moreover, Cardoso and Portela Souza are not able to disentangle transfers made by Bolsa Escola, the CCT program, from other income transfer programs in Brazil.

An evaluation also is available for the Programa Nacional de Becas Estudiantiles in Argentina. Heinrich (2007) uses matching methods to make two sets of comparisons: first, between children who were in the Becas program and other children, and, second, between children who were in the Becas program for one year only and other children

who were in the program for two years or longer. Following Behrman, Cheng, and Todd (2004), who analyze the impact of a preschool program in Bolivia, Heinrich refers to the first set of comparisons as estimates of “average” impacts of the program, and to the second set of comparisons as estimates of “marginal” effects. Heinrich argues that the marginal program effects are less likely to be biased if selection into the program is determined by student characteristics unobserved by the researcher, but the duration of program exposure is not. However, estimates of “marginal” program effects need not be free of endogeneity biases. More able students, or students who are different in hard-to-observe ways, may not only be more likely to receive Becas; they may also be likely to stay in the program for longer. Indeed, in an earlier version of the paper (Heinrich and Cabrol 2005) it appears that, after the first year, students who eventually would receive the Becas for two years had significantly lower grade repetition and significantly higher grade averages than those who would receive the program for only one year. This suggests that selection is a serious concern with this identification strategy. Also, interpretation of the estimated grade repetition effects reported by Heinrich (2007) may be problematic because there is anecdotal evidence that some teachers promoted Becas beneficiaries to ensure that they would remain eligible for the program, a point discussed by Heinrich.

In sum, there are many evaluations of CCT program effects. Broadly speaking, these evaluations can be grouped into four categories: First, there are evaluations of small-scale pilot programs, often based on random assignment. These evaluations generally have worked well, as is the case with the evaluations of the RPS and Atención a Crisis programs in Nicaragua and of the PRAF evaluation in Honduras. Random assignment appears to have equated the baseline characteristics of treatment and control groups effectively; attrition has been low. Under these circumstances, simple comparisons of means at follow-up between both groups provide credible estimates of program effects. The main limitation of these evaluations—and an important one—is the fact that the programs were small-scale pilot projects. For a variety of reasons, it is unclear how well the findings from these evaluations approximate the impacts of large, nationwide programs. Households participating in the pilot programs may be aware that they are participating in an experiment, and that may lead them to behave differently in a variety of ways—for example, they may be more likely to comply with the

conditions or be receptive to the program's social marketing; staff administering these pilots may be particularly motivated to demonstrate that the pilot programs work. As a result, these small-scale pilots may not be an accurate reflection of how well a much larger program administered by average staff would work in practice. Put differently, the evaluations of the small-scale, randomized pilot programs provide very accurate estimates of the impact of those pilots, but the external validity of the findings may be somewhat open to question.

Second, there are attempts to randomize large-scale programs for a period of time—often by randomizing the timing of the expansion of a program. That was the case with the rollout of the Oportunidades program in rural areas and the expansion of the coverage of the BDH program in Ecuador. Because both programs already had been implemented on a large scale, their evaluations face fewer questions about external validity than do the evaluations of pilot programs in Nicaragua and Honduras. Nevertheless, both evaluations also faced difficulties. Pressure to enroll all eligible beneficiaries shortened the period for which the original Oportunidades control communities did not receive transfers. Moreover, the rapid expansion of the program, even before households in the original control communities started receiving transfers, meant that control communities often literally were surrounded by other communities that were already receiving them. Under such circumstances, it is likely that households in the control communities may have expected to receive Oportunidades transfers before they actually started to receive them, which complicates interpretation of the estimated program effects. In the case of the BDH evaluation, there was substantial contamination of the control group in the sample used to estimate impacts on schooling, child labor, and consumption (see Edmonds and Schady 2008; Schady and Araujo 2008; Schady and Rosero 2008), although not in the sample that was used to estimate program effects on child health and development (Paxson and Schady 2008). The general point is that maintaining random assignment in a large-scale program is extremely difficult for political reasons. In addition, the Oportunidades data have faced other problems, including what appear to be very high levels of attrition and difficulties merging data across survey rounds, particularly for the anthropometric data.

Third, a number of evaluations have used RDD, including evaluations in Cambodia, Chile, Ecuador, Jamaica, Pakistan, and Turkey. A clear advantage of RDD is that it generally does not require program

administrators to alter the rules whereby potential beneficiaries are made eligible or ineligible for transfers. As a result, the pressure to incorporate households in the control group into the program tends to be less serious than with randomized experiments. Some contamination of the study design is not unusual (in a number of countries, some ineligible households received transfers and some eligible ones did not), but the solution to that problem is well known: estimating intent-to-treat effects on the basis of the initial assignment, or LATE estimates instrumenting program receipt with assignment. The main shortcoming of these RDD evaluations is that the estimated effect is “local,” applying only to households around the eligibility threshold. There appears to be considerable evidence of heterogeneity of CCT treatment effects (Maluccio and Flores 2005; Filmer and Schady 2008; Paxson and Schady 2008). For this reason, it is not clear that these estimates are relevant for other households whose value of the proxy means places them well below the threshold. This heterogeneity is perhaps less of a concern for the evaluations of programs in Chile and Turkey, where the CCT attempts to reach only a very small fraction of households (around 5 percent), than for evaluations in Ecuador, where the CCT attempts to make payments to fully 40 percent of households. A second potential disadvantage of RDD is that, as the value of the threshold becomes better known, households or sympathetic local program officials may attempt to manipulate scores to place some families who would normally have been ineligible for the program “just” below the eligibility threshold. Because that kind of manipulation is likely to be selective, affecting some households more than others (possibly on the basis of unobservable household characteristics), it could introduce serious biases into estimates of program effects.³

Finally, a number of evaluations have used differences-in-differences, often combined with matching, to estimate program effects. In some cases, as in the evaluation of the Familias en Acción program in Colombia, matching was done before the program began. In other cases, as with the second-generation Oportunidades evaluations discussed above, matching was done after the fact on the basis of administrative and retrospective data. This second approach adds a layer of uncertainty to the matches and the estimated program effects. Many of the more convincing evaluations using differences-in-differences also present a variety of validation exercises—for example, showing that preexisting trends are not different in the two groups of households or

communities, or showing that outcomes that one would not expect to have been affected by the treatment did not change differentially for the two groups. Attempts to triangulate the results with more than one source of data also can add to the credibility of the results.

CCTs truly have been unusual in how much and how seriously they have been evaluated. Few, if any, of these evaluations are without fault. However, the body of credible research on the impact of CCTs on a variety of outcomes is arguably without parallel in development. We conclude this appendix by discussing some areas that should receive high priority in impact evaluations (and research, more generally) on CCTs in the future.

First, much more needs to be known about the long-term effects of CCT programs in a variety of dimensions. Do CCTs lead to long-term reductions in poverty, as might be suggested by the results from Mexico that show households investing some of the transfer? Or does it take longer for households to respond to transfers by reducing their labor supply, in which case the short-term effects on consumption may overestimate the long-term effects? Do the children of families who received CCTs complete more schooling and eventually earn higher wages, or do the somewhat mixed and limited effects on learning and nutritional status translate into only small wage gains? Do families change their fertility and composition in the long term in response to transfers? Those are particularly difficult questions to answer because they involve revisiting households that received transfers many years earlier, and there is a great likelihood that they (or their children) have moved. Re-interview rates may be correspondingly low, and the possibility for substantial estimation biases is serious. Nevertheless, the returns to carefully constructing and studying long-term panels of this sort should be very high, and doing so is a priority for future evaluation work.

Second, although much is known about the effect of CCTs on some outcomes—such as consumption levels, school enrollment, health service utilization—much less is known about a variety of other important outcomes: Under what circumstances do CCTs affect learning outcomes, and how does that interact with the quality of the supply of schooling? Can CCTs be used to seek changes in sexual behaviors, as has been proposed in discussions about how best to limit the transmission of HIV/AIDS? In many countries, CCTs make payments through the banking system, and in some cases a fraction of payments is deposited directly into a savings account for a household. Have those payment

methods resulted in positive spillover effects in households' ability to access and use financial services?

Third, more needs to be done to unpack the CCT effects on outcomes. Are the changes that are observed a result of the cash, the conditions, the social marketing that generally accompanies the program, or the fact that transfers are made to women? How much, and for what outcomes, does the magnitude of the transfer matter? Understanding the answers to these and related questions is important for the design of efficient CCT programs in the future.