African Development Review
Revue africaine de développement

African Development Review, Vol. 24, No. 4, 2012, 302–315

Evaluating the Impact of Health Programmes on Productivity

Malcolm Keswell, Justine Burns and Rebecca Thornton*

***Abstract*:** This article reviews some of the key methodological approaches available to researchers interested in identifying a causal relationship between health interventions and economic indicators of productivity. We then discuss some of the empirical work that has utilized these techniques in making the case for a causal relationship from health interventions to productivity. We conclude that while considerable progress has been made in addressing concerns over attribution, much work remains to be done in expanding our knowledge of why certain interventions work whilst others do not.

1. Introduction

While the relationship 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, requiring that special attention be paid to identification strategies when trying to establish causal effects of health interventions on economic outcomes. Even when credible identification strategies are available, the evidence on this link tends to be limited in scope and generalizability; this is partly due to the fact that evaluating the impact of health interventions on individual welfare and productivity involves time lags between the intervention, often made during infancy or childhood, and welfare outcomes of interest, such as employment status, usually observed in adulthood.

In this paper, we review some of the key methodological approaches available to researchers interested in identifying a causal relationship between health interventions and economic indicators of productivity or success, and then discuss some of the empirical work that has utilized these techniques in making the case for a causal relationship from health interventions to productivity. We conclude by reflecting on those areas where the evidence is still limited in scope, thus providing opportunities for future research.

2. Methods of Establishing Attribution

An unbiased assessment of the impact of a programme or intervention requires that some type of inference be made about what the outcomes of the participants of the programme would have been had they not participated. Denote with y_1 an outcome of household i when it is exposed to the programme and denote y_0, had it not been exposed to the programme.[1] Researchers are interested in what difference the programme is likely to have on the outcome of interest, that is, the difference $\Delta = y_1 - y_0$. The problem is that either y_1 or y_0 is observed for each household, making it impossible to know Δ.

Let $D = 1$ denote households who participate in the programme and let $D = 0$ denote households who are not participants of the programme. Outcomes for both $D = 1$ and $D = 0$ households are observed. Further, let $\mathbf{x}$ denote a vector of observed characteristics of the household. The most basic parameter of interest to be estimated is the average treatment effect on the treated (ATT):

$$\begin{aligned} ATT &= E(\Delta | \mathbf{x}, D = 1) \\ &= E(y_i - y_0 | \mathbf{x}, D = 1) \\ &= E(y_1 | \mathbf{x}, D = 1) - E(y_0 | \mathbf{x}, D = 1) \end{aligned} \tag{1}$$

* Malcolm Keswell, School of Economics, University of Cape Town, Rondebosch, 7700; e-mail: malcolm.keswell@uct.ac.za. Justine Burns, School of Economics, University of Cape Town, Rondebosch, 7700; e-mail: justine.burns@uct.ac.za. Rebecca Thornton, Department of Economics, University of Michigan, 611 Tappan St., 213 Lorch Hall, Ann Arbor, MI, 48106, USA. The authors would like to thank especially the editor David Sahn. They also thank Robert Eastwood, Peter Glick, German Mwabu and Duncan Thomas for useful comments, as well as seminar participants at the AERC Collaborative Research workshop on Health, Economic Growth and Poverty Reduction held in Accra, Ghana, 20–22 April 2009.

The presence of the second term in the last line of Equation 1 summarizes the key identification problem that must be solved. In this section, we outline the most widely used empirical approaches to solving this identification problem.

2.1 Randomization

Randomizing assignment to the treatment group theoretically eliminates selection bias in the estimates of mean impact. Randomization involves a lottery process. By restricting attention to the $D = 1$ sub-population, and then randomizing assignment to the treatment group, one obtains a control group whose outcomes in the $D = 0$ state serve as the counterfactual outcome, while the mean outcomes of households in the $D = 1$ group (denoted the treatment group) serves as an estimate of first term. An advantage of this process is that it creates two groups that *ex-ante* are statistically similar among observables and unobservables, removing potential differences that could exist between the two groups for which social scientists have difficulty controlling, 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 not confounded by selection bias since treatment status is randomly allocated. We shall return in the next section to some of the main practical issues one has to contend with when designing a randomized treatment-control study. The main reference here is Duflo *et al.* (2008).

2.2 Matching

When randomization is not practically or politically feasible, or when there has been no *ex-ante* random assignment of the programme, more appropriate counterfactuals can be found by matching households that received a programme to similar households that did not. Matching treatment households to control households on the basis of **x** is one way to proxy for the missing counterfactual $E(y_0|\mathbf{x}, D = 1)$. The basic idea is to pair each treatment household with a household from the control group deemed to be observably similar in **x**, and then take the average difference in outcomes between each pairing to give the effect of the programme on the average treated household.

The ideal approach would be to match households directly on their characteristics.[2] However, exact matching is often not practical, first because some of the conditioning variables might be continuous and secondly because **x** might be of large dimension.[3] A standard non-experimental alternative to exact matching is the technique of propensity score matching (Rosenbaum and Rubin, 1983; Heckman *et al.,* 1997; Hahn, 1998). Under this approach, one tries to reduce the dimensionality of the problem by matching against a scalar index or propensity score — that is, the predicted probabilities that are computed from a regression where the outcome variable is a binary indicator of treatment. Formally, we can define the propensity score as the conditional probability of receiving treatment (i.e., where $D = 1$), given **x**:

$$p(\mathbf{x}) = \Pr(D = 1|\mathbf{x}) = E(D|\mathbf{x}) \tag{2}$$

Rosenbaum and Rubin (1983) provide a result (theorem 2) which establishes that if matching directly on **x** is an appropriate way of removing selection bias, then matching on $p(\mathbf{x})$ is equally appropriate. Specifically, they showed that the distribution of the covariates for treatment households and control households will be the same once one has conditioned on the propensity score (i.e., outcomes will be independent of treatment assignment, so treatment assignment is 'strictly ignorable'). If interest centres on the average treatment effect, then strict ignorability can be replaced with the weaker assumption of conditional mean independence:

$$E(y_0|\mathbf{x}, D = 1) = E(y_0|\mathbf{x}, D = 0) = E(y_0|\mathbf{x}) \tag{3}$$

A variety of matching methods have been proposed in the evaluation literature.[4] The main choice turns on how much information to use from the control group in defining a match for each treated household. The most commonly used method is based directly on the propensity score (Rosenbaum and Rubin, 1983).[5]

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. If assignment to treatment and control is random, that is, if one's position in the pipeline is randomly assigned, then the usual techniques as applied to a randomized controlled trial (RCT) can be used to assess impact. However, if position in the pipeline (or assignment to treatment versus control status) is not random, then matching methods can be utilized to measure impact. We describe such a pipeline matching process below.

Following Smith and Todd (2005), start by denoting S_p as the region of common support of $\hat{p}(\mathbf{x})$ between the $D = 1$ and $D = 0$ distributions. Let N_1 denote the set of households that have already received treatment through the programme, and let N_0 denote the set of households still awaiting treatment. Now denote as n_1 the number of treated households falling into the common support region of the estimated propensity score density; that is, the number of households falling into the set $N_1 \cap S_p$. The general form of the matching estimator is then given by

$$\begin{aligned}\delta &= (n_1)^{-1} \sum_{i \in N_1 \cap S_p} (y_{1i} - \hat{E}(y_{0i} | D_i = 1, p_i(x))) \\ &= (n_1)^{-1} \sum_{i \in N_1 \cap S_p} \left(y_{1i} - \sum_{j \in N_0} \omega(i, j) y_{0j} \right)\end{aligned} \tag{4}$$

where $i \in N_1 \cap S_p$ denotes the *i*th treated household in the common-support region for which a match is sought. The second term in this expression serves as a matched substitute for the outcomes of a randomized-out household of the treatment group, where the imputed counterfactual outcome is taken as a weighted average (given by $\sum_{j \in N_0} \omega(i, j) y_{0j}$) of the outcomes for a set of possible matches. This matching set is given by

$$A_i = \{ j \in N_0 | \, p_j(\mathbf{x}) \in C(p_i(\mathbf{x})) \}$$

where $c(p_i(\mathbf{x}))$ defines a neighborhood around each estimated propensity score for households in the treatment group, so that a *j*th match from the N_0 sample is drawn for every *i*th treated household if this *j*th control group household falls within the neighborhood.

A widely used method of constructing a matched counterfactual for every $i \in N_1 \cap S_p$ is to pick a *j*th control group household that satisfies both $j \in .N_0 | p_j \in C(p_i)$ and that has the smallest Euclidean distance to *i*, that is,

$$A_i(p(\mathbf{x})) = \{ p_j(\mathbf{x}) | \min_j \| p_i(\mathbf{x}) - p_j(\mathbf{x}) \| \} \tag{5}$$

Note that it must be the case that $\sum \omega(i, j) = 1$.The most restrictive implementation of this approach sets $\sum \omega(i, j) = 1$ when $j \in A_i(p(\mathbf{x}))$, and $\omega(i, j) = 0$ when $j \ni A_i(p(\mathbf{x}))$, so that a single match is selected from the N_0 sample and matched against a single household from the N_1 sample. A more efficient implementation would involve more than one nearest neighbour (for example the set $C(p_i(\mathbf{x}))$ might contain five members), but this comes at the cost of greater bias since equal weight would be given to these five nearest neighbours.

Another option — one that is more efficient, data permitting — is to match non-parametrically using a kernel-weighted average over multiple households in the control group, with declining weights for poorer matches (Heckman *et al.*, 1997, 1998a). In this instance the formula for the ATT remains as in Equation 4, but the weight given to the *j*th control group household in matching it to the *i*th treated household is determined by a kernel function of the form

$$\omega(i, j) = K \left(\frac{p_j(\mathbf{x}) - p_i(\mathbf{x})}{h_n} \right) / K \left(\frac{p_k(\mathbf{x}) - p_i(\mathbf{x})}{h_n} \right) \tag{6}$$

where K is a kernel function and h_n is a bandwidth parameter. 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 the specific functional form chosen for K. In practice, the choice of functional form for K tends to be less important that the choice of bandwidth.

2.3 Other Non-Experimental Methods

Two potential problems remain unexplored with the matching 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 the full set of variables influencing treatment status is not observed: instrumental variable estimation, regression discontinuity approaches and double-differencing.

Instrumental Variables

A key feature of this framework is that unobservables do not bias the treatment effect as long as an instrumental variable (IV) 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 one is dealing with a ‘broken’ experimental design premised on randomizing treatment, and there is concern that not all of the important variables predicting treatment can be observed given the survey instrument employed, IVs might offer a useful alternative.

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 the treatment dummy; y is the outcome variable; and i, j indexes villages/primary sampling units 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 T is instrumented 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

$$\begin{aligned}\hat{\delta} &= \frac{\left(\sum_{i=1}^{N}(P_{ij} - \bar{P})(y_{ij} - \bar{y})\right)}{\left(\sum_{i=1}^{N}(P_{ij} - \bar{P})(T_{ij} - \bar{T})\right)} \\ &= \frac{\left(\sum_{i=1}^{N}P_{ij}(y_{ij} - \bar{y})\right)}{\left(\sum_{i=1}^{N}P_{ij}(T_{ij} - \bar{T})\right)} \\ &= \frac{\bar{y}_1 - \bar{y}_0}{T_1 - T_0}\end{aligned}$$

The standard choice for an IV in this context is to use some indicator of eligibility. However, often the rules governing participation in a health programme 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 programme implementation. For example, if the health programme is targeted to poor villages but at a centralized 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.

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 South Africa, 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, 2003).

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 eligibility requirements for participation can be verified and measured.

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. This 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 has not been affected by a 'time persistent' shock that may manifest as a treatment effect (Bertrand *et al.*, 2004). 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. The assumption that selection bias is unchanging over time may 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.

The difference-in-difference approach has been successfully used to provide estimates of the impact of a number of health-related 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.

3. Concerns

In this section we outline a variety of concerns and practical considerations which often plague programme evaluations. While many of the issues outlined below are by now quite well known to practitioners and policy makers, they are likely to be especially important when evaluating health programmes, so we provide a brief overview of the main issues at stake. We frame the discussion from the point of view of RCTs, but of course the issues discussed apply irrespective of whether the data are experimental, quasi-experimental or observational.

3.1 Selective Compliance and Attrition

Bias may be introduced owing to selective compliance with or attrition from the randomly assigned status.[6] This occurs when individuals assigned to the control group take deliberate action in order to attain the benefits of treatment.

Similarly, individuals who benefit from an intervention may be less likely to drop out of an evaluation study than those who do not, resulting 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 programme. 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. When differential attrition occurs, it is typically dealt with using 'intention to treat' models (Imbens and Angrist, 1994). Here, 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. 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.

3.2 Externalities

A second important consideration in critically evaluating impact estimates is the presence of externalities generated by the 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, thereby affecting the magnitude of the impact estimates. Failure to account for positive or negative externalities associated with the intervention may lead to an under- or over-estimate of the intervention impact respectively. Thus, the choice of observational unit should reflect likely spillover effects (Ravallion, 2008).

3.3 Ethical objections

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, which may be particularly salient in the case of life-saving health interventions, may be combined with political concerns over service delivery to the electorate. 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 either used quasi-experimental methods to study the effects of treatment on economic behaviour (Habyarimana *et al.*, 2010; Thirumurthy *et al.*, 2008) or 'encouragement' designs, where treatment is not withheld from individuals but 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 further possibility would be 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.

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 at scale.

4. 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 behaviour (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.

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, collecting data on intermediate outcomes such as school enrolment 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.

4.1 Nutritional supplementation

There is overwhelming and consistent evidence that malnutrition during the early years of a child's life, which manifests as stunting later on, is associated with lower cognitive levels and academic achievement, as well as higher dropout rates (Grantham-McGregor *et al.*, 2007). 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 (Partnership for Child Development, 1999; Moock and Leslie, 1986), and more likely to attain lower grades for their age (Moock and Leslie, 1986; Jamison, 1986; Clark *et al.*, 1990; Hutchinson *et al.*, 1997). 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 dropout 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.

Randomized 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; Pollitt, 1996). 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, 1996).

Of course, provision of food supplements in the form of school meals may yield additional benefits over and above nutritional benefits, such as improved school attendance. Vermeersch and Kremer (2004) find that participation was 30 per cent higher in Kenyan pre-schools where a free breakfast was introduced, than compared to control pre-schools where no such intervention occurred. Moreover, test scores were 0.4 standard deviations higher in treatment schools, although this was conditional on teacher qualifications.

Schultz (2004) examines the impact of the PROGRESA programme in Mexico, which was designed to allow for a phase-in of conditional cash transfers. PROGRESA provides cash grants, given to women, conditional on children attending school regularly and utilizing preventative health measures (health care visits, nutritional supplements and participation in health education programmes). Children in treatment households experienced significantly better health outcomes (Gertler and Boyce, 2001) and enrolment rates (Schultz, 2004) than children in control households.

To the extent that health gains in early childhood translate into better earnings potential as an adult, Behrman and Hoddinott (2000) estimate that exposure to the PROGRESA treatment will result in an increase of 2.9 per cent in lifetime earnings.

4.2 Iron Supplementation

Walker *et al.* (2007) estimate that 44–66 per cent of all children aged 4 and below in developing countries suffer from anaemia, with half of these cases being attributable to iron deficiencies. Iron deficiency holds negative consequences for child outcomes, including lower mental, social-emotional, motor and brain functioning than infants without such deficiency (Lozoff *et al*., 2000; Grantham-McGregor *et al*., 2007). Importantly though, iron treatment in pre-school aged children with iron deficiency anaemia has yielded positive cognitive benefits consistently over a number of studies (Grantham-McGregor *et al*., 2007; Sachdev *et al*., 2005). There have been a number of large-scale RCT trials on iron supplementation in infants or young children in developing countries, including Zanzibar (Stoltzfus *et al*., 2001), Chile (Lozoff *et al*., 2000), Bangladesh (Black *et al*., 2004), Indonesia (Lind *et al*., 2004) and India (Black *et al*., 2004). 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.

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 4,000 children aged 2–6 years, through an existing pre-school network. The results to date show that children in treatment schools gained significant weight (0.6 kg 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 per cent reduction in the incidence of severe to moderate anaemia.

The longer-term benefits of iron supplementation are less clear, mainly due to insufficient evidence. The large scale randomized 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.

There is also evidence that suggests that iron supplements have a large effect on productivity of adult workers. Basta *et al.* (1979) found increased work output among anaemic 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. Using difference-in-difference methods, 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 haemoglobin levels (Thomas *et al*., 2003).

4.3 Deworming

Illness due to worms is a problem that affects approximately one-third of the world's population, and the incidence of such infection is highest amongst school-aged children (Watkins and Pollitt, 1997). There are relatively few studies of the impact of worm infections on child development, especially 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 Simeon *et al.* (1995) 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.

Miguel and Kremer (2004) 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.

4.4 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 access to 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.*, 2010; Thirumurthy *et al.*, 2008). Habyarimana *et al.* (2010) 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 antiretroviral (ARV) treatment to workers covers up to a third of the cost of treatment. Using longitudinal survey data from Western Kenya, Thirumurthy *et al.* (2008) show that within six months of beginning ARV treatment, adult ARV recipients are 20 per cent 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.* (2008) 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.

5. What We Don't (Yet) have Much Evidence On

One of the biggest challenges in evaluation work is separating out results that establish attribution from results that merely speak to the fractional contribution of a given intervention, as opposed to the contribution of other (perhaps simultaneously pursued) interventions aimed at affecting the same outcomes as the intervention of interest. Studies that focus on very narrowly defined interventions are especially susceptible to this criticism. Even if attribution can be established through a randomized design, important questions pertaining to the external validity and the opportunity cost of a given intervention are often overlooked.

We conclude in this section by outlining a number of areas where further work is needed.

5.1 External Validity

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 *et al*., 1998b). Scaling up may also fail if the socio-economic composition of local participants differs from the national demographic profile. Randomized 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, 2008).

Concerns over external validity may be ameliorated to the extent that interventions are replicated in different settings and at different scales (Duflo and Kremer, 2005; Duflo, 2003). The results from these replication studies provide evidence on the extent to which results can be generalized. Similarly, replication of RCTs in different settings and at different scales could also make a significant contribution to understanding the role of institutions (historical and contemporary) in producing successful health interventions, as well as understanding the extent to which health interventions are generalizable across countries (Duflo and Kremer, 2005; Duflo, 2003). 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 impacts in different settings gives some sense of how generalizable 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 *et al.* (2002), despite the fact that the Indian programme was modified to include iron supplementation.

It is also possible to calibrate results by using alternative methodological approaches to re-estimate results from specific studies. A replication study of this type, for example, is one that uses data that has been gathered from an RCT. It then 'de-randomizes' assignment to the treatment group, so to speak, by discarding the experimentally assigned control group status, and instead uses a control group from another source, external to the study sample. Using cutting-edge statistical and econometric techniques, the researcher then goes about trying to see which of the available methods produces the least amount of bias. The actual 'bias factor' is known because the true impact of the programme can be gleaned from the experimental data directly. When one has two interventions in a related area, one that is randomized and another that is not, valuable lessons might be learnt by conducting a replication study, and then using the best non-experimental technique so revealed (i.e., the one that comes the closest in replicating the estimate of impact derived from the RCT) to open up possibilities for conducting quantitative impact evaluations on the types of interventions for which RCTs are not well suited. Building up an evidence base of this sort could dramatically alter the landscape of evidence-informed policy making, by putting studies not using RCTs on the same evaluation plain as those that do.

The evidence on whether matching methods and randomization methods produce the same results is somewhat mixed. Agodini and Dynarski (2004) find no consistent evidence that propensity score matching can replicate RCT results of school dropout programmes in the US. In contrast, work by Heckman *et al.* (1998a) and Diaz and Handa (2004) suggests that matching works well as long as the survey instrument used for measuring outcomes is identical for treatment and control participants. A recent study by Diaz and Handa (2004) shows that with the collection of a large number of observables, propensity score matching can approximate results from RCTs. Buddelmeyer and Skoufias (2004) use cutoffs in PROGRESA's eligibility rules to measure impacts of the programme and find that discontinuity design gives a good approximation for almost all outcome indicators when compared to estimates obtained through randomization.

5.2 Comparative Cost-Effectiveness and Efficiency

An obvious difficulty of thinking about how generalizable 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, policy makers 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. Simply put, a study focused on a single intervention is concerned with establishing attribution of that intervention alone. It does not reveal whether or not the money could have been better spent elsewhere.

5.3 Timing

Returns on investments in health often take a long time to realize and often these investments are made at early ages. There may be significant lags in outcome responses to a specific health intervention. Ideally, 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 years. 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 knowledge of long-term effects of early health investments.

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).

6. 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 modelling 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 (1995) 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 programme 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 programme 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 PROGRESA, 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 non-experimental 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.

Notes

1. The unit of analysis could of course be the individual.
2. Angrist (1998) provides a good illustration.
3. In principal, $E(y_1|D = 0, \mathbf{x})$could be estimated non-parametrically, but as is well known, convergence can be hard if **x** is of very large dimension. One workaround to this problem is to split the sample into mutually exclusive bins and match within these bins, but then one has to contend with empty cell problems. For example, if **x** contains just five covariates and the sample is split into quintiles, then we would need $5^5 = 3125$ control group observations. This number rises to 100,000 if one matches within deciles instead. See Cochran (1968) and Rosenbaum (2004) for more on this technique.
4. See for example, Frolich (2004), Zhao (2004) and Smith and Todd (2005).

5. Dehejia and Wahba (1999, 2002) have popularized this method to such an extent that it now rivals more traditional solutions to solving identification problems such as IVs. Imbens and Wooldridge (2009) provide an excellent summary of the details behind this approach. See also Ravallion (2008).

6. 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).

References

Agodini, R., and M. Dynarski (2004), 'Are Experiments the Only Option? A Look at Dropout Prevention Programs', *Review of Economics and Statistics*, Vol. 86, No. 1, pp. 180–94.

Angrist, J. (1990), 'Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records', *American Economic Review*, Vol. 80, pp. 313–35.

Angrist, J. (1998), 'Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants', *Econometrica*, Vol. 66, No. 2, pp. 249–88.

Basta, S., S. Soekirman, D. Karyadi and N. S. Scrimshaw (1979), 'Iron Deficiency Anemia and the Productivity of Adult Males in Indonesia', *American Journal of Clinical Nutrition*, Vol. 32, No. 4, pp. 916–25.

Beasley, N. M. R., A. Hall, and A. M. Tomkins (2000), 'The Health of Enrolled and Not Enrolled Children at School Age in Tanga, Tanzania', *Acta Tropica*, Vol. 76, pp. 223–29.

Behrman, J., and J. Hoddinott (2000), 'An Evaluation of the Impact of Progresa on Pre-school Child Height', International Food Policy Research Institute, Working Paper, July.

Berkman, D. S., A. G. Lescano, R. H. Gilman, S. L. Lopez and M. M. Black (2002), 'Effects of Stunting, Diarrhoeal Disease, and Parasitic Infection during Infancy on Cognition in Late Childhood: A Follow-up Study', *Lancet*, Vol. 359, pp. 296–300.

Bertrand, M., E. Duflo and S. Mullainathan (2004), 'How Much Should We Trust Differences-in-differences Estimates?' *The Quarterly Journal of Economics*, Vol. 119, No. 1, pp. 249–75.

Black, M. M., S. Sazawal, R. F. 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*, Vol. 113, pp. 1297–305.

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

Buddelmeyer, H. and E. Skoufias (2004), 'An Evaluation of the Performance of Regression Discontinuity Design on Progresa', Policy Research Working Paper Series 3386, The World Bank, September.

Clark, N., S. Grantham-McGregor and C. Powell (1990), 'Health and Nutrition Predictors of School Failure in Kingston, Jamaica', *Ecological Food Nutrition*, Vol. 26, pp. 1–11.

Cochran, W. G. (1968), 'The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies', *Biometrics*, Vol. 24, pp. 205–13.

Dehejia, R. H. and S. Wahba (1999), 'Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs', *Journal of the American Statistical Association*, Vol. 94, No. 448, pp. 1053–62.

Dehejia, R. H., and S. Wahba (2002), 'Propensity Score Matching Methods for Nonexperimental Causal Studies', *Review of Economics and Statistics*, Vol. 84, pp. 151–61.

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 Annual Bank Conference in Development Economics in Bangalore.

Duflo, E. and M. Kremer (2005), 'Use of Randomization in the Evaluation of Development Effectiveness', in O. Feinstein, G. Pitman and G. Ingram (eds.), *Evaluating Development Effectiveness*, Transaction Publishers, New Brunswick, NJ.

Duflo, E., R. Glennerster and M. Kremer (2008), 'Using Randomization in Development Economics Research: A Toolkit', in T. P. Schultz and J. A. Strauss (eds.), *Handbook of Development Economics*, vol. 4 ch. 61, pp. 3895–962, Elsevier North Holland, Amsterdam.

Frolich, M. (2004), 'Finite Sample Properties of Propensity Score Matching and Weighting Estimators', *Review of Economics and Statistics*, Vol. 86, No. 1, pp. 77–90.

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*, Vol. 113, No. 1, pp. 83–119.

Gertler, P. J., and S. Boyce (2001), 'An Experiment in Incentive-based Welfare: The Impact of Progresa on Health in Mexico', University of California, Berkeley, CA.

Grantham-McGregor, S. M., C. A. Powell, S. P. Walker and J. H. Himes (1991), 'Nutritional Supplementation, Psychosocial Stimulation, and Mental Development of Stunted Children: The Jamaican Study', *Lancet*, Vol. 338, No. 8758, pp. 1–5.

Grantham-McGregor, S., Y. B. Cheung, S. Cueto, P. Glewwe, L. Richter and B. Strupp (2007), 'Developmental Potential in the First 5 Years for Children in Developing Countries', *Lancet*, Vol. 369, No. 9555, pp. 60–70.

Habyarimana, J., B. Mbakile and C. Pop-Eleches (2010) 'The Impact of HIV/AIDS and ARV Treatment on Worker Absenteeism: Implications for African Firms', *Journal of Human Resources*, Vol. 45, No. 4, pp. 809–39.

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 Semiparametric Estimation of Average Treatment Effects', *Econometrica*, Vol. 66, No. 2, pp. 315–31.

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*, Vol. 64, 605–54.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998a), 'Characterizing Selection Bias Using Experimental Data', *Econometrica*, Vol. 66, pp. 1017–98.

Heckman, J., and J. Smith (1995), 'Assessing the Case for Social Experiments', *Journal of Economic Perspectives*, Vol. 9, No. 2, pp. 85–110.

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

Hutchinson, S. E., C. A. Powell, S. P. Walker, S. M. Chang and S. M. Grantham-McGregor (1997), 'Nutrition, Anaemia, Geohelminth Infection and School Achievement in Rural Jamaican Primary School Children', *European Journal of Clinical Nutrition*, Vol. 51, No. 11, pp. 729–35.

Imbens, G., and J. Angrist (1994), 'Identification and Estimation of Local Average Treatment Effects', *Econometrica*, Vol. 62, No. 2, pp. 467–75.

Imbens, G. W., and J. M. Wooldridge (2009), 'Recent Developments in the Econometrics of Program Evaluation', *Journal of Economic Literature*, Vol. 47, No. 1, pp. 5–86.

Jalan, J., and M. Ravallion (1998), 'Are There Dynamic Gains from a Poor-area Development Program', *Journal of Public Economics*, Vol. 67, No. 1, pp. 65–86.

Jamison, D. T. (1986), 'Child Malnutrition and School Performance in China', *Journal of Development Economics*, Vol. 20, pp. 299–309.

Lanjouw, P., and M. Ravallion (1999), 'Benefit Incidence and the Timing of Program Capture', *World Bank Economic Review*, Vol. 13, No. 2, pp. 257–74.

Lind, T., B. Lönnerdal, H. Stenlund, I. L. Gamayanti, D. Ismail, R. Seswandhana and L.-A. Persson (2004), 'A Community-based Randomized Controlled Trial of Iron and Zinc Supplementation in Indonesian Infants: Effects on Growth and Development', *American Journal of Clinical Nutrition*, Vol. 80, No. 3, pp. 729–36.

Lozoff, B., E. Jimenez, J. Hagen, E. Mollen and A. W. Wolf (2000), 'Poorer Behavioral and Developmental Outcome More than 10 years after Treatment for Iron Deficiency in Infancy', *Pediatrics*, Vol. 105, No. 4, pp. e51.

Miguel, E., and M. Kremer (2004), 'Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities', *Econometrica*, Vol. 72, No. 1, pp. 159–217.

Moock, P. R., and J. Leslie (1986), 'Child Malnutrition and Schooling in the Terai Region of Nepal', *Journal of Development Economics*, Vol. 20, pp. 33–52.

Oberhelman, R. A., E. S. Guerrero, M. L. Fernandez, M. Silio, D. Mercado, N. Comiskey, G. Ihenacho, and R. Mera (1998), 'Correlations between Intestinal Parasitosis, Physical Growth, and Psychomotor Development among Infants and Children from Rural Nicaragua', *American Journal of Tropical Medicine and Hygiene*, Vol. 58, No. 4, pp. 470–75.

Partnership for Child Development, The (1999), 'Short Stature and the Age of Enrolment in Primary School: Studies in Two African Countries', *Social Science and Medicine*, Vol. 48, No. 5, pp. 675–82.

Pollitt, E. (1996), 'Timing and Vulnerability in Research on Malnutrition and Cognition', *Nutrition Reviews*, Vol. 54, No. 2, Pt 2, pp. S49–S55.

Ravallion, M. (2008), 'Evaluating Anti-poverty Programs', in R. E. Evenson and T. P. Schultz (eds.), *Handbook of Development Economics*, Vol. 4, Elsevier North-Holland, Amsterdam, pp. 3787–846.

Rosenbaum, P. R. (2004), 'Matching in Observational Studies', in Andrew Gelman and Xiao-Li Meng (eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Wiley, Chichester, pp. 15–24.

Rosenbaum, P. R., and D. Rubin (1983), 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika*, Vol. 70, No. 1, pp. 41–55.

Sachdev, H. P. S., T. Gera and P. Nestel (2005), 'Effect of Iron Supplementation on Mental and Motor Development in Children: Systematic Review of Randomised Controlled Trials', *Public Health Nutrition*, Vol. 8, No. 2, pp. 117–32.

Schultz, T. P. (2004), 'School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program', *Journal of Development Economics*, Vol. 74, No. 1, pp. 199–250.

Simeon, D. T., S. M. Grantham-McGregor, J. E. Callender and M.S. Wong (1995), 'Treatment of Trichuris Trichiura Infections Improves Growth, Spelling Scores and School Attendance in Some Children', *Journal of Nutrition*, Vol. 125, pp. 1875–83.

Smith, J., and P. Todd (2005), 'Does Matching Overcome Lalonde's Critique of NX Estimators', *Journal of Econometrics*, Vol. 125, No. 12, pp. 305–53.

Stoltzfus, R. J., J. D. Kvalsvig, H. M. Chwaya, A. Montresor, M. Albonico, J. M. Tielsch, L. Savioli and E. Pollitt (2001), 'Effects of Iron Supplementation and Anthelmintic Treatment on Motor and Language Development of Preschool Children in Zanzibar: Double Blind, Placebo Controlled Study', *British Medical Journal*, Vol. 323, No. 7326, pp. 1389–93.

Thirumurthy, H., J. G. Zivin and M. Goldstein (2008), 'The Economic Impact of Aids Treatment: Labor Supply in Western Kenya', *Journal of Human Resources*, Vol. 43, No. 3, pp. 511–52.

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 Com- Petition: A Randomized Evaluation', World Bank Policy Research Paper 3523, Washington, DC.

Waber, D. P., L. Vuori-Christiansen, N. Ortiz, J. R. Clement, N. E. Christiansen, J. O. Mora, R. B. Reed and M. G. Herrera (1981), 'Nutritional Supplementation, Maternal Education, and Cognitive Development of Infants at Risk of Malnutrition', *American Journal of Clinical Nutrition*, Vol. 34 (Suppl 4), pp. 807–13.

Walker, S. P., S. M. Chang, C. A. Powell and S. M. Grantham McGregor (2005), 'Effects of Early Childhood Psychosocial Stimulation and Nutritional Supplementation on Cognition and Education in Growth-stunted Jamaican Children: Prospective Cohort Study', *Lancet*, Vol. 366, pp. 1804–807.

Walker, S. P., T. D. Wachs, J. Meeks Gardener, B. Lozoff, G. A. Wasserman, E. Pollitt and J. A. Carter (2007), 'Child Development: Risk Factors for Adverse Outcomes in Developing Countries', *Lancet*, Vol. 369, pp. 145–57.

Watkins, W. E., and E. Pollitt (1997), ' "Stupidity or Worms": Do Intestinal Worms Impair Mental Performance?' *Psychology Bulletin*, Vol. 121, No. 2, pp. 171–91.

Zhao, Z. (2004), 'Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence', *The Review of Economics and Statistics*, Vol. 86, No. 1, pp. 91–107.