Evaluating the Impact of Health Programmes *

Justine Burns, Malcolm Keswell, and Rebecca Thornton

June 22, 2009

Abstract

This paper has two broad objectives. The first objective is broadly methodological and deals with some of the more pertinent estimation issues one should be aware of when studying the impact of health status on economic outcomes. We discuss some alternatives for constructing counterfactuals when designing health program evaluations such as randomization, matching and instrumental variables. Our second objective is to present a review of the existing evidence on the impact of health interventions on individual welfare.

1 Introduction

There are a number of mechanisms through which health can affect productivity (Strauss and Thomas 1998; Bloom et al. 2004; Weil 2006). Improved health can have a direct effect by increasing the productivity of healthy workers, as well as an indirect effect by affecting savings and investment (Muney and Jayachandran 2008; Yaari 1965). Most of the research on the indirect effects involves macroeconomic studies of aggregate changes in life-expectancy on savings, investment, and GDP, whereas the bulk of research on the direct effects of health policy on the other hand is largely micro-focused. Moreover, this evidence is limited in scope and generalizability partly because evaluating the impact of health interventions on individual welfare and productivity involves time lags between the intervention, often made during infancy or childhoood, and welfare outcomes of interest, such as employment status, usually observed in adulthood.

A further difficulty concerns the reliability of the evidence which is available. While the connection between income levels and health status has long been recognized as crucial for economic growth, the *causal* relationship between income and health is harder to establish. Plausibly, many economic outcomes of interest (productivity for instance) and an individual's health status are simultaneously determined. Thus, establishing the causal effects of health interventions on economic outcomes requires that special attention be paid to identification strategies.

This paper has two broad objectives. The first objective is broadly methodological and deals with some of the more pertinent estimation issues one should be aware of when studying the impact of health status on economic outcomes. If the analyst wishes to estimate the direction and magnitude of the impact of a particular program or policy intervention on beneficiaries of the intervention, it is necessary to assess the welfare outcomes of program beneficiaries against some type of counterfactual. The paper begins in section 2 by discussing why the need for counterfactuals arise in the first instance. Sections 3-5 then presents some alternatives for constructing counterfactuals. We begin in section 3 with the "gold standard" of randomization and then move to quasi-experimental matching approaches in section 4. Although methods like propensity score matching were designed primarily as a solution to problems of identification in observational studies, the method is most useful when used in a complimentary

*University of Cape Town, University of Stellenbosch, and University of Michigan, respectively. The authors would like to thank Duncan Thomas for useful comments. We also thank Robert Eastwood, Peter Glick, German Mwabu as well as participants of the AERC Collaborative Research workshop on Health, Economic Growth and Poverty Reduction held in Accra, Ghana, April 20-22, 2009. Comments can be sent to keswell@sun.ac.za

fashion to imperfect experimental or quasi-experimental evaluations. We therefore treat this approach separately from other well known non-experimental methods in order to motivate its use as a solution to the problems of "internal validity" that often compromise otherwise well designed randomized experiments. Section 5 deals with other non-experimental methods such as IVs, double-differencing, and regression discontinuity. Even though the structure of these estimators are well known, we include a brief outline of them here since these approaches are potentially quite useful in contexts where the assumptions underlying randomization or matching do not hold.

Our second objective in this paper is to present a review of the existing evidence on the impact of health interventions on individual welfare. The task of establishing the external validity of health interventions has to be confronted irrespective of the the underlying methodology used. Section 6 discusses some of the issues pertinent to external validity. This is followed in section 7 by a review of how the range of available methods outlined in the preceding sections has been employed in different settings and what is known about the impact of health interventions on individual productivity. Section 8 concludes the paper with an assessment of where opportunities for further study might lie.

2 Identification of Program Impact

At the heart of the evaluation problem is the desire to know whether a particular program or policy intervention has made a difference in the lives of those individuals or communities affected by it, and if so, what the magnitude of this impact has been. In order to make this kind of judgment, it is necessary to assess the welfare outcomes of program beneficiaries against the counterfactual, namely:

(1) How would those people who benefited from an intervention have fared in the absence of the intervention?

(2) How would those people who were not beneficiaries of an intervention have fared in the presence of the intervention?

More specifically, consider the following hypothetical problem: let y_{1i} refer to average outcomes across all households in a given "community" i if the community has received some health intervention, and let y_{0i} refer to average outcomes across all households in this same community i where no intervention took place. We are interested in what difference the receipt of treatment has made to the average outcomes of households in this community; i.e., the difference $y_{1i} - y_{0i}$. The problem is that we will never have a given community both with and without treatment at the same time.

Imagine that we have data on many communities, where some communities have received treatment and others not. If we had this type of data, we could approximate $y_{1i} - y_{0i}$ with $\delta = E[y_{1i}|T = 1] - E[y_{0i}|T = 0]$. This estimate, known as the single-difference estimate, is confounded by the presence of selection bias. To see why this is so, imagine that we could observe the counterfactual $E[y_{0i}|T = 1]$ - i.e., we can compute the average outcome of interest across all households in non-beneficiary communities in an alternative state of the world in which these communities were part of the beneficiary group. Now add and subtract this conditional mean from the one used previously to give:

$$\delta = \underbrace{E[y_{1i}|T = 1] - E[y_{0i}|T = 1]}_{\textit{treatment effect}} - \underbrace{E[y_{0i}|T = 0] + E[y_{0i}|T = 1]}_{\textit{selection bias}}$$

The first term in this expression is what we want to try to isolate: the effect of the intervention on those that received it. We call this the *treatment effect*, or more precisely, the average treatment effect on the treated (ATT). The last two terms together constitute selection bias and picks up systematic unobservable differences between treatment and control households.

The inability to separate out the treatment effect from selection bias is the identification problem we are confronted with if we simply regress the outcome variable on the treatment dummy. In this type of model, selection bias arises because the treatment variable is correlated with unobserved characteristics. A natural solution therefore would seem to use proxy variables of these unobservables in the outcome regression. Characterizing the problem in this way suggests that many of the standard techniques that deal with endogenous regressors can be used as potential solutions. However finding plausible ways of extracting exogenous variation in treatment status in non-experimental settings often rests on *a priori* reasoning that might be contestable or quite specific to some sub-population of the sample affected to participate in the treatment as a result of the exogenous variable/s such models rely on. We return to these non-experimental techniques in section 5.

3 Randomization

3.1 Motivation

Randomizing assignment to the treatment group (from a sample of potential participants in a program) theoretically eliminates the confound from selection bias in estimates of mean impact. Randomization involves a lottery process. Individuals from some well-defined population are randomly selected into either the treatment group or the control group. An advantage of this process is that it removes potential differences that could exist between the two groups for which social scientists cannot control, or for which they find it difficult to control, such as ability, work ethic, psychological disposition and so forth. Importantly, the observed and unobserved attributes of individuals in the treatment and control groups prior to the intervention must be independent of assignment to the treatment or control group. If this condition does not hold, this will result in differences in mean outcomes ex-post that would falsely be attributed to the intervention. However, when randomization is successfully implemented, the treatment effect is unconfounded since treatment status is randomly allocated. [1]

3.2 Internal Validity

Internal validity examines whether the specific design of an evaluation study generates reliable estimates of counterfactual outcomes in a specific context (Ravallion, 2008). Despite the simplicity involved in randomization, there may be a number of reasons why evaluation estimates derived from this method lack internal validity. Bias may be introduced owing to selective compliance with or attrition from the randomly assigned status. This occurs when individuals assigned to the control group take deliberate action in order to attain the benefits of treatment. For example, if an intervention is regionally based, or school based, individuals in the control group may actively move schools or locations in order to be counted part of the treatment group. Differential attrition in the treatment and control groups will also lead to biased estimates. Since individuals who benefit from an intervention may be less likely to drop out of an evaluation study than those who do not (control, group), this can result in differential attrition between control and treatment groups. On the other hand, individuals randomly assigned to the treatment group may choose not to comply with the treatment (for example, they may neglect to take their pills, they may choose not to collect a social grant or to utilize a voucher and so on), or, because they feel healthier, may stop complying with the requirements of the

[1] It is important bear in mind that random assignment in general does not "eliminate" selection bias because participation is generally not open to all individuals in a population. Random assignment in such instances only apply to a subset of the population. Under this sort of scenario, Heckman and Smith (1996) show that estimates of mean impact will be unbiased because the effect of randomization is to balance the bias between the treated and not-treated, so that the bias is differenced out when computing δ. However, when interest lies in some other measure of central tendency, or higher order moments of the distribution of impacts, then randomisation alone does not remove the effect of selection bias on estimates of impact. In this instance, combining social experimentation with non-experimental methods of dealing with selection bias is a more appropriate strategy.

programme. [2] Institutional or political factors that delay randomized assignment may also promote selective attrition (Ravallion, 2008; Heckman and Smith, 1995).

In each case, this leads to a difference between the actual allocation and the intended allocation, and to the extent that this is not controlled for, will result in biased estimates of impact. Program design should try to anticipate this, and put processes in place to minimize attrition that might occur. In the Balsakhi program, for example, Banerjee et al (2007) ensured that when students did not appear at schools, the data collectors went to their homes in order to ensure that they collected the data and could track the individuals. This resulted in lower attrition from the study.

However, when attrition or selective compliance is present, researchers typically deal with these kinds of problems through intention to treat models (Imbens and Angrist, 1994), whereby the differences in outcomes for treatment and control groups (as per the original assignment) are scaled up by dividing the difference in outcomes by the difference in the probability of actually receiving treatment in the two groups. This gives an estimate of the average treatment effect for those induced to participate by randomization (Ravallion, 1995). Importantly, this differs from the average treatment effect in the population as a whole, where this kind of selective compliance does not occur. Rather, intention to treat models account for the fact that individuals who anticipate benefiting from a programme may be the most likely to take advantage of it. Arguably, these may be precisely the kinds of individuals that policy makers are most interested in.

A second important consideration in critically evaluating impact estimates is the presence of externalities generated by the programme or intervention itself. Externalities may plague the credibility of impact evaluation estimates if policy makers or aid agencies reallocate their spending priorities to compensate some communities or individuals for their non-participation in the intervention. This is difficult to know but vitally important to keep track of, since to the extent that such re-allocation of spending priorities may occur, this will influence the magnitude of the impact estimates. In addition, to the extent that an intervention confers positive externalities on individuals outside of the treatment group, failure to account for these externalities may lead to an under-estimate of the intervention impact. For example, in the Miguel and Kremer (2004) study of mass deworming programmes in Kenya, they argue that a randomized intervention targeted at the individual level in which some children within the same school were treated while others were not would result in a serous underestimate of treatment effects, since the control children would enjoy reduced disease transmission by virtue of being in contact with treated children. [3] Hence, they chose to randomize at the school level.

The presence of externalities generated by an intervention thus points to the need for careful thought to be given about the level at which randomization should occur as well as the need to collect detailed information to control for these possible spillovers in arriving at credible impact estimates. For example, despite randomizing at the school level, Miguel and Kremer (2004) still find evidence of positive spillovers in the deworming project in that children attending neighbouring non-treatment schools also enjoy reduced incidence of intestinal worms through reduced transmission of disease when interacting with children in treatment schools. Since detailed spatial information about the distance between schools was collected as part of the evaluation survey, Miguel and Kremer (2004) are able to utilize this data to control for these spillovers.

Thus, the choice of observational unit should reflect likely spillover effects (Ravallion, 2005).

[2] Not only does partial compliance by individuals holds implications for the credibility of the impact estimates, but it also holds implications for sampling. For example if there were to be approximately an 80% level of compliance by the treated group, then the entire sample would have to be approximately 50% larger in order to get commensurable effects relative to a group that had 100% compliance. Thus, in designing a study, one has to weigh up the costs and benefits between a study that requires high compliance rates but lower sample sizes relative to a program with lower compliance levels but requiring larger sample sizes in order to have the same level of power from the results. It may happen that a more comprehensively considered program with higher (predicted) compliance levels might in fact be less expensive to implement than a project with lower compliance levels.

[3] Similarly, failure to account for negative externalities imposed by an intervention would result in an over-estimate of the programmes benefits.

Once this decision has been made, it is important to ensure that the sample size is as large as possible *at the level at which the randomization has occurred*. For example, if randomization has occurred at a group level (e.g. school), it is important to have as large a sample of schools as possible. It is not the case that increasing the sample size of individuals within a school gives more power to the evaluation. Rather, at the margin, the evaluator will gain more information from the addition of a cluster or group (in this case, school) than they will through the addition of a new individual to an already existing group. This is because individuals within a given community or school could all negatively (positively) be affected by some shock, with the consequence that their individual outcomes could be correlated as a result. The addition of new groups helps to cater for the possibility of intra-group shocks that could affect a number of individuals in a significant manner.

Randomization bias may also plague impact assessment estimates (Heckman and Smith, 1995). This arises if there is a significant difference in the kinds of individuals who would choose to participate in a programme compared to those individuals who are randomly assigned to participate in a programme. Consequently, the intervention that is evaluated is different than the intervention that is implemented in practice, making it difficult to know what to make of the estimates (Ravallion, 2008).

Finally, randomized evaluations may confront ethical objections that the method of randomization by its very nature will exclude some individuals that could potentially benefit from the intervention, and will include some individuals in the treatment group that do not need the intervention as much. These objections may be combined with political concerns over service delivery to the electorate. While ethical objections should be addressed, the short-term loss of being excluded from the benefits of an intervention may be small in relation to the long-term benefits once a programme that has been properly evaluated is implemented and scaled up (Ravallion, 2008). Moreover, randomization may be the fairest method of allocating scarce resources, when it is simply not possible to deliver a programme to everyone. For example, the PROGRESSA programme, launched in 1998, provided social grants to households conditional on the enrollment and attendance of children at school, and their participation in preventative health care programmes. Since budget constraints made it impossible to reach all of the 50 000 potential beneficiary communities, the Mexican government made a conscious choice to begin with a pilot project of 506 communities, of which, half were randomly selected to receive the grants while the others did not (Gertler and Boyce, 2001). The project was later scaled up considerably.

4 Propensity Score Matching

When randomization is not practically or politically feasible, or when the results from a randomized intervention are not internally valid, more appropriate counterfactuals can be found by matching treatment households to control households. The ideal approach is to match treated household to control households directly on their characteristics (see for example Angrist (1998)) but this approach is often not practical when some of the more important variables we wish to condition on are continuous, or when the number of covariates we wish to match on is of large dimension.

Propensity score matching is a useful alternative to exact matching. The idea here is to match not on the multidimensional vector of covariates but rather on a scalar index (propensity score) of predicted probabilities computed from a regression where the outcome variable is a binary indicator of treatment (see Rosenbaum and Rubin, 1983; Heckman and Robb, 1985; Heckman, LaLonde and Smith, 1999).[4]

Formally, if we let $\mathbf{x}$ be a vector of pre-treatment variables, then we can define the propensity

[4]Hirano and Imbens (2004) provide a generalization of this approach to the case where treatment is not binary but continuous. This approach is potentially quite useful for many health interventions where one would be interested in not only the effect of treatment but the dosage of treatment among the treated (e.g., ARV treatment).

score as the conditional probability of receiving the treatment T, given $\mathbf{x}$

$$p(\mathbf{x}) = \Pr[T = 1|\mathbf{x}] = E[T|\mathbf{x}]$$

For the purposes of the analysis to follow, two key results first introduced by Rosenbaum and Rubin (1983) are noteworthy:

Lemma 1 *(Balance): If $p(\mathbf{x})$ is the propensity score, then $\mathbf{x} \perp T|p(\mathbf{x})$. Stated differently, the distribution of the covariates for treatment and control is the same once we condition on the propensity score: $F(\mathbf{x}|T = 1, P(\mathbf{x})) = F(\mathbf{x}|T = 0, P(\mathbf{x}))$*

Lemma 2 *(Ignorability): If there is no omitted variable bias once $\mathbf{x}$ is controlled for, then assignment to treatment is unconfounded given the propensity score.*

The first result says that once we condition on the propensity score, assignment to the treatment group is random. In other words, for two identical propensity scores, there should be no statistically significant differences in the associated $\mathbf{x}$ vector, independent of how these scores are distributed between the treatment group and the control group. This property must be met if we are to move forward after computing the propensity score.

The second result says that selection into treatment depends only on what we can observe, i.e., $\mathbf{x}$. In other words, while the propensity score balances the data (i.e., removes the influence of the observables on assignment to the treatment group), it also assumes no confounding on the basis of unobservables. Whether or not this assumption is plausible rests on whether the specification of the propensity score regression accurately reflects the key factors that might influence the process of treatment assignment.

A key challenge in getting the right specification for the propensity score is making sure that the balancing property is satisfied. Practically speaking, the balancing property of the propensity score implies that we need to make sure that the control group and beneficiary group are not statistically different from each other, once we've conditioned on $\mathbf{x}$. This requires that we check that $\mathrm{E}(p(\mathbf{x})|T = 1) = \mathrm{E}(p(\mathbf{x})|T = 0)$ as well as that $\mathbf{x} \perp T_i|p(\mathbf{x})$. One way to accomplish this test is to aggregate the estimated propensity score $\hat{p}(\mathbf{x})$, into mutually exclusive intervals (blocks) over its distribution and then check that the average propensity score within each block is uncorrelated with treatment assignment. Then using this same procedure, we can also check that each covariate is uncorrelated with treatment assignment within each block.

This obviously means that the balancing property can only be tested in a proximate sense. Dehejia and Wahba (1999, 2002) and the associated STATA implementation of Becker and Ichino (2002) is one very widely used algorithm for testing that the estimated propensity score balances the covariates of treatment status.[5]

4.1 Stratification

If lemma 1 (the balance property) is satisfied, a somewhat natural way to compute the treatment effect then is to take the difference between the mean outcomes of the treated and control groups within each stratum of the propensity for which the covariates are balanced, and weight each of these differences by the distribution of the treated households across the strata in order to get the average treatment effect for the treated households. Formally, let i denote the ith treated household; let j denote the jth control household, and let b denote the bth block (stratum). Then a block-specific treatment effect is

$$\mathrm{ATT}_b = (N_{b,1})^{-1} \sum_{i \in I(b)} y_{1i} - (N_{b,0})^{-1} \sum_{j \in I(b)} y_{0j}$$

[5]The approach works by arbitrarily grouping the data by blocks (intervals) of the propensity score, where initially the scores within a block are quite similar. An equality of means test between treatment and control observations is performed for each of the regressors contained in $\mathbf{x}$. If there are no statistically significant differences between treatment and control for each of the covariates in the propensity score regression, then the regressors are balanced. If a particular regressor is unbalanced for a particular block, then that block is split into further groups and the test is conducted again. This iterative process continues until all the regressors are balanced or the test fails.

where I_b is the set of households in the bth block, and where $N_{b,1}$ and $N_{b,0}$ are the subsets within I_b that fall either into the treatment group or control group. To get the average treatment effect by the method of stratification, we simply weight each of these block-specific treatment effects by the proportion of treated households falling into each block, and then sum the resulting weighted block-specific treatment effects over all strata Thus,

$$ATT^{Strat} = \sum_{b=1}^{6} \text{ATT}_b \times \frac{\sum_{i \in I_b} D_i}{\sum D_i}$$

4.2 Nearest-Neighbor Matching

One very attractive feature of matching on the propensity score is that we need not assume a specific functional form for the underlying distribution of the treatment effect since the (average) treatment effect can be computed semi-parametrically.

One such approach is to match each treated household to the control household that most closely resembles it. There are various ways in which this can be done, one of which is to match directly on $\mathbf{x}$, but given Lemma 1, a better way to proceed is to match on the propensity score. Since $p(\mathbf{x})$ is a scalar index, this method has the advantage of permitting a greater number of matches than matching directly on $\mathbf{x}$ would allow.

Formally, we can define the set of potential control group matches (based on the propensity score) for the ith household in the treatment group with characteristics $\mathbf{x}_i$ as

$$A_i(p(\mathbf{x})) = \{p_j | \min_j |p_i - p_j|\}$$

The matching set will usually contain more than one control group household that could potentially feature in the calculation of the average treatment effect. The most restrictive form of the nearest neighbor method would select a unique control group household for every treatment group household on the basis of computing the absolute value of the difference in propensity scores for every pairwise match considered, and then selecting as a match the jth household with the smallest absolute difference in propensity scores. Alternatively, all observations in the set $A_i(p(\mathbf{x}))$ could be matched against household i. In this case, a differential weight would be applied to each match falling into the matching set. The average treatment effect would then be computed as follows:

$$\text{ATT}^{NN} = (N_1)^{-1} \sum_{i \in \{T=1\}} (y_{1i} - \Sigma_j \omega(i,j) y_{0j})$$

where j is an element of $A_i(p(\mathbf{x}))$ and $\omega(i,j)$ is the weight given to j. For the restrictive one-to-one match mentioned above, we would then have $\omega(i,j) = 1$ when $j \in A_i(p(\mathbf{x}))$, and $\omega(i,j) = 0$ when $j \ni A_i(p(\mathbf{x}))$.

4.3 Kernel Matching

A closely related approach to nearest-neighbour matching is to match non-parametrically using a kernel function. In this instance our formula for the ATT is as above, but the weight given to the jth control group household in matching it to the ith treated household is determined as follows

$$\omega(i,j) = \frac{K(p(\mathbf{x}_j) - p(\mathbf{x}_i))}{\sum_{j=1}^{N_{0j}} K(p(\mathbf{x}_j) - p(\mathbf{x}_i))}$$

$$K = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{p(\mathbf{x})^2}{2\sigma^2}}$$

where K is the Gaussian (normal) kernel. This method has the benefit of using the entire sample for each prediction with decreasing weights for more distant observations, where the rate of decline of these weights is determined by σ. In principle, ω could be determined in other ways (e.g., tri-cubic, caliper etc.)

4.4 Pipeline Matching

Delays in implementation of a programme may also facilitate the formation of a comparison group. In these studies, usually termed pipeline studies, the control group comprises those individuals who have applied for a programme but not yet received it (for example, see Chase, 2002; Galasso and Ravallion 2004). For example, in PROGRESSA, one third of eligible participants didn't receive a transfer for 18 months during which they formed control group. Thereafter, they were phased into the programme. Similarly, in a Kenyan deworming programme studied by Miguel and Kremer (2004), while 75 schools were chosen to participate, in the first year of the study, only 25 schools were treated, while the other 50 schools formed the control group. In year two, a further 25 schools were phased into treatment and by the third year, all 75 schools were receiving the treatment. The advantage of this method is that it deals with selection bias even on unobservable characteristics to some extent, since the successful applicants not yet receiving treatment will be very similar in most respects to beneficiaries of the programme. A key assumption though in pipeline studies is that the timing of treatment be random given application.

4.5 Comparison with Randomization

The evidence on whether PSM methods and RE methods produce the same results is somewhat mixed. Agodini and Dynarski (2004) find no consistent evidence that PSM can replicate RE results of school dropouts programmes in the US. In contrast, work by Heckman et al (1997a, 1998) and Diaz and Handa (2004) suggests that PSM works well as long as the survey instrument used for measuring outcomes is identical for treatment and control participants (Diaz and Handa, 2004; Heckman et al, 1997a, 1998). A recent study by Diaz and Handa (2007) shows that with the collection of a large number of observables, propensity score matching can approximate RE results.

Hence, the success of PSM hinges critically on the data available, as well as the variables used for matching. The key challenge for PSM methods is to identify all potentially relevant covariates and differences between treatment and control groups. If treatment is assigned on the basis of an unobservable trait, then the estimates obtained will be biased.

The choice of variables should be based on some theoretical reasoning and/or facts about the intervention and its context, as well as any relevant socio-economic and political considerations. In this regard, additional qualitative work may be useful (Jalan and Ravallion, 2003b; Godtland et al, 2004). Ex-post, it is important to test for differences in the covariates between treatment and comparison groups to ensure that covariate balance can be achieved (Smith and Todd, 2005a). Importantly, then, PSM estimates will be limited to a matched sample and not the full sample. However, matched sample estimates tend to be less biased and more robust to misspecification error (Rubin and Thomas, 2000).

5 Other Non-Experimental Methods

Two potential problems remain unexplored with the propensity score approach. The first, discussed already, concerns the possibility of remaining omitted variable biases. The propensity score regression uses proxies for the unobserved/omitted variables under the assumption that the omitted variables are redundant in explaining treatment assignment once their proxies are accounted for. Matching methods are of little use when such proxies do not exist. Observational studies – even those based on quasi-experimental designs – with this type of problem are said to exhibit selection on unobservables. This section deals with three widely used alternatives to randomization and/or matching when we do not observe the full set of variables influencing treatment status: instrumental variable estimation, regression discontinuity approaches and double-differencing.

5.1 Instrumental Variables

A key feature of this framework is that unobservables don't bias the treatment effect as long as an instrumental variable can be found that is non-trivially related to treatment assignment but is uncorrelated with other variables which are omitted from the outcome equation of interest. Thus if we are dealing with a "broken" experimental design premised on randomizing treatment, and we have a concern that not all of the important variables predicting treatment can be observed given the survey instrument employed, IVs might offer a useful alternative

5.1.1 Wald Estimator: Binary Treatment-Binary IV

Consider once again the single difference estimator introduced earlier. A regression equivalent of that estimator is:

$$y_{ij} = \alpha + \delta T_{ij} + u_{ij}$$

where T is our treatment dummy; y is our outcome variable; and i, j indexes villages/PSUs and households respectively.

A simple alternative to this naive approach is the Wald estimator (Angrist, 1990). This estimator is a special case of the local average treatment estimator or LATE (Imbens and Angrist, 1994) where we instrument T with a binary variable.

Let this variable be denoted as P_{ij}. Then as long as P_{ij} does not perfectly predict T_{ij}, it can be shown that δ is simply equal to the ratio of the difference in means for y (between households with $P = 1$ and $P = 0$) to the difference in means for T (between households with $P = 1$ and $P = 0$). For the most parsimonious case given above where we use a single IV, the IV estimate of the slope can be written as

$$\begin{aligned}
\hat{\delta} &= \frac{(\sum_{i=1}^{N}(P_{ij} - \bar{P})(y_{ij} - \bar{y}))}{(\sum_{i=1}^{N}(P_{ij} - \bar{P})(T_{ij} - \bar{T}))} \\
&= \frac{(\sum_{i=1}^{N} P_{ij}(y_{ij} - \bar{y}))}{(\sum_{i=1}^{N} P_{ij}(T_{ij} - \bar{T}))} \\
&= \frac{\bar{y_1} - \bar{y_0}}{\bar{T_1} - \bar{T_0}}
\end{aligned}$$

The complete derivation is given in appendix A1. The standard choice for an IV in this context is to use some indicator of eligibility.

5.1.2 IV Estimator: Continuos Treatment-Binary IV

Often the rules governing participation in a health program might invalidate the use of eligibility as an IV. For example, many health interventions are deliberately targeted to poorer segments of a population. If the outcome of interest is some type of welfare metric (say consumption), then a model such as the one above will have an implausible exclusion restriction since a variable such as P is likely to covary with y (the outcome variable of interest). However, exogenous variation can sometimes by extracted through innovative use of prior information about rollout or other features of program implementation. For example, if the health programme is targeted to poor villages but at a centralised location such as a clinic, then spatial information such as the distance from sampled households to the clinic could in principal be used to construct a model with more plausible exclusion restrictions.

How exactly might this might work? Let D refer to a measure of distance such as the one just discussed and let P be defined as in the previous model. Now let's imagine we are interested in estimating the impact of some health intervention which is best understood as a "dose".[6] As before, denote treatment (this time assumed continuous) as T. Plausibly, D, P

[6]For example, the treatment for iron deficiency anemia ranges from 3-12 months and then has to be complemented for the rest of the patient's life by a more iron-enriched diet than was the case prior to the onset of treatment.

and T all belong in the structural model. Individuals that live on the fringes of the village boundary might be relatively more cut-off from the centre of economic activity so that their spatial location covaries with their outcomes. Likewise, if the program is means-tested and a baseline survey is not available, then P might also belong in the structural model. However, there is no obvious reason to expect that the interaction between D and P belongs in the structural equation. Thus, a more plausible data sampling process might be:

$$y_{ij} = \alpha + \beta D_i + \gamma P_{ij} + \delta T_{ij} + \underbrace{\{\eta(D_i \times P_{ij}) + v_i + \epsilon_{ij}\}}_{\text{composite error}}$$

where $i = 1, \ldots, N$ indexes villages, $j = 1, \ldots, M^i$ indexes the M^i sampled households in village i, and v_i and ϵ_{ij} are project and household-specific error terms respectively. As before, y_{ij} is a measure of consumption. Under this type of data sampling process, if $(D_i \times P_{ij})$ is to be considered a valid IV, we must assume $\eta = 0$, otherwise it could be the case that $cov((D_i \times P_{ij}), u_{ij}) \neq 0$, where $u_{ij} = v_{ij} + \epsilon_{ij}$. On the other hand, if we assume $\eta = 0$, we can then construct a Wald type of estimator using $D_i \times P_{ij}$ as an IV for T_{ij}. We show in appendix A.2 that this IV turns out to resemble a Wald type of estimator that consistently estimates the average treatment effect. Formally,

$$\begin{aligned} \tilde{\delta}_{IV} &= \frac{\Delta_{y|D,P}}{\Delta_{T|D,P}} \\ &\xrightarrow{p} \delta + \frac{\eta}{\Delta_{T|D,P}} \end{aligned}$$

where $\Delta_{y|D,P}$ and $\Delta_{T|D,P}$ are defined explicitly in appendix A.2.

5.2 Regression Discontinuity Design

With this approach, researchers take advantage of extant discontinuities that occur as the result of the policy itself to try and identify the impact of the programme. Discontinuities may be generated by programme eligibility criteria, thereby making it possible to identify impact by comparing differences in the mean outcomes for individuals on either side of the critical cutoff point determining eligibility. For example, in Israel, if a class size exceeds forty students, a second class is introduced to cater for this increase in student numbers. Hence there is a discontinuity between the levels of 40 students and 41 students in a grade respectively, or 80 and 81, and so forth. This allows researchers to observe differences immediately above and immediately below the threshold level (Angrist and Lavy, 1999). Similar work has been done in South Africa with respect to welfare responses resulting from access to the state Old Age Pension which has an age eligibility criteria. Health outcomes for children, girls in particular, are shown to be significantly better in households that have pension-eligible members (aged 60 and above) as opposed to households that do not (with household members aged 55-59) (Duflo, 2001). As with PSM, regression discontinuity only gives the mean impact for a selected sample of participants, namely those in the neighbourhood of the cutoff point.

A key identifying assumption is that there is no discontinuity in counterfactual outcomes at the point of discontinuity. This is made difficult if the discontinuity is generated by an eligibility requirement that is geographically specific or one that coincides with political jurisdiction, since this in itself might suggest pre-existing differences in the outcomes of interest. Moreover, it is assumed that the evaluator knows the eligibility requirements for participation and that these can be verified and measured. Where eligibility is based on some criteria such as age, this is relatively easy to do. However, if eligibility for a programme relies on a means-test, verification of pre-intervention status becomes more difficult since incomes are only observed ex-post in a cross-sectional survey. In these instances, a baseline survey helps to control for pre-intervention differences.

Buddelmeyer and Skoufias (2004) use cutoffs in PROGRESSAs eligibility rules to measure impacts of the program and find that discontinuity design gives a good approximation for almost all outcome indicators when compared to estimates obtained through randomization.

5.3 Difference-in-difference Analysis

This method contrasts the growth in the variable of interest between a treatment group and a relevant control group. This approach requires that participants be tracked over time, beginning with a pre-intervention baseline survey, followed up by subsequent surveys of participants and non-participants. The estimate of treatment impact is given by the difference in outcomes for individuals before and after the intervention, and then the difference between that mean difference for participants and non-participants. The key assumption underlying this method is that selection bias is invariant over time.

Difference-in-difference estimates may be appropriate where an argument can be made that outcomes would not have been different over time in regions that received the programme compared to those that did not, had the programme not been introduced. If this case can be made, then one can compare differences in the growth of the variable of interest between programme and non-programme areas. However, this approach requires long-standing time-series data in order to ensure that the groups are as similar as possible, and to project that they would have behaved similarly without the presence of the treatment. Moreover, one must be certain that no other programmes were introduced concurrently, and that a region may has not been affected by a time persistent shock that may manifest as a treatment effect (Bertrand, Duflo and Mullainathan, 2003).

A further benefit of the difference-in-difference approach is that it can be used to address bias in the estimates obtained from a randomized evaluation study if there has been selective compliance or attrition, and they minimize bias that might arise due to measurement error. Even so, there can be additional biases to the standard errors from using this method. At the time of the baseline survey, it may not be apparent which individuals will participate in the programme and which will not, and hence, the researcher must make their best guess when drawing a random sample for the baseline survey. This may hold implications for sample representativeness ex-post, so to minimize this source of possible bias, the researcher should use any information they have about the details and context of the proposed programme to help guide their sampling choices, and then over-sample from the likely participant group, in order to ensure a good comparison group. Secondly, the assumption that selection bias is unchanging over time may also be problematic, especially if changes in outcome variables due to the intervention are a function of initial conditions which influenced progamme assignment to begin with (Ravallion, 2008; Jalan and Ravallion, 1998). In other words, if poor regions are targeted for intervention because of their poverty status, and if treatment impact depends on the level of poverty, this will bias impact estimates. Consequently, the researcher needs to control for initial conditions in deriving their impact estimates (Ravallion, 2008).

Since difference-in-difference estimates require longitudinal data, the researcher will have to consider the trade-off between relying on a single survey estimate and utilizing PSM to find a comparable control group, as opposed to incurring the cost of tracking individuals over time in order to be able to utilize difference-in-difference estimators. Ravallion (2008) argues that such a decision should be made based on how much is known ex ante about programme placement. If a single cross-sectional survey is able to provide comprehensive data in this regard, then this may be a more feasible alternative that collecting longitudinal data.

The difference-in-difference approach has been successfully used to provide estimates of impact in a number of interventions. For example, Thomas et al (2003) show that iron supplementation amongst iron-deficient individuals, males in particular, yields improved economic productivity, as well as improved psycho-social and physical health outcomes. Galiani et al (2005) use difference-in-difference estimates to show that the privatization of water services in Argentina reduced child mortality.

6 External validity

The non-experimental methods reviewed above may assist in dealing with concerns that arise over the internal validity of impact estimates based on randomization alone. However, in addi-

tion to concerns about the internal validity of impact evaluation estimates, concerns may arise about external validity, and these concerns arise irrespective of the evaluation methodology adopted. External validity concerns the extent to which results derived from a specific evaluation study can be generalized to other contexts, and whether lessons can be taken away for the future. In particular, can one expect the same outcomes once the programme is scaled up, and can policy makers base their own decisions on the introduction of new policies and programmes on the experience of previous interventions in other contexts?

There are a number of reasons why the answer to such questions may be no. The first relates to the fact that estimates for an evaluation study will only produce partial equilibrium effects, and these may be different from general equilibrium effects (Heckman, Lochman and Taber, 1998). In other words, the scale of the programme may affect estimated treatment effects. If an intervention study is limited to a specific region or area, or if participation is means-tested in some way, then taking that same programme and replicating it at the national level may lead to very different results. This concern will be even more justified if the success of the intervention is tied to the existence of specific institutions. For example, if a specific intervention rests on the activities of a local NGO, then the impact when the programme is scaled up to the national level may be quite different (Duflo and Kremer, 2005). Moreover, scaling a programme up to the national level may alter the way that markets work, thereby affecting the operation of the programme itself. For example, a wage subsidy programme tested at a local level may show promising results, but when this same intervention is scaled nationally, it may alter the operation of labour markets, and produce a different outcome (Ravallion, 2008).

Scaling up may also fail if the socio-economic composition of local participants differs from the national demographic profile. Randomised interventions tested at a local level tend to under-estimate how pro-poor a programme will be, since initial benefits of an intervention tend to be captured by local elites (Lanjouw and Ravallion, 1999). However, as the programme is scaled up, the incidence of benefits tends to become more pro-poor as the benefits are extended to greater numbers of individuals (Ravallion, 2004a).

An obvious difficulty of thinking about how generalisable the results from a specific intervention are is that the counterfactual is typically posed in terms of how participants would have fared in the absence of the intervention. However, policymakers are typically trying to choose amongst alternative programmes, not between whether to intervene or not. Hence, while a specific intervention may fare well against a counterfactual of no intervention, it need not be the case that the same intervention would fare as well when compared against a different policy option.

Concerns over external validity may be ameliorated to the extent that interventions are replicated in different settings and at different scales (Duflo and Kremer, 1995; Duflo, 2003). The results from these replication studies provide evidence on the extent to which results can be generalized. Since different contexts will require adaptations and changes to programmes, the robustness of the programme or intervention is revealed by the extent to which it survives these changes. Moreover having multiple estimates of programme estimates in different settings gives some sense of how generalisable the results really are. For example, the findings from the mass deworming intervention in Kenya reported by Miguel and Kremer (2004) were largely vindicated in a study in India, reported by Bobonis, Miguel and Sharma (2002), despite the fact that the Indian programme was modified to include iron supplementation.

Concerns also arise over the length of the evaluation period. To the extent that the evaluation period coincides with the project period, any impacts that continue after the completion of the project or only materialize in the long run will fail to be captured in the evaluation. In short, there may be significant lags in outcome responses to an intervention. With health care programs as an example, the interventions will only have effects once better health care outcomes (BMI, height-weight ratios, incidence of absenteeism, etc) can be definitively measured. Thus the length of the program hinges on what the outcome variable of concern is, and whether there is sufficient time in the program for there to be a change in the outcome variable. One solution to this concern is to design an intervention to include the tracking of participants for a significant period of time, perhaps even after the programme or intervention has ended. Of

course, this is costly, but the advantage is that it yields a lot of data that allows one to unpack the causal mechanisms explaining changes in the outcomes of interest. However, since tracking may not always be a viable option, an alternative is to simply collect data on intermediate indicators of long term impact in a cross-sectional survey (Ravallion, 2008).

Crucial to dealing with concerns over external validity is the need to properly understand the programme context. This requires data, especially administrative data. Data also allows us to understand the causal processes that underline the differences in outcomes. A researcher may collect detailed information about the specific setting, and use survey data to try and unpack why the outcomes occur as they do, and allow one to infer what might work in a different context. Ravallion (2008) suggests that one should focus on intermediate behavioural variables and not just outcome variables in this regard. In addition, it is important o have a process evaluation conducted alongside the evaluation itself, that is, an evaluation of whether the programme is being implemented as envisaged, whether monies are being spent as they should, and to obtain feedback from stakeholders that might be used to adapt and improve delivery on the ground. This kind of data is also vitally important for policy makers considering going to scale.

Despite these concerns over external validity, policymakers frequently do use lessons from past successful health policy interventions in designing new policies and programmes. In Section 7, we provide a review of some of the existing evidence concerning the impact of health interventions on individual welfare outcomes. While evidence emanating from Africa is scarce (with the exception of Kenya perhaps), the available evidence does suggest that health interventions aimed at combatting geohelminth infections, malnutrition, and iron deficiencies have significant positive impacts on individual productivity. In terms of other kinds of health interventions, the evidence is less well-established, suggesting scope for additional research in these areas.

7 Existing Evidence of Health Impacts

Most evaluations in developing countries that focus on health examine either the uptake of a certain health input (e.g., such as getting tested for HIV, using a mosquito net, going to the clinic) or look at ways to change health behavior (e.g. through increased education or knowledge, bargaining power). However, there are relatively few studies that look at the effects of health on economic variables such as productivity.

There are several reasons for the limited number of studies on this topic that are related to both the difficulty of this research question itself, as well as the context of Africa itself. As discussed in detail in section 2 above, causal inference is particularly difficult with estimating the relationship between health and wealth and there is a vast literature outlining these challenges (Smith 1999, Strauss 1986, Strauss and Thomas 1998). While randomized controlled trials provide one research strategy to mitigate the challenges of causal measurement of this question, there are additional challenges that make evaluating the relationship between health and economic outcomes difficult, especially in Africa. We discuss each of these challenges briefly.

Returns on investments in health often take a long time to realize and often these investments are made at early ages. Therefore, empirical analyses of the effects of early investments in health require longitudinal data collection on individuals that can measure health inputs and productivity after several decades. Alternatively, if only cross sectional data is available, this requires that data be collected on intermediate indicators of long term success (Ravallion, 2008). While the number of longitudinal studies in Africa is increasing, the number is still limited. Existing studies such as the Cape Area Panel Study, the Malawi Diffusion and Ideational Change Study, and the Kenya Life Panel Survey are among some examples of panel surveys that follow individuals over time.

Other surveys, such as the Demographic Surveillance Surveys, follow individuals over time, but often lack rich economic data; they instead focus on demographic and health indicators. Investment in longitudinal studies would help to build our knowledge of long-term effects of

early health investments. Political stability and funding are two challenges to conducting these longitudinal surveys, and the studies in South Africa, Malawi, Kenya, and Ghana are examples of countries that have had stable governments and the presence of researchers; however, more effort should be made to expand this list and to further understand linkages in other regions and countries.

There are further challenges to conducting studies that evaluate the causal relationship between health and productivity. In conducting randomized controlled trials, it is important to consider ethical implications of withholding some treatment from the control group. Internal review boards consider it unethical to withhold life-saving treatment from a study population and thus interventions must carefully consider these implications for the research. For example, we have limited evidence of the effects of ARVs on economic productivity of HIV-infected individuals. One of the reasons for this is that it could be viewed as unethical to have a study population of HIV positive individuals for whom some of them are in a control group, receiving no ARVs. Researchers who have examined this research question have used quasi-experimental methods to study the effects of treatment on economic behavior (Habyarimana et al. 2008; Thirumurthy et al. 2005). In addition to using these non-experimental methodologies, there are several other possibilities for researchers. First, encouragement designed evaluations can be conducted where treatment is not withheld from individuals; rather, individuals are given randomized encouragement such as subsidies or reminders to get their treatment. The randomized subsidy can then be used as an instrument for the treatment itself. A second approach that could be useful to explore is to partner with medical randomized controlled studies to study economic outcomes. For example, following individuals in phase III medical trials over time could be one promising avenue. If a drug or vaccine is found to be effective, these individuals could be followed over time to study longer-term effects of good health.

7.1 What have we learned from health evaluations to date?

One of the difficulties with evaluating the impact of health interventions on individual welfare and productivity is the time lag involved between the intervention, often made at a relatively young age, and welfare outcomes of interest, such as employment, income and poverty status in adult life. Consequently, in this arena, collecting data on intermediate outcomes such as school enrollment rates, labour market participation, and test scores aimed at measuring cognitive ability becomes important. Insofar as positive outcomes in these respects are associated with better long term prospects as an adult, they provide some evidence for the impact of health interventions on productivity. In this section, we briefly review some of the available evidence concerning the impact of health interventions on individual productivity.[7] The evaluation methods used in these studies encompass the entire range of evaluation methods reviewed earlier in this paper.

7.1.1 Nutritional supplementation

There is overwhelming and consistent evidence that malnutrition during the early years of a childs life is associated with lower cognitive levels and academic achievement, as well as higher dropout rates (Grantham-McGregor, 2007). Malnutrition which occurs in utero, or during the early years of a childs life can have serious and long lasting impacts on child development outcomes, and most often manifests itself as stunting. Longitudinal studies in developing countries have indicated that stunted children are less likely to be enrolled in school (Beasley et al, 2000), more likely to enrol late (Brooker et al, 1999; Moock and Leslie, 1986), and more likely to attain lower grades for their age [8] (Moock and Leslie, 1986; Jamison, 1986; Clark et al, 1990; Hutchinson et al, 1997). Part of the advantage that well nourished children enjoy is that they enter school earlier and thus have more time to learn, and they also appear to

[7] This section draws heavily on Burns (2007)

[8] The relationship between stature and age-appropriate grade is reduced with progression through school, which is compatible with a higher dropout rate for more stunted children.

enjoy greater learning productivity per year in the form of school attendance and homework completion. Young children who are malnourished also tend to show less positive affect, be less attentive, more apathetic, have poor social skills and have lower levels of play than healthy children (Gardner et al, 1999; Graves, 1978; Galler and Ramsey, 1989, Richardson et al, 1972). [9]

Randomised trials that have provided food supplements to improve the nutritional status of children have yielded gains of between 6 and 13 developmental quotient points for treatment children compared to those in the control group with regards to motor development, mental development and cognitive development (Waber et al, 1981; Grantham-McGregor et al, 1991; Pollit et al, 1993; Pollitt et al, 2000). /footnote A longitudinal study in Kenya (Sigman et al, 1989) documented that children who were better nourished achieved higher scores on a test of verbal comprehension and higher scores in Ravens matrices. Improved attention spans were particularly evident for well-nourished girls. Sigman et al (1991) also examine the extent to which cognitive abilities of 5 year olds in Kenya was affected by nutritional status. They show that food intake during the first two and a half years of life, and physical stature at two and half, was associated with better cognitive skills at age 5. Less information is available on the long term benefits of nutritional supplementation to children who are already malnourished, and the evidence that does exist has been the product of flawed research designs. These include low take-up of nutritional supplements, small sample sizes, and a follow-up period that was too short for any real benefits to have accrued. However, evidence from a study in Guatemala where food supplementation was begun during pregnancy and continued until the child was aged 2 suggest significant benefits, with these infants exhibiting less anxiety at age 6-8 and greater social skills (Pollitt et al, 1993; Barrett et al, 1982).

Provision of food supplements in the form of school meals may yield additional benefits over and above nutritional benefits. There is some evidence to suggest that this may also encourage attendance at school. Vermeersch and Kremer examine the effect of school meals on school participation in Kenya, and find that participation was 30% higher in Kenyan pre-schools where a free breakfast was introduced, than compared to control pre-schools where no such intervention occurred. Despite the fact that the provision of meals reduced teaching time, they also show that test scores were 0.4 standard deviations higher in treatment schools, although this was only the case if the teachers had good qualifications prior to the implementation of the programme. Alderman et al (1997) find that in Pakistan, a childs health and nutritional status is a significant predictor of school enrolment, and this is particularly the case for girls, thereby closing the gender education gap. Attanasio and Vera-Hernandez (2007) conduct an evaluation of a large scale community nursery programme in rural Colombia, which was implemented with the specific aim of providing nutritional supplementation and childcare to poor households. [10] Attanasio et al (2007) demonstrate that this programme had large and significant effects, both on the outcomes of the children, but also in terms of a labour supply effect for mothers in particular. More specifically, they show that a 6 year old boy who had been enrolled in this programme since birth would be 4.36 centimetres taller on average than boys who had not benefited from this programme, with an estimate of 4.41cms for girls. Moreover, mothers whose children were enrolled in this programme were 31% more likely to have been employed than mothers whose children were not enrolled.

Schultz (2007) examines the impact of the PROGRESSA programme in Mexico, which

[9] A longitudinal study by Berkman et al (2002) in Peru demonstrates that stunting at age 2 impacts negatively on cognitive outcomes measured at age 9, while a study in the Philippines demonstrated that stunting at age 2 led to higher drop out rates, later enrolment ages, higher grade repetition, and lower IQ scores amongst children at age 8 and 11. Walker et al (2005) provide evidence from Jamaica that shows that stunting before age 2 is associated with lower cognitive abilities and school achievement and higher dropout rates at age 17.

[10] In rural communities, eligible parents were asked to form local parents associations, and each association then elected a community mother. The community mother provided the childcare, and received up to a maximum of 15 children (all children of parents who were members of the parents association) in her home, in return for which the parents paid a small monthly fee to her. In addition, the state provided funds to provide food which was delivered on a weekly basis to the community mothers home. The children received three nutritionally balanced meals a day, lunch and two snacks as well as a nutritional drink. These food supplements were designed by a nutritionist to provide 70% of the daily recommended caloric intake for these children.

was designed to allow for a phase-in of conditional cash transfers. PROGRESSA provides cash grants, given to women, conditional on children attending school regularly and utilising preventative health measures (health care visits, nutritional supplements and participation in health education programmes). The programme was launched in 1998, but budgetary constraints made it impossible to roll the programme out nationally. Hence, the Mexican authorities rolled the programme out randomly, and used this phase-in design to help evaluate the project.[11] Schultz (2007) finds that enrollment increases by 3.4% for students in Grades 1-8, with the increase being larger for girls. In addition, participants who received the transfers enjoyed improved health outcomes. Gertler and Boyce (2001) demonstrate that the incidence of illness was reduced by 23% amongst recipient children, and the incidence of anemia was reduced by 18%. Moreover, children experienced a 1-4% increase in height. Behrman and Hoddinott (2000) demonstrate that for children aged 1-3 years, those who receive the treatment experience higher growth rates and are significantly less likely to be stunted. They estimate that treatment children experience an increase in growth rates of 16% of the mean growth rate relative to those who do not receive the treatment, and that these effects are larger for children from relatively poorer households. To the extent that health gains in early childhood translate into better cognitive development and academic performance at school, better health status and thus earnings potential as an adult, Berhman and Hoddinott (2000) estimate that exposure to the PROGRESSA treatment will result in an increase of 2.9% in lifetime earnings.

Given the success of the PROGRESSA programme, similar conditional cash transfer programmes have been implemented elsewhere. PROGRESSA was replicated in Colombia, although there the programme was called *Familias en Accion (FA)*. In this programme, mothers of children aged 0-17 were eligible to receive assistance. Beneficiary families with children under the age of 5 are eligible to receive a cash subsidy for nutrition, but to qualify for this, mothers must take their children for regular clinic visits. In addition, mothers are encouraged to participate in local education sessions on health and hygiene, and contraception. Households with children aged 6-17 receive a separate monthly grant per child, conditional on the child attending at least 80% of their classes. Attanasio et al (2005) demonstrate that FA had large and significant impacts on school attendance for children aged 12-17, increasing attendance by 10.1% in rural areas, and 5.2% in urban areas. The effect amongst children aged 8-11 was negligible, and they argue that this is mainly due to the fact that attendance amongst this cohort was high even prior to the introduction of the programme. FA also increased household consumption (and thus household welfare) significantly by 19.5% in rural areas, and by 9.3% in urban areas, with the bulk of this increased expenditure being devoted to food and clothing and footwear for children. Since FA requires children to visit clinics regularly, it is perhaps unsurprising to find that this significantly increased the number of children aged 0-2 who had an up-to-date schedule of health care visits, from 17% to 40%. Amongst children aged 2-4 years, this figure increased from 33.6% to 66.8%. FA also reduced the incidence of diarrhoea by approximately 10% for children aged 0-4 in rural areas.

In short, the evidence suggests that nutritional supplementation has significant and positive impacts on child development outcomes, and may yield added benefits in the form of higher school attendance, better academic performance and lower dropout rates.

7.1.2 Iron supplementation

Walker et al (2007) estimate that 44-66% of all children aged 4 and below in developing countries suffer anemia, with half of these cases being attributable to iron deficiencies. Iron deficiency holds negative consequences for child outcomes. From a survey of 21 articles, 19 report that young children with iron deficiency anemia have lower mental, social-emotional, motor and brain functioning than infants without (Lozoff et al, 2006; Grantham-McGregor, 2001). Importantly though, iron treatment in pre-school aged children with iron deficiency anemia has yielded

[11] While conditional cash transfer programmes have become increasingly popular as vehicles of development, their success does require that the conditionality be enforced. This involves an additional level of monitoring and an evaluation of the process.

positive cognitive benefits consistently over a number of studies (Grantham-McGregor and Ani, 2001; Sachdev et al, 2005). There are a number of large-scale trials on iron supplementation in infants or young children in developing countries, including Zanzibar (Stoltzfus et al 2001), Chile (Lozoff et al, 2003), Bangladesh (Black et al, 2004), Indonesia (Lind et al, 2003) and India (Black et al, 2002). Four of these aforementioned studies include infants at risk for stunting, while the fifth includes well nourished infants. All five studies report positive benefits of iron supplementation for motor skills, while the studies in India, Bangladesh and Chile also report social-emotional benefits. Finally, the Zanzibar and Chile studies also demonstrate cognitive language benefits for children receiving iron supplementation. It is worth noting that the Chilean study yields the largest number of beneficial outcomes, and this was the only study to target healthy infants. [12]This simply serves as a reminder that the outcome of an intervention will in part be a function of the characteristics of the target population.

Bobonis et al (2002) report results from the Balwadi Health project in India, in which they evaluate the impact of a non-governmental organization (NGO) pre-school nutrition and health project implemented in Delhi. This programme provides iron supplementation and deworming drugs to over 4000 children aged 2-6 years, through an existing pre-school network. The pre-schools in the study were randomly divided into three groups, and the schools were gradually phased into the program as it expanded over the course of two years. The results to date show that children in treatment schools gained significant weight (0.6 kgs on average) compared to children in control schools, and that average pre-school participation rates increased by 6.3 percentage points among assisted children, reducing pre-school absenteeism by roughly one-fifth. Moreover, they found an almost 50% reduction in the incidence of severe to moderate anemia. The longer-term benefits of iron supplementation are less clear, mainly due to insufficient evidence. The large scale randomised trials suggest that cognitive, social, emotional and motor development can all be positively affected by iron supplementation, at least in the short run, which is promising in terms of longer term effects.

In addition to potential effects on school attendance, there is evidence that suggests that iron supplements have a large effect on productivity of adult workers. Basta et al. (1979), found increased work output among anemic workers in Indonesia who were given iron supplements. However, while this study was a randomized controlled trial, their estimates are likely biased upwards due to problems of attrition. Another large-scale study of iron supplements in Indonesia found gains in adult productivity (as measured by earnings) especially among those who already had low hemoglobin levels (Thomas et al 2003).

7.1.3 Deworming

Illness due to worms is a problem that affects approximately one third of the worlds population, and the incidence of such infection is highest amongst school-aged children (Watkins and Pollitt, 1987). There are relatively few studies of the impact of worm infections on child development, and particularly for pre-schoolers, but arguably, poor health due to geohelminth infections not only has negative health effects but may also limit participation in pre-school activities. Hutchinson et al (1997) conduct a cross-sectional study of 800 children aged 9-13 in Jamaica and find an association between low academic achievement and mild levels of malnutrition and geohelminth infections. Oberhelman et al (1998) demonstrate a correlation between geohelminth infections and poor language development, while Callander et al (1998) show that treatment of children with trichuris dysentery syndrome produced improvements in mental and motor development after 4 years. These kinds of statistical associations suggest a compelling case for interventions aimed at improving school performance in developing countries to target the health and nutritional status of children. Bleakley (2007) finds that hookworm eradication campaigns in the southern United States in the early 1900s resulted in increased school enrollment and attendance. In that study, adults exposed to the deworming campaign as children

[12]It is possible that additional benefits were not seen in the other studies that targeted at-risk infants if these additional benefits required complementary activities, such as parental stimulation, nutritional supplements and so on.

were more likely to be literate as adults. Other studies in Jamaica and China found deworming improved children's scores on memory and cognition tests (Simeon et al. 1995; Nokes et al. 1999).Miguel and Kremer (2001) evaluate a programme of bi-annual school based treatment for worms with inexpensive deworming drugs in Kenyan schools. In this impact evaluation, 75 schools were phased into the programme in random order. They show that health and school participation increased at treatment schools, but that positive externalities were also generated for nearby control schools through reduced disease transmission. Absenteeism in treatment schools was significantly lower than in control schools, and they estimate that the programme increased schooling by 0.15 years per treated person. Finally, they also argue that what makes deworming such an attractive intervention strategy is that it is very cost effective relative to other interventions that provide free uniforms, textbooks or nutritional supplementation. [13]Bobonis et al (2002) find similar results in India as reported above.

7.1.4 HIV/AIDS

Given the AIDS pandemic across most African countries, this is one area where understanding the link between health and productivity becomes especially important. There have been a number of papers that have examined the economic effects of HIV/AIDS or the provision of ARVs on productivity. These studies are complicated with the difficulty of randomizing HIV status or of ARVs due to obvious ethical issues. Several studies have used other approaches to examine the long run effects such as matching or using quasi-experimental techniques (Habyarimana et al. 2008; Thirumurthy et al. 2005). Habyarimana et al (2008) find a significant reduction in worker absenteeism in the year following the introduction of ARVS in the workplace, and argue that for the typical manufacturing firm in East and Southern Africa, the benefit of providing ARV treatment to workers covers up to a third f the cost of treatment. Using longitudinal survey data from Western Kenya, Thirumurthy et al (2005) show that within six months of beginning ARV treatment, adult ARV recipients are 20% more likely to participate in the labour force, and they increase their weekly work hours by a third. Moreover, they argue that these estimates are, in fact, an underestimate, since in the absence of treatment, worker productivity would have declined even further. Hence, the upper bound of the impact of treatment is larger. Thirumurthy et al (2005) also find that once adult AIDS patients within the household begin treatment, young boys within the household work fewer hours in the labour market, thereby potentially yielding positive outcomes for school attendance and attainment. Evidence concerning the impact of HIV status on child outcomes is scant, but Brown et al (2000) argue that HIV status in children is associated with delays in language acquisition, and to the extent that this translates into educational penalties, will affect later labour market prospects. Moreover, many children have been orphaned by AIDS, and thus find themselves vulnerable and often living in chronic poverty. This impacts their developmental potential since they have reduced access to resources and must deal with a great deal of psychological stress. Case and Ardington (2006) show that orphans are less likely to be enrolled in school, and if they are in school, they lag behind children of the same age.

7.1.5 Other health interventions

There are numerous other kinds of health interventions that might potentially also yield positive impacts on productivity and incomes later in life. For example, the effects of indoor air pollution due to use of cooking fuel within a household has been suggested to be an important factor in economic productivity (Duflo, Greenstone and Hanna 2008). Malaria may also reduce productivity and there have been a number of papers that have examined the effects through non-experimental methods (Ashraf, Fink and Weil 2009). In terms of supplementation of other micronutrients, the evidence is either insufficient with more randomised control trials being

[13]Since several programme interventions were conducted in Kenya in similar environments, they are able to make cost-benefit comparisons of these different kinds of interventions. They show that deworming costs $3.50 per additional year of school participation, compared to $99 for the provision of free uniforms, and $36 for nutritional supplementation. (the latter programme was targeted to pre-schools specifically)

needed to make an strong causal statements or the evidence that is available is simply not compelling. For example, evidence of Vitamin A deficiency is scant, and Walker et al (2007) argue that this can be ignored as a priority since there is little evidence to suggest that vitamin A supplements would have a large impact on the development outcomes for young children. By way of contrast, zinc deficiencies are estimated to affect one third of the worlds population, yet the evidence of the role of zinc in child development is unclear. Importantly,zinc supplementation may produce negative outcomes if provided to children who are not lacking zinc to begin with, since it affects the balance of other micronutrients. However, that being said, zinc supplementation has been associated with better motor development and behaviour amongst children in a Bangladesh study (Stoltzfus et al, 2001) but no such effect was found in India or Indonesia (Grantham-McGregor et al 2007).

Iodine forms part of thyroid hormones and as such, is crucial for the functioning of the central nervous system. It also aids in regulating physiological processes, and deficiency can lead to mental retardation. Despite a worldwide campaign to combat iodine deficiency through salt iodisation, this deficiency is still considered a risk factor. A 1994 meta-analysis of 18 studies of children and adolescents concluded that IQ scores were 13.5 points lower amongst children with iodine deficiency (Walker et al, 2007). Another meta-analysis in 2005 (based on publications in Chinese journals) showed that IQ scores were 12.5 points lower for children living in iodine-deficient areas, and who had lived there during their childhood years. Moreover, children who received iodine supplementation both pre- and post-natally had IQ scores that were 8.7 points higher on average than children who did not receive such supplementation (Walker et al, 2007). Finally, a longitudinal study in China suggests that iodine supplementation during the first and second trimesters of pregnancy may be more effective than supplementation during the third trimester, or during infancy.

There have been other studies of potential direct effects of health interventions aimed at improving water, sanitation and infrastructure. While the number of randomized controlled trials are increasing, there still only remains a limited number that examine longer term or economic effects. Access to clean water and proper sanitation reduces the risk of diarrhoeal disease for young children. Diarrhoea is especially prevalent during the first 2 years of life, making it an important risk factor, although Walker et al (2007) argue that there is no proper evidence concerning the link between diarrhoeal disease and child development per se. While two small-scale studies in Brazil suggest there is an association between the incidence of diarrhoea in the first two years of life and cognitive outcomes, a larger cohort study in Peru that controls for other covariates does not find any such association (Berkman et al, 2002; Guerrant et al, 1999; Niehaus et al, 2002). The lack of evidence in this regard does not mean that no link exists, simply that there is insufficient documented evidence to be persuasive that an intervention on this front yields substantial developmental benefits.

8 Conclusions

Randomization is often viewed as the ideal method to deal with the problem of selection bias. When appropriate to the policy context, the results of randomized evaluations are relatively easy to communicate because they generally do not require substantial qualifying assumptions. An added advantage is the transparency associated with choosing a control group *ex-ante*. However, these advantages of randomization justify its use to the exclusion of other methods only when interventions are of such a nature that they affect an entire population. In the case of a health intervention, if participation is rendered mandatory and the intervention is rolled out randomly across districts, then randomization at the district level will yield population-wide average treatment effects that are unconfounded by selection bias.

However since participation in health interventions is most often voluntary, randomization alone is usually insufficient. Under this more realistic scenario more explicit modeling exercises are required to identify treatment effects. Propensity score matching has been shown to be quite effective when coupled with less-than-perfect experimental designs. Heckman and Smith (1996)

have also argued that randomizing eligibility could be coupled with instrumental variables. This type of quasi-experimental design works quite well when the eligibility rules of the program are not compromised during implementation. When eligibility is correlated with outcomes however, the analyst might be forced to look for IVs elsewhere. In such instances, detailed knowledge of the institutional environment as well as the administration of the program could prove useful in constructing alternative IVs.

While experimental designs are always desirable when evaluating health impacts, they are not a panacea to all data problems. Identification strategies that rely solely on randomizing treatment assignment have to contend with the problem of selective compliance and attrition from both the treatment and control groups. Guarding against such problems will often involve combining methods and/or building into studies additional rules concerning participation This may require conditionality to be imposed on participants, as was the case with PROGRESSA, or may require significant investments of time and energy by the research team in establishing good working relationships with survey participants, as well as the ability to maintain contact over time in the case of longitudinal studies. Moreover, interventions that are simple to administer and for participants to adhere to have a stronger chance of success than interventions that require a complex bureaucratic structure in order to be administered, or where the intervention requires significant education or time commitment on the part of participants.

Where health investments are made at early ages, longitudinal data is ideally required to assess longer term health impacts on productivity. When the collection of longitudinal data is not possible, intermediate indicators of long term success should be collected in cross-sectional surveys. Given the costs involved in data collection exercises, collecting such data might best be accomplished by partnering with medical randomized controlled studies.

In sum, the evaluation problem is really one of missing data. The credibility of impact estimates will only ever be as good as the data upon which they are based. Randomized evaluations that do not control adequately for selective compliance and attrition will necessitate the use of NX methods as well as substantial collection of good quality data, including administrative and process data to provide important insights about the context and inner workings of the programme, so that additional analytical options are available if important aspects of the experimental design of a program are prone to unravelling.

A Appendix

A.1 Derivation of the Wald Estimator

Our derivation follows Wooldridge (2002). The the numerator can be written as $\sum_{i=1}^{N} P_i(y_i - \bar{y}) = \sum_{i=1}^{N} P_i y_i - (\sum_{i=1}^{N} P_i)\bar{y} = N_1\bar{y_1} - N_1\bar{y} = N_1(\bar{y_1} - \bar{y})$ where $N_1 = \sum_{i=1}^{N} P_i$ is the number of observations in the sample with $P_i = 1$ and $\bar{y}_i$ is the average of the y_i over the observations with $P_i = 1$. Next write $\bar{y}$ as a weighted average: $\bar{y} = \frac{N_0}{N}\bar{y_0} + \frac{N_1}{N}\bar{y_1}$, where the zero/one subscripting refers to treatment and control. After some algebra it can be shown that $\bar{y_1} - \bar{y} = (\frac{N-N_1}{N})\bar{y_1} - (\frac{N_0}{N})\bar{y_0} = (\frac{N_0}{N})(\bar{y_1} - \bar{y_0})$. So the numerator of the IV estimate is $(\frac{N_0 N_1}{N})(\bar{y_1} - \bar{y_0})$. The same argument shows that the denominator is $(\frac{N_0 N_1}{N})(\bar{T_1} - \bar{T_0})$. Taking the ratio completes the proof.

A.2 Derivation of the Probability Limit of the Wald Estimator Using $D \times P$ as an IV

We begin by computing the following conditional expectations:

$$\begin{aligned}
E(y_{ij}|D_i=1,P_{ij}=1) &= \alpha+\beta+\gamma+\delta E(T_{ij}|D_i=1,P_{ij}=1) \\
&+ \eta+E(v_i|D_i=1) \\
E(y_{ij}|D_i=1,P_{ij}=0) &= \alpha+\beta+\delta E(T_{ij}|D_i=1,P_{ij}=0)+E(v_i|D_i=1) \\
E(y_{ij}|D_i=0,P_{ij}=1) &= \alpha+\gamma+\delta E(T_{ij}|D_i=0,P_{ij}=1)+E(v_i|D_i=0) \\
E(y_{ij}|D_i=0,P_{ij}=0) &= \alpha+\delta E(T_{ij}|D_i=0,P_{ij}=0)+E(v_i|D_i=0)
\end{aligned}$$

We will also need to compute:

$$\begin{aligned}
&E(T_{ij}|D_i=1|P_{ij}=1) \\
&E(T_{ij}|D_i=1|P_{ij}=0) \\
&E(T_{ij}|D_i=0|P_{ij}=1) \\
&E(T_{ij}|D_i=0|P_{ij}=0)
\end{aligned}$$

We can now construct difference-in-difference estimators for the effect of D and P on consumption, as well as on the dose variable:

$$\begin{aligned}
\hat{\Delta}_{y|D,P} &= [E(y_{ij}|D_i=1,P_{ij}=1)-E(y_{ij}|D_i=1,P_{ij}=0)] \\
&- [E(y_{ij}|D_i=0,P_{ij}=1)-E(y_{ij}|D_i=0,P_{ij}=0)]
\end{aligned}$$

$$\begin{aligned}
\hat{\Delta}_{T|D,P} &= [E(T_{ij}|D_i=1,P_{ij}=1)-E(T_{ij}|D_i=1,P_{ij}=0)] \\
&- [E(T_{ij}|D_i=0,P_{ij}=1)-E(T_{ij}|D_i=0,P_{ij}=0)]
\end{aligned}$$

Taking the ratio of these two estimators produces a Wald estimator with probability limit,

$$\begin{aligned}
\tilde{\delta}_{IV} &= \frac{\Delta_{y|D,P}}{\Delta_{T|D,P}} \\
&\xrightarrow{p} \delta+\frac{\eta}{\Delta_{T|D,P}}
\end{aligned}$$

References

Agodini, R., and M. Dynarski (2004): "Are Experiments the Only Option? A Look at Dropout Prevention Programs," *Review of Economics and Statistics*, 86(1), 180–194.

Alderman, H., J. Behrman, V. Lavy, and R. Menon (1997): "Child Nutrition, Child Health, And School Enrollment: A Longitudinal Analysis," World Bank Policy Research Working Paper No. 1700.

Angrist, J. (1990): "Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records," *American Economic Review*, 80, 313–335.

Angrist, J., and J. Hahn (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *Review of Economics and Statistics*, 86(1), 58–72.

Angrist, J. D., and A. B. Krueger (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card. Elsevier, Amsterdam, North Holland.

Attanasio, O., E. Battistin, E. Fitzsimons, A. Mesnard, and M. Vera-Hernandez (2005): "How Effective Are Conditional Cash Transfers Evidence From Colombia," Institute For Fiscal Studies, Briefing Note No. 54.

Attanasio, O., and A. M. Vera-Hernandez (2004): "Medium and Long Run Effects of Nutrition and Child Care: Evaluation of a Community Nursery Programme in Rural Colombia," Working Paper EWP04/06, Centre for the Evaluation of Development Policies, Institute of Fiscal Studies London.

Attanasio, O., and M. Vera-Hernandez (2007): "Nutrition And Child Care Choices: Evaluating A Community Nursery Programme In Rural Colombia," Institute For Fiscal Studies Working Paper EWP04/06.

Beasley, N., A. Hall, and A. Tomkins (2000): "The Health of Enrolled And Not Enrolled Children At School Age In Tanga, Tanzania," *Acta Tropica,*, 76, 223–229.

Becker, S., and A. Ichino (2002): "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, 2, 358–377.

Behrman, J., Y. Cheng, and P. Todd (2004): "Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach," *Review of Economics and Statistics*, 86(1), 108–32.

Behrman, J., and J. Hoddinott (2000): "An Evaluation of The Impact Of Progressa On Pre-School Child Heigh," International Food Policy Research Institute, Working Paper, July.

Behrman, J., P. Sengupta, and P. Todd (2002): "Progressing through PROGESA: An Impact Assessment of a School Subsidy Experiment in Mexico," University of Pennsylvania.

Berkman, D., A. Lescano, R. Gilman, S. Lopez, and M. Black (2002): "Effects Of Stunting, Diarrhoeal Disease, And Parasitic Infection During Infancy On Cognition In Late Childhood: A Follow-Up Study," *Lancet*, 359, 296–300.

Black, M., S. Sazawal, R. Black, S. Khosia, J. Kumar, and V. Menon (2004): "Cognitive And Motor Development Among Small For Gestational Age Infants: Impact Of Zinc Supplementation, Birth Weight And Care Giving Practices," *Pediatrics*, 113, 1297–305.

Bleakley, H. (2007): "Disease and Development: Evidence from Hookworm Eradication in the American South," *Quarterly Journal of Economics,*, 122, 73–117.

Bloom, D. E., D. Canning, and J. Sevilla (2004): "The Effect of Health on Economic Growth: A Production Function Approach," *World Development*, 32(1), 1–13.

Bobonis, G., E. Miguel, and C. Sharma (2002): "Iron Supplementation and Early Childhood Development: A Randomized Evaluation in India," University of California, Berkeley.

Buddelmeyer, H., and E. Skoufias (2004): "An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA," Policy Research Working Paper 3386, World Bank, Washington DCr.

Burtless, G. (1995): "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9(2), 63–84.

Chase, R. (2002): "Supporting Communities in Transition: The Impact of the Armenian Social Investment Fund," *World Bank Economic Review*, 16(2), 219–240.

Clark, N., S. Grantham-McGregor, and C. Powell (1990): "Health And Nutrition Predictors Of School Failure In Kingston, Jamaica," *Ecological Food Nutrition,*, 26, 1–11.

Cochran, W. G. (1968): "The effectiveness of adjustment by subclassification in removing bias in observational studies," *Biometrics*, 24, 205–213.

Deaton, A. (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. John Hopkins University Press, Baltimore, MD.

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

——— (2002): "Propensity Score Matching Methods for Nonexperimental Causal Studies," *Review of Economics and Statistics*, 84, 151–161.

DeLong, J. B., and K. Lang (1992): "Are All Economic Hypotheses False?," *Journal of Political Economy*, 100(6), 1257–72.

Diaz, J. J., and S. Handa (2004): "An Assessment of Propensity Score Matching as a NX Impact Estimator: Evidence from a Mexican Poverty Program," University of North Carolina Chapel Hill.

Duflo, E. (2003): "Scaling Up and Evaluation," Paper prepared for the ABCDE in Bangalore.

Duflo, E., M. Greenstone, and R. Hanna (2008): "Indoor Air Pollution, Heatlh and Economic Well-being," MIT Working Paper.

Duflo, E., and M. Kremer (2005): "Use of Randomization in the Evaluation of Development Effectiveness," in *Evaluating Development Effectiveness*, ed. by O. F. George Pitman, and G. Ingram. Transaction Publishers, New Brunswick, NJ.

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

Frankenberg, E., W. Suriastini, and D. Thomas (2005): "Can Expanding Access to Basic Healthcare Improve Childrens Health Status? Lessons from Indonesias 'Midwife in the Village Program," *Population Studies*, 59(1), 5–19.

Galasso, E., and M. Ravallion (3): "Social Protection in a Crisis: Argentinas Plan Jefes y Jefas," *World Bank Economic Review*, 18, 367–399.

Galasso, E., M. Ravallion, and A. Salvia (2004): "Assisting the Transition from Workfare to Work: Argentinas Proempleo Experiment," *Industrial and Labor Relations Review*, 57(5), 128–142.

Galiani, S., P. Gertler, and E. Schargrodsky (2005): "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy*, 113(1), 83–119.

Gertler, P. (2004): "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment," *American Economic Review, Papers and Proceedings*, 94(2), 336–41.

Gertler, P., and S. Boyce (2001a): "An Experiment In Incentive-Based Welfare: The Impact Of PROGRESSA On Health In Mexico," University of California, Berkley.

Gertler, P. J., and S. Boyce (2001b): "An experiment in incentive-based welfare: The impact of PROGRESA on health in Mexico," University of California, Berkeley.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz (2004): "Retrospective vs. Prospective Analysis of School Inputs: The Case of Flip Charts in Kenya," *Journal of Development Economics*, 74, 251–268.

Godtland, E., E. Sadoulet, A. D. Janvry, R. Murgai, and O. Ortiz (2004): "The Impact of Farmer Field Schools on Knowledge and Productivity: A Study of Potato Farmers in the Peruvian Andes," *Economic Development and Cultural Change*, 53(1), 63–92.

Grantham-Mcgregor, S., Y. Cheung, S. Cueto, P. Glewwe, L. Richter, and B. Strupp (2007): "Developmental Potential In The First 5 Years For Children In Developing Countries," *Lancet*, 369.

Habyarimana, J., B. Mbakile, and C. Pop-Eleches (2000): "HIV/AIDS, ARV Treatment and Worker Absenteeism: Evidence from a Large African Firm," Unknown.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparamentric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

Heckman, J., and J. Hotz (1989): "Choosing Among Alternative NX Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–874.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J., H. Ichimura, and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, 64, 605–654.

Heckman, J., R. Lalonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics: Volume 3A*, ed. by O. Ashenfelter, and D. Card, pp. 1865–2097. Elsevier, Amsterdam, North Holland.

Heckman, J., L. Lochner, and C. Taber (1998): "General Equilibrium Treatment Effects: A Study of Tuition Policy," NBER Working Paper 6426.

Heckman, J., and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman, and B. Singer. Cambridge University Press, Cambridge, UK.

Heckman, J., and J. Smith (1995): "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9(2), 85–110.

Heckman, J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for heterogeneity in Programme Impacts," *Review of Economic Studies*, 64(4), 487–535.

Hirano, K., and G. Imbens (1973): "The Propensity Score with Continuos Treatments," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. by A. Gelman, and X.-L. Meng, pp. 73–84. Wiley, West Sussex, UK.

Hoddinott, J., and E. Skoufias (2004): "The Impact of PROGRESA on Food Consumption," *Economic Development and Cultural Change*, 53(1), 37–61.

Imbens, G., and J. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62(2), 467–475.

Jacoby, H. G. (2002): "Is There an Intrahousehold 'Flypaper Effect'? Evidence from a School Feeding Programme," *Economic Journal*, 112(476), 196–221.

Jalan, J., and M. Ravallion (1998): "Are There Dynamic Gains from a Poor-Area Development Program," *Journal of Public Economics*, 67(1), 65–86.

Jamison, D. (1986): "Child Malnutrition And School Performance In China," *Journal Of Development Economics*, 20, 299–309.

Lalonde, R. (1986): "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604–620.

Lanjouw, P., and M. Ravallion (1999): "Benefit Incidence and the Timing of Program Capture," *World Bank Economic Review*, 13(2), 257–274.

Leamer, E. (1983): "Lets take the Con Out of Econometrics," *American Economic Review*, 73(1), 31–43.

Manski, C. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 531–542.

Miguel, E., and M. Kremer (2004): "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72(1), 159–217.

Moock, P., and J. Leslie (1986): "Child Malnutrition And Schooling In The Terai Region Of Nepal," *Journal Of Development Economics*, 20, 33–52.

Muney, A. L., and S. Jayachandran (2006): "Longevity and human capital investments: evidence from declines in maternal mortality," Policy responses in Health.

Newman, J., M. Pradhan, L. B. Rawlings, G. Ridder, R. Coa, and J. L. Evia (2002): "An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund," *World Bank Economic Review*, 16, 241–274.

Nokes, C., S. McGarvey, L. Shiue, G. Wu, H. Wu, D. Bundy, and G. Olds (1999): "Evidence for an improvement in cognitive function following treatment of Schistosoma japonicum infection in Chinese primary schoolchildren," *American Journal of Tropical Medicine and Hygiene*, 60, 556–565.

Ravallion, M. (1973): "Evaluating Anti-Poverty Programs," in *Handbook of Development Economics: Volume 4*, ed. by R. E. Evenson, and T. P. Schultz, pp. 3787–3846. Elsevier, Amsterdam, North-Holland.

Rosenbaum, P. R. (1973): "Matching in Observational Studies," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. by A. Gelman, and X.-L. Meng, pp. 15–24. Wiley, West Sussex, UK.

Rosenbaum, P. R., and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

Rosenzweig, M., and K. Wolpin (1986): "Evaluating the Effects of Optimally Distributed Public Programs: Child Health and Family Planning Interventions," *American Economic Review*, 76, 470–480.

Rubin, D. B., and N. Thomas (2000): "Combining propensity score matching with additional adjustments for prognostic covariates," *Journal of the American Statistical Association*, 95, 573–585.

Schultz, T. P. (2004): "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program," *Journal of Development Economics*, 74(1), 199–250.

Sigman, M., M. McDonald, C. Neumann, and N. Bwibo (1991): "Prediction Of Cognitive Competence In Kenyan Children From Toddler Nutrition, Family Characteristics And Abilities," *Journal Of Child Psychology And Psychiatry*, 32.

Sigman, M., C. Numann, A. Jansen, and N. Bwibo (1989): "Cognitive Abilities Of Kenyan Children In Relation To Nutrition, Family Characteristics And Education," *Child Development*, 60, 1463–74.

Simeon, D. T., S. M. Grantham-McGregor, J. E. Callender, and M. Wong. (1995): "Treatment of Trichuris trichiura Infections Improves Growth, Spelling Scores and School Attendance in some Children," *Journal of Nutrition*, 125, 1875–1883.

Skoufias, E. (2005): "PROGRESA and Its Impact on the Welfare of Rural Households in Mexico," Research Report 139, International Food Research Institute, Washington DC.

Smith, J. (1999): "Healthy Bodies and Thick Wallets: the dual relation between health and economic status," *Journal of Economic Perspectives*, 13(2), 145–166.

Smith, J., and P. Todd (2005): "Does Matching Overcome LaLonde's Critique of NX Estimators," *Journal of Econometrics*, 125(12), 305–353.

Strauss, J. (1986): "Does Better Nutrition Raise Farm Productivity?," *Journal of Political Economy*, 00, 297–320.

Strauss, J., and D. Thomas (1998): "Health, Nutrition, and Economic Development," *Journal of Economic Literature*, 36(2), 766–817.

Thirumurthy, H., J. G. Zivin, and M. Goldstein (2005): "The Economic Impact of AIDS Treatment: Labor Supply in Western Kenya," NBER Working Paper 11871.

Thomas, D., E. Frankenberg, J. Friedman, et al. (2003): "Iron Deficiency and the Well-Being of Older Adults: Early Results from a Randomized Nutrition Intervention," Paper Presented at the Population Association of America Annual Meetings, Minneapolis.

Vermeersch, C., and M. Kremer (2004): "School Meals, Educational Achievement And School Competition: A Randomized Evaluation," *World Bank Policy Research Paper*, 3523.

Walker, S., S. Chang, C. Powell, and S. G. Mcgregor (2005): "Effects Of Early Childhood Psychosocial Stimulation And Nutritional Supplementation On Cognition And Education In Growth-Stunted Jamaican Children: Prospective Cohort Study," *Lancet*, 366, 1804–07.

Walker, S., T. Wachs, J. M. Gardener, B. Lozoff, G. Wasserman, E. Pollitt, and J. Carter (2007): "Child Development: Risk Factors For Adverse Outcomes In Developing Countries," *Lancet*, 369.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Massachusetts.

Yaari, M. E. (1965): "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *The Review of Economic Studies*, 32(2), 137–150.