Review of Educational Research
September 2015, Vol. 85, No. 3, pp. 353–394
DOI: 10.3102/0034654314553127

Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments

Patrick J. McEwan
Wellesley College

I gathered 77 randomized experiments (with 111 treatment arms) that evaluated the effects of school-based interventions on learning in developing-country primary schools. On average, monetary grants and deworming treatments had mean effect sizes that were close to zero and not statistically significant. Nutritional treatments, treatments that disseminated information, and treatments that improved school management or supervision, had small mean effect sizes (0.04–0.06) that were not always robust to controls for study moderators. The largest mean effect sizes included treatments with computers or instructional technology (0.15); teacher training (0.12); smaller classes, smaller learning groups within classes, or ability grouping (0.12); contract or volunteer teachers (0.10); student and teacher performance incentives (0.09); and instructional materials (0.08). Metaregressions suggested that the effects of contract teachers and materials were partly accounted for by composite treatments that included training and/or class size reduction. There are insufficient data to judge the relative cost-effectiveness of categories of interventions.

Keywords: meta-analysis, randomized experiment, school effectiveness, learning, developing countries

There is a vast nonexperimental literature that analyzes student achievement in primary schools of developing countries (for influential reviews, see Fuller & Clarke, 1994; Glewwe, 2002; Hanushek, 1995; Lockheed & Verspoor, 1991; Velez, Schiefelbein, & Valenzuela, 1993). This literature provided rich descriptions of developing-country school systems, cataloged inequalities in the distribution of resources and learning, inspired theoretical work across the social sciences, and catalyzed new empirical research (including the experiments reviewed in this meta-analysis). It also confronted two empirical challenges. First, regression analysis with nonexperimental data could not always distinguish between the causal effects of schools and the confounding effects of the children and families that happen to attend those schools (Glewwe, 2002; Glewwe, Kremer, Moulin, & Zitzewitz, 2004). Second, many empirical studies used proxies of school quality,

such as teacher credentials and pupil-teacher ratios, that did not encompass the wider menu of investment choices available to policymakers.

A growing number of randomized, controlled experiments have addressed both challenges. Random assignment of students or schools to school-based treatments improves the internal validity of causal inferences (Duflo, Glennerster, & Kremer, 2008; Glewwe & Kremer, 2006). Moreover, researchers have evaluated policy-relevant treatments that encompass (a) instructional interventions that incorporate teacher training, textbooks, computers and technology, and/or changes in the size and composition of classes; (b) school-based health and nutrition interventions, such as deworming, school meals, and micronutrient supplementation; and (c) interventions that modify stakeholder incentives to improve learning, such as information dissemination, student or teacher performance incentives, flexible teacher contracts, and reforms that affect school management and supervision.

I conducted a literature search in economics, education, and public health, identifying 77 published and unpublished experiments that include 111 treatment arms. Researchers randomly assigned children, schools, or entire villages to receive a school-based treatment, versus "business as usual" in the similar settings. I initially sought to identify studies that used a regression-discontinuity design, since well-designed studies have strong internal validity (Lee & Lemieux, 2009; What Works Clearinghouse, 2011), but only a handful of papers fulfilled the criteria for sample inclusion (Chay, McEwan, & Urquiola, 2005; McEwan, 2013; Urquiola, 2006). I coded effect sizes and standard errors for language and mathematics outcomes. I further coded study attributes that describe features of the treatment, the experimental context and sample, the outcome measures, and the study quality.

Two categories of interventions—monetary grants and school-based deworming—have mean effect sizes that are close to zero and not statistically significant (based on random effects models). School-based nutritional treatments, treatments that provide information to parents or students, and treatments that improve school management and supervision tend to have small mean effect sizes—from 0.04 to 0.06 standard deviations—that are not always robust to controls for study moderators in metaregressions. The largest average effect sizes are observed for treatments that incorporate instructional materials (0.08); computers or instructional technology (0.15); teacher training (0.12); smaller classes, smaller learning groups within classes, or ability grouping (0.12); contract or volunteer teachers (0.10); and student and teacher performance incentives (0.09). However, it bears emphasis that the categories are not mutually exclusive. Metaregressions that control for treatment heterogeneity and other moderators suggest that the effects of materials and contract teachers, in particular, are partly accounted for by overlapping treatments. For example, instructional materials have few effects on learning in the absence of teacher training (e.g., Glewwe et al., 2004; Glewwe, Kremer, & Moulin, 2009), and contract and volunteer teacher interventions overlap with class size reduction or other instructional treatments (e.g., Banerjee, Cole, Duflo, & Linden, 2007; Bold, Kimenyi, Mwabu, Ng'ang'a, & Sandefur, 2012). The metaregression estimates are surprisingly robust to controls for other moderators.

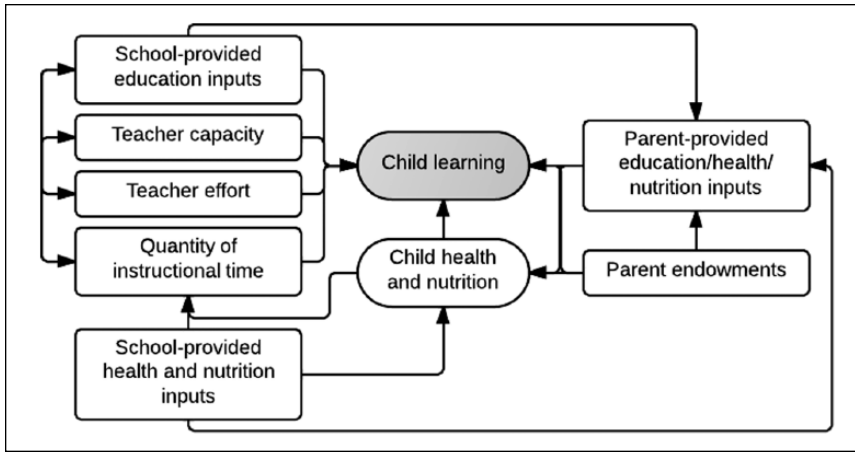

FIGURE 1. *The direct and indirect determinants of child learning.*

Most experiments contain minimal data cost on costs, complicating an assessment of whether specific treatments in the meta-analytic sample—or categories of treatments—are relatively more cost-effective despite smaller effect sizes (or less so despite larger ones). As an alternative, I combine effect sizes with auxiliary cost estimates for 15 treatment arms that are analyzed in Kremer, Brannen, and Glennerster (2013). The results suggest that some interventions are relatively less cost-effective than others, such as computer-assisted instruction in India and class size reduction in Kenya. However, the conclusions are tempered by the small samples and the inability to statistically distinguish between ranked cost-effectiveness ratios (CERs).

Method

Production Functions and Policy Effects

The field of development economics couches experimentation in the framework of the education production function (Glewwe & Kremer, 2006; Glewwe & Miguel, 2008). Figure 1 illustrates a stylized production function for child learning, in which learning is directly influenced by four groups of variables: (a) parent endowments such as schooling and ability, (b) parent-provided education inputs like supplemental instruction, (c) child endowments such as nutrition and health but potentially including a wider range of cognitive or noncognitive abilities, and (d) school and teacher inputs.

Within the fourth category, school-provided inputs may include textbooks and related instructional materials, computers and software, and school equipment and facilities. Teacher capacity denotes a teacher's ability to deliver or facilitate classroom instruction. Capacity itself may be influenced by preservice training in pedagogy or content, by in-service training and experience, or by innate talents. Teacher effort is the intensity or time devoted to lesson plan preparation, classroom instruction, or other activities directly related to learning. It may be influenced by extrinsic

incentives (e.g., rewards or sanctions for poor performance) or intrinsic ones (e.g., a desire to improve student learning). Finally, the quantity of instructional time is a function of the mandated number of instructional days, length of the school day (including after-school instruction), the proportion of time devoted to learning-related activities, and the local decisions of households and teachers.

In early research, researchers estimated the parameters of production functions using nonexperimental data and regression analysis (Fuller & Clarke, 1994; Glewwe, 2002; Hanushek, 1995; Velez et al., 1993). They typically regressed a measure of child learning on variables such as mother's schooling, the availability of textbooks, and teachers' years of experience and education credentials. Reviewers cautioned against potential biases in the estimates of key parameters (Glewwe, 2002; Glewwe & Kremer, 2006). Suppose, for example, that schools with higher income families also tend to have greater endowments of textbooks and that higher levels of either variable increase student achievement, all else equal. Further suppose that income is imperfectly measured, or even omitted from the regression analysis. In this case, the textbook effect is plausibly biased upward (Glewwe et al., 2004).

The random assignment of an education or health intervention strengthens causal interpretations, since child and household variables will be balanced across treatment and control groups, on average. However, experimental treatment effects rarely have a straightforward interpretation as production function parameters. In most experiments, the effects are policy (or reduced-form) effects rather than structural estimates of production function parameters, since they encompass the direct and indirect effects of interventions on learning (Glewwe & Kremer, 2006; Todd & Wolpin, 2003).

In Figure 1, suppose that the provision of school meals improves child nutrition or health, which directly improves learning. Meals are an in-kind transfer to children, conditional on regular school attendance, so they may also spur attendance and affect learning by increasing the quantity of instructional time. Households could react to the availability of free lunches by reallocating food within the household, perhaps toward needier siblings (H. G. Jacoby, 2002). In this example, the policy effect of a school-based intervention potentially encompasses multiple direct or indirect effects on learning. This meta-analysis summarizes a broad range of policy effects but cannot disentangle the causal mechanisms of those effects.

Literature Search

I conducted a literature search between August 2012 and February 2013, first examining the references of two meta-analyses of randomized and nonrandomized evaluations of education interventions in developing countries (Glewwe, Hanushek, Humpage, & Ravina, 2011; Petrosino, Morgan, Fronius, Tanner-Smith, & Boruch, 2012). Each review employed a keyword search of scholarly databases such as Econlit, Eric, and Medline, although Petrosino et al. focused on studies with at least one attainment outcome (e.g., enrollment), thereby excluding studies focusing exclusively on learning outcomes. I next examined narrative reviews of the education production function literature in developing countries (Evans & Ghosh, 2008; Fuller & Clarke, 1994; Glewwe, 2002; Glewwe & Kremer, 2006;

Glewwe & Miguel, 2008). Two of the most recent emphasize randomized impact evaluations (Kremer et al., 2013; Kremer & Holla, 2009). Finally, I searched *Dissertation Abstracts* using the keywords randomized (or randomised), experiment(s), and school(s).

To increase coverage of school-based health and nutrition interventions, I consulted meta-analyses and narrative reviews in nutrition and public health. The treatments included deworming medications (Dickson, Awasthi, Williamson, Demellweek, & Garner, 2000; Taylor-Robinson, Maayan, Soares-Weiser, Donegan, & Garner, 2012), iron supplementation (Falkingham et al., 2011; Grantham-McGregor & Ani, 2001; Hermoso et al., 2011), multiple micronutrient supplementation (Best et al., 2011; Eilander et al., 2010), malaria medications (S. D. Fernando, Rodrigo, & Rajapakse, 2010), and school feeding programs (Jomaa, McDonnell, & Probart, 2010; Kristjansson et al., 2006). Finally, I consulted World Bank reports on learning in developing countries (Bruns, Filmer, & Patrinos, 2011; Vegas & Petrow, 2008).

To maximize the coverage of unpublished research, I searched the websites and working papers of the Abdul Latif Jameel Poverty Action Lab at MIT; the Center for Effective Global Action at the University of California, Berkeley; Innovations for Poverty Action; the Inter-American Development Bank; the National Bureau of Economic Research; RTI International; the Rural Education Action Program at Stanford University; and the World Bank. I examined the results of keyword searches for randomized, randomised, or random assignment. I applied the same keyword searches to the *American Economic Review, American Economic Journal: Applied Economics, Comparative Education Review, Economic Development and Cultural Change, Journal of Development Economics, Journal of Development Effectiveness*, and the *International Journal of Educational Development.*

Criteria for Study Inclusion and Exclusion

I included studies if they (a) were conducted in a low- to upper middle-income country, as defined by the World Bank in 2012[1]; (b) were conducted in primary schools, broadly defined to include Grades 1 to 8 (or ages 6 to 14, if the grades were not reported); (c) randomly assigned children (or clusters of children) to an education or health intervention in a school setting, or "business-as-usual" in the same setting; (d) reported results for at least one continuously measured learning outcome in language or reading, mathematics, or a composite assessment including either outcome; and (e) reported sufficient data to calculate the treatment's effect size and standard error, in the full experimental sample.

After identifying the initial sample of studies, I excluded studies if they did not meet at least one of the criteria. Several studies were conducted in preschool grades (He, Linden, & MacLeod, 2009; Jukes et al., 2006) or secondary grades (Angrist, Bettinger, Bloom, King, & Kremer, 2002; Blimpo, 2010; Liu et al., 2013). Some studies did not randomly assign units to treatment and control groups (Heyneman, Jamison, & Montenegro, 1984; Inamdar, 2004; Nitsaisook & Anderson, 1989; Rosas et al., 2003). Some authors used a method of quasi-random assignment, such as alternating treatment-control assignment from an alphabetized list of clusters (Miguel & Kremer, 2004). I included these studies but coded the study attribute for

subsequent analysis in metaregressions. Berry (2012) compared two incentive treatments but did not include a pure control group. Banerjee, Banerji, Duflo, Glennerster, and Khemani (2010) report results for binary indicators of learning outcomes, whereas the outcome measures in several papers did not include language or mathematics (Beuermann, Cristia, Cruz-Aguayo, Cueto, & Malamud, 2012; Clarke et al., 2008; Kvalsvig, Cooppan, & Connolly, 1991; Lien et al., 2009; Newman et al., 2002; Seshadri & Gopaldas, 1989).

A remaining set of studies met the previous criteria but did not report sufficient data to estimate effect sizes and/or standard errors. Several studies did not report sufficient data to estimate the mean difference between treatment and control groups at each follow-up (Pollitt, Hathirat, Kotchabhakdi, Missell, & Valyasevi, 1989; Sungthong, Mo-suwan, Chongsuvivtwong, & Geater, 2004; Whaley et al., 2003), in one case because statistically nonsignificant results were not reported in the paper (Nga et al., 2011). In other cases, I could not convert mean differences (or a regression coefficient estimating a similar parameter) to effect sizes, given the lack of data on the standard deviation of the outcome variable (Adelman, Alderman, Gilligan, & Lehrer, 2008; Kazianga, de Walque, & Alderman, 2012; Pandey, Goyal, & Sundararaman, 2009; Vazir, Nagalla, Thangiah, Kamasamudram, & Bhattiprolu, 2006). Finally, I excluded cluster-randomized experiments in which standard errors did not correctly account for the unit of assignment (Chandler, Walker, Connolly, & Grantham-McGregor, 1995; Lai, Zhang, Hu, et al., 2012; Piper & Korda, 2011).

Coding of Experiments

Experiments and Papers

I defined six groups of variables that describe experiments, papers, treatment arms, follow-ups, outcome measures, and effects. For coding purposes, I defined a single experiment as one or more treatment arms and the single control group against which they are contrasted. I coded variables that are shared across experiments, including the random assignment procedure, the size of the control group, and the dates of baseline data collection. The modal experiment consists of one control group and one treatment arm, with results reported in a single paper. For example, Watkins, Cruz, and Pollitt (1996) randomly assigned 125 primary-grade students in Guatemala to receive deworming medication, and 125 to receive a placebo.

Sometimes one experiment is reported in multiple papers. Muralidharan and Sundararaman (2011) randomly assigned 500 Indian schools to four treatment arms and a control group. The cited article includes results on two performance incentive treatments, whereas Muralidharan and Sundararaman (2010a) and Das et al. (2011) report evidence on a contract teacher treatment and a block grant treatment, respectively. A fourth article posits that the control group in the aforementioned experiment is itself an informational treatment, since students were tested and teachers received this performance feedback (Muralidharan & Sundararaman, 2010b). The authors randomly selected a separate control group of schools that were not tested until follow-up. I code this as a separate experiment, although analyses account for statistical dependencies across the effect sizes from the two experiments, given the shared samples.

In cases where results from a single experiment are replicated in more than one paper, I used a preferred set of estimates. In one experiment, for example, the putative baseline occurred several months after the start of the treatment. I report estimates from Glewwe and Maïga (2011), which treats the baseline as a follow-up, in contrast to Lassibille, Tan, Jesse, and Nguyen (2010). Conversely, a single paper sometimes reports results of more than one experiment, conducted in different sites or time periods but usually on similar treatments (Banerjee, Banerji, Duflo, & Walton, 2012; Banerjee et al., 2007; He, Linden, & MacLeod, 2008). Pradhan et al. (2011) randomly assigned 520 school committees to a control group and eight treatments: grants, grants and training, grants and elections of committee members, grants and village–committee linkages, and four combinations thereof. The paper reports the grant/control contrast but otherwise discards the pure control group, and reports six contrasts in which grant receipt is balanced across the two groups of schools. I code this as seven experiments, given the varying composition of each control group.

Treatment Arms

Each experiment contained at least one treatment arm and sometimes as many as seven. The coded attributes of treatment arms included sample sizes, treatment duration, the implementing agency (whether a government, nongovernmental organization [NGO], or researcher), and the location of treatments within a typology of school-based treatments. The typology included three main categories, and sub-categories within each. The categories are not mutually exclusive, and they do not exhaustively describe potential school-based treatments, since they reflect the preferences and constraints of experimental researchers.

The first category includes treatments that endow schools with monetary grants, instructional materials such as textbooks, computers or other instructional technology, and teacher training. It also includes interventions that manipulate the size or composition of learning groups within schools, via class size reduction, small-group instruction, or ability group tracking. The second category includes health and nutrition treatments administered in schools, such as iron and micronutrient supplements, school-provided meals or beverages, and deworming or malaria prevention drugs. A residual health category provides diverse treatments such as vision screening and menstrual cups.

The third category of treatments modifies incentives for students, parents, or school personnel to improve student learning. Some treatments disseminate information on student performance to teachers or school officials, to school management committees or parents, or directly to students, often via a report card.[2] Other treatments link student or teacher rewards to performance measures based on teacher attendance, student test scores, or student health. Many treatments encourage the recruitment and hiring of teachers with flexible labor contracts, often locally hired contract teachers outside the civil service. In other cases, teachers are hired and trained by NGOs or work as volunteers.[3] Finally, a diffuse subcategory of treatments attempts to improve the management and supervision of schools by providing training to school officials or local school committees in management and in the hiring, monitoring, and assessment of teacher performance.

Follow-Ups and Outcome Measures

Each experiment conducts as many as three follow-up data collections. I coded variables on each follow-up, including the date of data collection and attrition from the baseline sample at the time of follow-up. Most experiments report at least two outcomes, although some report as many as five. It is common for experiments to report results for a main assessment, in addition to subtests consisting of items within the main assessment. In these cases, I only coded the main outcome measure, unless the paper reports only subtest results. I further coded whether the outcome measures language or reading, mathematics, or a composite, as well as the source of assessment items (whether a government or school exam, an off-the-shelf commercial assessment, or an evaluator-designed test).

Coding of Effect Sizes

I calculated at most two effect sizes for each unique combination of experiment, treatment arm, follow-up, and outcome. I refer to the first as an unconditional effect, in that it is the unconditional mean difference between the treatment and control group (or that it controls, at most, for dummy variables indicating experimental strata). I refer to the second as a conditional effect. It is usually obtained from a least squares regression that controls for variables that are plausibly unaffected by the treatment, such as a pretest.

Unconditional Effect Sizes

The literature on meta-analysis emphasizes that effects for continuous variables (e.g., test scores) should be expressed in comparable units. The most common is the effect size, often called Cohen's *d* (Borenstein, 2009). It is simply the mean difference in the outcome (*Y*)—measured at the follow-up—between treatment and control groups, divided by the sample standard deviation of the pooled treatment and control samples: $d = \frac{\bar{Y}_T - \bar{Y}_C}{s_{\text{pooled}}}$. The samples used to calculate the means include all members of the original treatment and control groups, regardless of their eventual exposure to the treatment. This is commonly referred to as an intention-to-treat estimate (Duflo et al., 2008). Its standard error can be calculated as $SE_d = \sqrt{\frac{n_T + n_C}{n_T n_C} + \frac{d^2}{2(n_T + n_C)}}$, where the additional terms are the student sample sizes in treatment and control groups. I also apply a small-sample correction to *d* and its standard error, resulting in Hedges's *g* (Borenstein, 2009). In practice, the correction makes no difference in this article's results.

In randomized experiments conducted by economists, it is common that authors report effects based on the linear regression:

$$O_{ij} = \beta_0 + \beta_1 T_{ij} + \varepsilon_{ij}, \quad (1)$$

where O_{ij} is the outcome of student *i* in school *j*, T_{ij} is a dummy variable indicating assignment to the treatment group (vs. the control), and β_1 represents the

mean difference between treatment and control groups, also interpreted as the effect of the intention-to-treat. If the dependent variable is expressed as a *z* score, with mean zero and standard deviation one, then $\hat{\beta}_1$ is a handy estimator of the effect size. Otherwise, one can divide the estimated effect size by a pooled standard deviation reported in the article. In cases where some of the treatment group refused treatment and/or some of the control group obtained it anyway, it is common to report instrumental variable estimates of the local average treatment effect on those whose treatment status was influenced by random assignment to the treatment group (Duflo et al., 2008). I excluded one paper because it reports only instrumental variable estimates (Evans, Kremer, & Ngatia, 2009).

It is common to stratify the units of assignment—whether students or schools—by pretreatment characteristics of the sample, such as location or poverty. Then, random assignment occurs within each stratum (Bruhn & McKenzie, 2009; Duflo et al., 2008). Sometimes authors form pair-wise matches across all units (i.e., multiple pairs of observably-similar students), and then randomly assign one unit to the treatment within each pair. The goal is to ensure that treatment and control groups are balanced on stratifying variables, thus increasing the precision of estimated effects. Bruhn and McKenzie (2009) show that Equation (1) yields overly conservative standard errors in the presence of stratification or pairwise matching. A more suitable specification would control for dummy variables indicating strata or pairs:

$$O_{ijk} = \beta_0 + \beta_1 T_{ijk} + \delta_k + \varepsilon_{ij}, \tag{2}$$

where δ_k represent separate intercepts for each stratum or pair *k*.

In Equations (1) and (2), authors typically calculate standard errors that take account of the unit of randomization. In the case of student assignment, authors usually report heteroskedasticity-consistent, Huber–White standard errors. In cluster-randomized experiments, researchers usually report cluster-adjusted Huber–White standard errors or apply an alternative, such as generalized least squares with a group random effect.

In this article, I code estimates and standard errors from Equation (2) when the experiment employs stratified or pairwise randomization. If the dependent variable is not already a *z* score, I divide the treatment effect and its standard error by the pooled standard deviation of O_{ijk}. Some authors standardize the dependent variables by the mean and standard deviation of the control group but do not report sufficient data to calculate the pooled standard deviation (Barrera-Osorio & Linden, 2009). In lieu of a better alternative, I code the regression coefficient and its standard error. For remaining cluster-randomized experiments, I use the coefficient estimate and standard error from Equation (1), dividing both by the pooled standard deviation of the dependent variable. For the remaining studies—all student-level randomized experiments in the nutrition and medical literature—I estimate Hedges's *g* and its standard error, using group-specific means and sample sizes, as well as the pooled standard deviation.

I code effects based on the full experimental sample, rather than subgroups defined by baseline achievement, geography, or other variables. Sometimes this leads me to prefer estimates different from those emphasized by authors. In a

Jamaican evaluation of school breakfast, for example, I calculate Hedges's *g* in the full experimental sample, aided by descriptive statistics that are disaggregated by the nutritional status of children (Powell, Walker, Chang, & Grantham-McGregor, 1998). In Kremer, Miguel, and Thornton (2009), a Kenyan evaluation of merit scholarships for girls, I include the full-sample effects but not effects disaggregated by district.

Conditional Effect Sizes

If students or schools are randomly assigned to treatment and control groups, then unconditional effects are unbiased. In practice, most authors also report estimates of the following regression:

$$O_{ijk} = \beta_0 + \beta_1 T_{ijk} + X_{ijk}\gamma + \varepsilon_{ij}, \quad (3)$$

where X_{ijk} is a vector of control variables that often includes a baseline pretest.

The main rationale for including control variables is that, in general, it reduces the standard error of the estimated treatment effect (Duflo et al., 2008). Indeed, the literature on power analysis emphasizes that the use of relevant covariates can reduce the minimum detectable effect size (MDES) in randomized experiments, all else equal (Dong & Maynard, 2013). On the other hand, controlling for a "kitchen sink" of covariates could increase standard errors if they do not explain variation in the dependent variable. A second rationale is that control variables adjust for imbalance in observed variables just after randomization, as might occur when a small number of students or clusters is randomized. Controls also adjust for imbalances in observed variables introduced after assignment by nonrandom attrition from the treatment and/or control groups. Deaton (2010) is less sanguine about the virtues of including controls, noting that it may encourage researchers to search over various regression specifications until the treatment is shown to "work," and that it could introduce biases, particularly in small samples, from the covariance between heterogeneity in treatment effects and the squares of included covariates.

I define an experiment as a single control group and one or more treatment arms. Thus, in an experiment with two treatment arms, two posttreatment follow-ups, and three outcomes, I potentially coded 12 unconditional and 12 conditional effect sizes (though many experiments do not report one or the other).

Statistical Analysis of Effect Sizes

The statistical analysis uses a single effect size per outcome, with a preference for the conditional effect size. This variable, $\hat{\theta}_{ijk}$, is equal to the *i*th effect size estimated in experiment *j* that is clustered within study *k*. Two or more experiments are defined as belonging to the same study if they have overlapping samples and/or identical instructional treatments. For example, three experiments on deworming medication used different samples of children within a common set of schools (Gardner, Grantham-McGregor, & Baddeley, 1996; Simeon, Grantham-McGregor, Callender, & Wong, 1995; Simeon, Grantham-McGregor, & Wong, 1995). He et al. (2008) report two experiments on similar instructional treatments implemented by an Indian NGO, conducted a year apart in different regions of India.

To estimate the mean effect size, I use a random effects model (Raudenbush, 2009; Ringquist, 2013). Suppose that one knows the true effect sizes (θ_{ijk}) of various treatments. One could estimate the following:

$$\theta_{ijk} = \theta + u_{ijk}, \quad u_{ijk} \sim N(0, \sigma_\theta^2), \tag{4}$$

where θ represents the mean effect size and u_{ijk} is a normally distributed error term that captures variation due to unobserved features of, say, treatments or samples. In fact, we observe an estimate of θ_{ijk}, such that

$$\hat{\theta}_{ijk} = \theta_{ijk} + e_{ijk}, \quad e_{ijk} \sim N(0, v_{ijk}). \tag{5}$$

The estimate, $\hat{\theta}_{ijk}$, has an estimation error with a zero mean and variance v_{ijk}. Substituting Equation (4) into (5) yields the following:

$$\hat{\theta}_{ijk} = \theta + e_{ijk} + u_{ijk}, \quad e_{ijk} + u_{ijk} \sim N(0, v_{ijk} + \sigma_\theta^2). \tag{6}$$

One can efficiently estimate θ as a weighted average of $\hat{\theta}_{ijk}$, applying inverse variance weights of $\frac{1}{v_{ijk} + \hat{\sigma}_\theta^2}$ (Ringquist, 2013). v_{ijk} is the square of the standard error of each effect size estimate, and $\hat{\sigma}_\theta^2$ is separately obtained with a restricted maximum likelihood estimator. I further adjust these weights so that some experiments do not exert undue influence on the mean simply because they measure more outcome variables or conduct more follow-ups. Specifically, I apply weights equal to $\frac{1}{v_{ijk} + \hat{\sigma}_\theta^2} \times \frac{1}{n_{ijk}}$, where n_{ijk} is equal to the number of effect sizes coded within experiment-by-treatment arm cells.

Ringquist (2013) suggests estimating Equation (6) by weighted least squares. Doing so facilitates the calculation of Huber–White standard errors that are clustered by study, as defined previously, to allow for statistical dependencies across effect sizes due to overlapping samples or treatments. It also permits the extended specification:

$$\hat{\theta}_{ijk} = \theta + X'_{ijk}\gamma + e_{ijk} + u_{ijk}, \tag{7}$$

where X_{ijk} is a vector of moderators that potentially explain variation in effect sizes. I use four categories of moderators that describe (a) subcategories of treatments, (b) country contexts and experimental samples, (c) outcome variables, and (d) the quality of experimental design and implementation.

Results

Descriptive Data

Experiments

Table 1 reports descriptive data on 77 experiments, divided by three categories of treatments. First, it confirms that the use of randomized experiments has grown

TABLE 1

Characteristics of experiments

	Instructional ($N = 39$), M (SD)	Health or nutrition ($N = 22$), M (SD)	Incentives ($N = 34$) M (SD)
Number of treatment arms per experiment	1.462 (0.85)	1.409 (0.80)	1.706 (1.27)
Number of follow-ups per experiment	1.359 (0.58)	1.273 (0.55)	1.353 (0.60)
Number of outcomes per experiment	2.051 (1.08)	2.091 (1.02)	2.059 (1.15)
Year			
Pre-1990s	0.026	0.091	0
1990s	0.103	0.591	0.058
Post-1990s	0.872	0.318	0.941
Region			
Africa	0.282	0.182	0.235
Latin America and Caribbean	0.128	0.318	0.029
East Asia and Pacific	0.282	0.409	0.382
South Asia	0.308	0.091	0.353
Gross domestic product per capita in baseline year, US$ (2000)	1,284 (1265)	2,012 (1491)	1,088 (1210)
Grades included at baseline			
Grades 1–4 only	0.564	0.318	0.559
Grades 5–8 only	0.077	0.091	0.118
Both	0.359	0.409	0.294
Uncertain	0	0.182	0.029
Published in			
Economics journal	0.205	0.182	0.265
Medical or nutrition journal	0	0.682	0
Psychology journal	0.026	0.046	0
Unpublished (working paper, report, etc.)	0.769	0.091	0.735
Convenience sample (vs. random sample)	0.769	0.864	0.706
Power analysis reported	0.179	0.409	0.206
Alternating list assignment (vs. randomized)	0.103	0.091	0.059
Cluster randomization (vs. student)	0.949	0.227	0.912
Stratification or pair-wise matching	0.667	0.818	0.618

Note. Each experiment consists of a control group and one or more treatment arms. See the text for details.

rapidly in the past decade (for further illustration, see Figure 2). More than two third of instructional and incentive experiments are still unpublished (in part

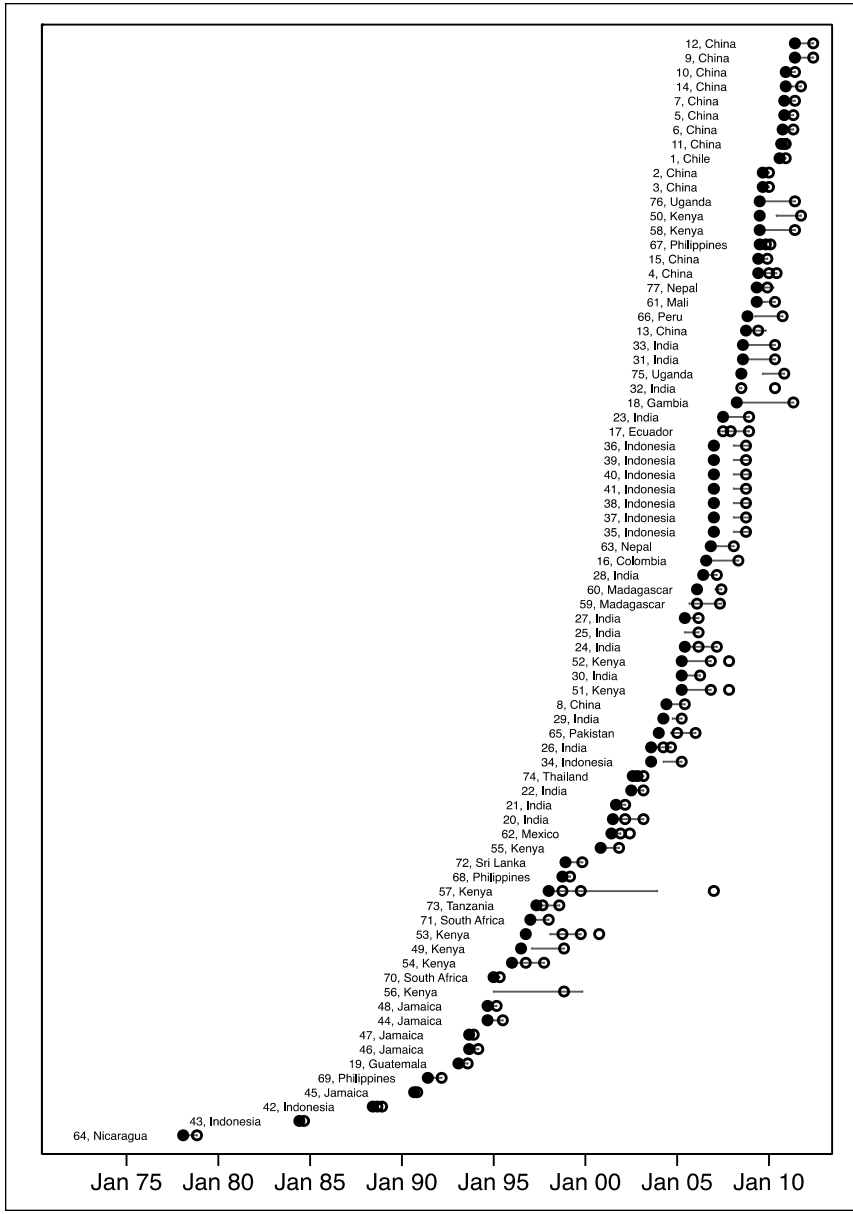

FIGURE 2. *Dates of treatment and data collection in 77 experiments.* Numerical codes identify experiments listed in the appendix available in the online journal. Solid circles indicate the date of baseline data collection in each experiment (defined as a control group and one or more treatment arms). Hollow circles indicate follow-up data collection(s), and lines indicate the duration of the treatment(s), in months.

because they are so recent), and most of the rest are published in economics journals. Impressionistically, economists have embraced school-based experimentation, whereas scholars in comparative and international education have not (for early exceptions, see Friend, Searle, & Suppes, 1980; Jamison, Searle, Galda, & Heyneman, 1981). The majority of health-related experiments are published in medical and nutrition journals.

Second, the smallest proportion of instructional (13%) and incentive experiments (3%) occur in Latin America, and none in the largest country of Brazil. This may reflect the preferences and funding of scholars affiliated with active research centers such as the Abdul Latif Jameel Poverty Action Lab. There may also be institutional constraints to conducting randomized experiments in public schools of relatively higher income developing countries.

Third, the majority of experiments begin with nonrandom, convenience samples of schools, often chosen because of geographic convenience, high poverty, or low achievement or because school officials consented to participate. Even Table 1 may understate the degree to which experimental samples are nonrepresentative of a large and well-defined population of children, since random samples are sometimes drawn from a population of schools whittled down by geographic and other exclusions (Kleiman-Weiner et al., 2013). A notable exception is a multiarm experiment that drew a representative sample of schools in the large Indian state of Andhra Pradesh (Muralidharan & Sundararaman, 2011). Whether convenience or random samples of schools, almost all of them include children from early primary Grades 1 to 4, and less than 10% focus exclusively on later grades.

Fourth, fewer than half of experiments mention that a power analysis guided the choice of sample size, especially in instructional and incentive experiments. Fifth, fewer than 10% of experiments use quasi-random assignment, such as the alternating selection of schools from alphabetized lists. Sixth, researchers employ stratification or pairwise matching in the majority of experiments, but only 43% appear to have actually controlled for strata or pair dummy variables. Bruhn and McKenzie (2009) find similar results in a broader sample of randomized experiments in development economics.

Treatment Arms

NGOs implement the majority of instructional and incentive treatments, university researchers are far more likely to administer health treatments, and experimentation in collaboration with governments is rarer in all cases (see Table 2). Instructional interventions often combine more than one type of input. For example, 45% of instructional treatments provide materials such as textbooks (Table 2). Of these, 79% also included teacher in-service training. Twenty-eight percent of instructional treatments use technology (Table 2), but 73% of those also involve training. Still, in some cases, materials or computers are provided directly to schools with little in the way of complementary inputs (Cristia, Ibarrarán, Cueto, Santiago, & Severín, 2012; Glewwe et al., 2009). Health and nutrition treatments mainly focus on the provision of micronutrients, school meals, or deworming medications. Malaria-related treatments are rarer, with only one included in the table's sample of studies (D. Fernando, de Silva, Carter, Mendis, & Wickremasinghe, 2006). Moreover, health and nutrition treatments are rarely

TABLE 2

Characteristics of treatment arms

	Instructional ($N = 53$) M (SD)	Health or nutrition ($N = 28$) M (SD)	Incentives ($N = 51$) M (SD)
Implementer			
Government	0.151	0.107	0.275
Nongovernmental organization	0.623	0.037	0.490
University or researcher	0.226	0.857	0.235
Instructional inputs			
Materials	0.453	0	0.176
Computer or technology	0.283	0	0.039
Grants	0.113	0	0.059
Teacher training	0.547	0	0.176
Class size, small-group instruction, tracking	0.226	0	0.196
Health inputs			
Food, beverage, and/or micronutrients	0	0.643	0.020
Deworming drugs	0	0.250	0
Malaria drugs	0	0.036	0
Other health inputs	0	0.071	0
Incentives			
Information for students, parents or schools	0.057	0	0.373
Performance incentives for students or schools	0.038	0	0.196
Contract or volunteer teachers	0.245	0	0.255
School management or supervision	0.094	0.036	0.275
For cluster-randomized experiments			
No. of clusters in treatment arm	51 (36)	20 (15)	66 (43)
No. of clusters in control	51 (30)	21 (10)	74 (53)
For student randomized experiments			
No. of students in treatment arm	295 (205)	171 (117)	461 (19)
No. of students in control	295 (205)	156 (111)	461 (19)

applied in concert with either instructional or incentive-based treatments (for an exception, see Sylvia et al., 2012). Treatments with informational components and performance incentives rarely include instructional inputs, in contrast to contract and volunteer teacher treatments. Among contract or volunteer teacher

TABLE 3

Characteristics of follow-ups and outcomes

	Instructional		Health or nutrition		Incentives	
	M (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*
Months of treatment exposure at follow-up	12.9 (8.8)	70	9.2 (11.1)	35	10.5 (6.9)	68
Follow-up conducted >1 month after treatment?	0.100	70	0.114	35	0.074	68
Attrition at follow-up (proportion)	0.199 (0.17)	52	0.088 (0.10)	29	0.136 (0.07)	44
Absolute value of differential attrition (proportion)	0.040 (0.06)	40	0.023 (0.03)	23	0.024 (0.02)	31
Content of outcome						
Language or reading	0.506	77	0.614	44	0.431	58
Mathematics	0.429	77	0.295	44	0.397	58
Composite score	0.065	77	0.091	44	0.172	58
Source of test items						
Evaluator or nongovernmental organization	0.416	77	0.114	44	0.310	58
Commercial or international	0.091	77	0.545	44	0.190	58
Government or school test	0.247	77	0.318	44	0.379	58
Uncertain	0.247	77	0.023	44	0.121	58

treatments, 61% are combined with materials, 59% with teacher training, and 56% with class size reduction or small-group instruction.

The vast majority of instructional and incentive treatments are evaluated in cluster-randomized experiments, with an average of 51 to 66 clusters per treatment arm and 51 to 74 in the control group. Health treatments mainly employ student-level randomization. I later assess whether typical experiments have adequate power to detect plausible effects.

Follow-Ups and Outcomes

Follow-ups are usually conducted after an academic year of exposure to treatments (Table 3). Despite the appeal of estimating longer run effects, it is rare that follow-ups occur more than a month after the treatment ends (Table 3). This is illustrated in Figure 2, in which solid circles indicate the date of baseline data collection, horizontal lines indicate the duration of treatments, and hollow circles indicate the date(s) of follow-up data collection. In a few cases, follow-ups occur one year after treatment (Duflo, Dupas, & Kremer, 2012). Baird, Hicks, Kremer,

& Miguel (2012) tracked an experimental cohort of Kenyan children who received deworming medication about 9 years after the baseline.

For each follow-up, I coded the proportion of the full experimental sample lost due to attrition, and the differential attrition between the treatment arm(s) and the control group. As the sample sizes in Table 3 make clear, data on full-sample attrition are missing for 17% to 34% of experimental follow-ups, with the smallest percentage in the sample of health treatments. Missing attrition data are less likely in noneconomics journals, which often enforce CONSORT guidelines (http://www.consort-statement.org), including the presentation of an experimental flow diagram.

The reasons for missing data vary. First, some cluster-randomized experiments do not conduct a baseline, including the experiments without a solid circle in Figure 2. In these, researchers conduct follow-up testing among a cohort of students, but without assurances that the same cohort was enrolled at the start of the treatment. Second, some experiments conduct a baseline in one cohort of students and a follow-up in another, perhaps with partial overlap but without student identifiers that could be used to calculate attrition (Friedman, Gerard, & Ralaingita, 2010). Both are potential threats to internal validity, since student enrollment and dropout—after the randomization of schools but before follow-up testing—may create imbalance in observed or unobserved variables that affect outcomes. Third, some experiments do not report attrition data for unstated reasons. In all cases, one might regard missing attrition data as a proxy of study quality.

The median full-sample attrition at follow-up is less than 15%. The median differential attrition (treatment minus control) is negative but close to zero. In this article, I remain agnostic about whether there are threshold levels of attrition that are "too high" (What Works Clearinghouse, 2011) or whether a study has adequately ruled out nonrandom attrition as a threat to internal validity. Instead, I specify four variables as potential moderators of study quality: full-sample attrition, the absolute value of differential attrition between a treatment arm and the control group, and dummy variables indicating missing data for each variable.

Last, Table 3 describes features of the outcome assessments. In general, instructional and incentive treatments were most likely to use tests designed by evaluators, NGOs, or governments, whereas a slim majority of health treatments used off-the-shelf assessments from a firm, university, or international agency.

Mean Effect Sizes by Treatment

Instructional Treatments

Figure 3 illustratively reports effect sizes for one treatment category: computers or technology. Figures S1 to S10 (see the online appendix), report data for 10 additional categories. Diamonds and brackets indicate effect sizes and their 95% confidence interval, respectively. Note that the size of diamonds is proportional to the random effects weights used to calculate the mean (see Equation 6 and its discussion). Effect sizes are down-weighted, for example, if they are imprecisely estimated and/or if authors reports many effect sizes within a single treatment arm (due to multiple follow-ups or outcome measures).

The mean effect size for computer-related treatments is 0.15. The p value in Figure 3 is .003; it is estimated with the wild cluster bootstrap-t, suitable for the

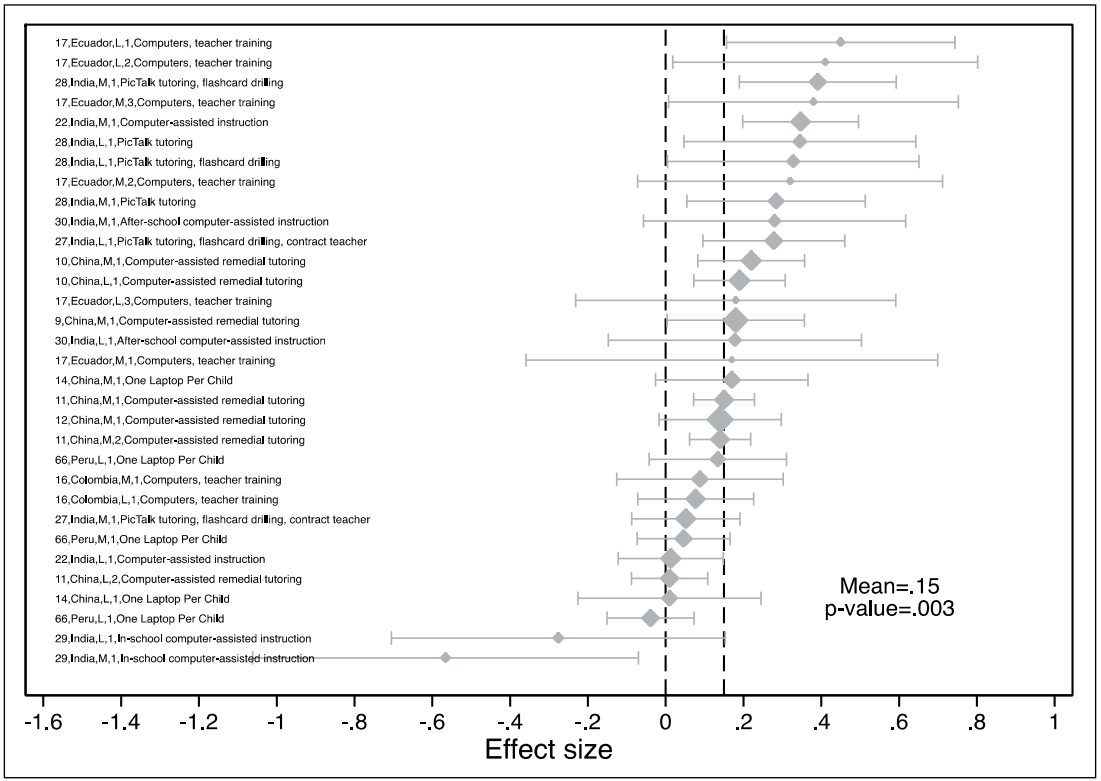

FIGURE 3. *Effect sizes of treatments with computers or technology.* Labels indicate the following: (a) experiment codes listed in the online appendix; (b) the country; (c) whether the outcome is language or reading, mathematics, or a composite; (d) the number of follow-up, from 1 to 3; and (e) a treatment description. Diamonds and brackets indicate effect sizes and 95% confidence intervals, respectively. The relative size of diamonds is proportional to their weight in the mean (see the text for discussion of weights). The mean effect size is estimated with Equation (6). The *p* value is calculated using the wild cluster bootstrap-*t*, allowing for clustering by study (see the text for details).

TABLE 4

Mean effect sizes by subcategory of treatment

	Mean effect size	*p*	Sample sizes		
			Studies	Experiments	Effect sizes
Instructional					
Computers or technology	0.150	.003	10	13	32
Teacher training	0.123	<.001	17	23	75
Class size or composition	0.117	.018	6	8	34
Instructional materials	0.078	<.001	15	19	69
Monetary grants	−0.011	.723	4	4	14
Health or nutrition					
Food, beverages, and/or micronutrients	0.035	.054	12	12	38
Deworming drugs	0.013	.388	5	7	22
Incentives					
Contract or volunteer teachers	0.101	<.001	8	11	41
Student/teacher performance incentives	0.089	.044	8	9	26
School management or supervision	0.055	.168	5	10	32
Informational treatments	0.049	.240	7	8	28

Note. The mean effect size is obtained with Equation (6). The *p* value is obtained with the wild cluster bootstrap-*t*, clustering by the number of studies. A single study may include more than one experiment if the experiments share samples and/or treatments. A single experiment is defined as one or more treatment arms and one control group. Each experiment contains $T * F * O$ effect sizes, where T is the number of treatment arms, F is the number of follow-ups, and O is the number of outcome measures.

small number of study clusters within most categories (Cameron, Gelbach, & Miller, 2008; Ringquist, 2013). The experiments in Figure 3 are notable for their geographic reach, with encouraging results in China (Lai, Zhang, Qu, et al., 2012; Mo et al., 2013), Ecuador (Carillo, Onofa, & Ponce, 2010), and India (Banerjee et al., 2007; He et al., 2008; Linden, 2008). Treatments apparently have smaller effects when laptop distribution is unaccompanied by parent or student training (Cristia et al., 2012), when computer use appears to substitute away from useful instructional time during school hours (He et al., 2008), or when the treatment does not incorporate consistent strategies for improving learning (Barrera-Osorio & Linden, 2009).

Table 4 summarizes mean effect sizes and *p* values for all 11 categories of treatments. The next largest means are for treatments that include teacher training (0.12), modify the size or composition of learning groups (0.12), or distribute

instructional materials (0.08). Recall that treatment categories are not mutually exclusive. That said, instructional materials alone do not improve learning (Glewwe et al., 2009; Kremer, Moulin, & Namunyu, 2003) but appear most effective when combined with teacher training and the use of a well-articulated instructional model (Banerjee et al., 2007; Friedman et al., 2010; Lucas, McEwan, Ngware, & Oketch, 2014). Class size reduction is often implemented in tandem with a contract teacher intervention (Bold et al., 2012; Duflo, Dupas, et al., 2012; Muralidharan & Sundararaman, 2010a), complicating efforts to isolate its effects. There is only one instance of a statistically significant effect size of monetary grants on learning (Das et al., 2011). The overall mean across four grant experiments is close to zero (−0.01) and statistically indistinguishable from zero.

Health and Nutrition Treatments

The mean effect size of providing food, beverages, or micronutrients in school settings is 0.04 (and statistically different from zero at 10%). The mean does not include a number of excluded studies, although the available data from these studies shows statistically nonsignificant effects for micronutrients[4] and zero to small effects for school meals.[5] Across all nutrition-related experiments, iron and micronutrient interventions have the greatest potential to increase learning, at least in Asian countries (Kleiman-Weiner et al., 2013; Luo et al., 2012; Soemantri, 1989; Soemantri, Pollitt, & Kim, 1985). The Kleiman-Weiner et al. (2013) study is particularly relevant, because it finds no effects of a food-based intervention in a second treatment arm. The mean effect size of deworming drugs is close to zero (0.01) and not statistically significant across seven experiments.[6] Despite the findings on learning outcomes, nutritional and deworming treatments often have positive effects on measures of enrollment and attainment not included in this meta-analysis (Baird et al., 2012; Miguel & Kremer, 2004; Petrosino et al., 2012). I later discuss the implications of these conflicting findings.

Incentive Treatments

The mean effects of performance incentives and informational treatments are 0.09 and 0.05, respectively, although the latter is not statistically distinguishable from zero. The two categories have the least overlap with others, permitting a more transparent horserace between popular incentive-based programs. The evidence on performance incentives is encouraging, particularly in India (Muralidharan & Sundararaman, 2011), but a Kenyan experiment found that effects were focused on the tests used in performance formulas (Glewwe, Ilias, & Kremer, 2010). Student performance incentives are more mixed, with positive effects in a Kenyan experiment (Kremer et al., 2009) but no evidence that cash incentives raise students' achievement in Nepali or Chinese classrooms, unless combined with peer tutoring (Li, Han, Rozelle, & Zhang, 2010; Sharma, 2010).

School report cards are sometimes effective, particularly when designed to be most useful to parents (Andrabi, Das, & Khwaja, 2009; Barr, Mugisha, Serneels, & Zeitlin, 2012), but many effects are small and imprecisely estimated. Providing information to students on the economic returns to schooling improves achievement in an oft-cited Magagascar experiment (Nguyen, 2008), but several other

treatment arms in the same experiment find information to be less effective when combined with apparently innocuous information about local role models.

The mean effect size of contract or volunteer teacher treatments is 0.10. It reflects a heterogeneous and overlapping set of interventions. Some are accompanied by instructional treatments. When these treatments rely mainly on volunteers, they appear less likely to improve learning (Banerjee et al., 2012; Cabezas et al., 2011). The effective use of contract teachers is often accompanied by smaller class sizes (Bold et al., 2012; Duflo, Dupas, et al., 2012; Muralidharan & Sundararaman, 2010a). To disentangle the effects, Duflo, Dupas, et al. (2012) included a treatment arm with class size reduction implemented with regular civil service teachers. They found small and imprecisely estimated effects of class size reduction alone. Even so, it is not clear whether a contract teacher intervention, in the absence of class size reduction, would be equally effective.

Finally, the diversity of school management and supervision treatments cautions against broad conclusions. Duflo, Dupas, et al. (2012) suggest that well-trained parent committees, when tasked specifically with managing teachers, can improve the effectiveness of both civil service and contract teachers in smaller classes. However, ambitious experiments in Gambia, Indonesia, and Madagascar showed few effects of school-based management and supervision reforms (Blimpo & Evans, 2011; Glewwe & Maïga, 2011; Pradhan et al., 2011), except for attempts to creates linkages between school committees and local governments (Pradhan et al., 2011).

Moderators of Effect Sizes

Table 5 reports random effects metaregressions that control for treatment subcategories, country contexts and experimental samples, outcome variables, and study quality.[7] The sample includes 259 effect sizes in 76 experiments, clustered within 57 studies. Given the constraints of sample size, particularly the number of study clusters, I pool effect sizes across all categories of interventions. The pooled sample excludes three effect sizes, including a radio mathematics treatment that is both the earliest treatment as well as the largest effect size of 1.5 (Jamison et al., 1981). The sample also excludes language and math effect sizes from a malaria prevention treatment that was the only such treatment eligible for inclusion in the sample (D. Fernando et al., 2006).

Column 1 controls for a series of dummy variables indicating treatment categories, whereas columns 2 to 6 include controls for additional moderators. Column 6 is the most complete specification. Treatments with either a training component or a component related to class size and composition have remarkably robust effects on effect sizes across all specifications. Computer-related treatments and performance incentives have consistently positive effects, which are larger and significant only when controlling for country context and sample moderators (notably the gross domestic product [GDP] per capita of countries). Nutritional treatments present a special case. The coefficient in column 6 is not statistically significant, but the positive coefficient (0.04) is quite imprecisely estimated and consistent with negative or even much larger effects. Finally, materials, grants, deworming, information, contract and volunteer teachers, and management and supervision have anywhere from negative to statistically nonsignificant effects.

TABLE 5

Random effects metaregressions

	1	2	3	4	5	6
Materials	−0.051* (0.028)	−0.064*** (0.022)	−0.024 (0.021)	−0.064*** (0.022)	−0.077*** (0.026)	−0.033 (0.022)
Computers or technology	0.023 (0.041)	0.044 (0.041)	0.068* (0.035)	0.047 (0.041)	0.032 (0.046)	0.076** (0.037)
Grants	−0.062 (0.039)	−0.054* (0.028)	−0.039 (0.025)	−0.045* (0.026)	−0.048 (0.029)	−0.030 (0.033)
Teacher training	0.091*** (0.019)	0.104*** (0.021)	0.080*** (0.023)	0.098*** (0.022)	0.117*** (0.027)	0.086*** (0.021)
Class size, small-group instruction, tracking	0.075*** (0.021)	0.089*** (0.021)	0.072** (0.032)	0.079*** (0.026)	0.077*** (0.023)	0.069* (0.039)
Food, beverage, micronutrients	−0.017 (0.032)	0.039 (0.040)	0.023 (0.038)	0.039 (0.043)	0.009 (0.055)	0.044 (0.048)
Deworming	−0.042 (0.032)	−0.000 (0.035)	−0.014 (0.037)	−0.003 (0.037)	−0.021 (0.045)	−0.018 (0.043)
Other health	0.018 (0.037)	0.017 (0.043)	0.082 (0.057)	0.034 (0.046)	0.025 (0.051)	0.070 (0.059)
Information	−0.007 (0.032)	0.017 (0.036)	−0.002 (0.032)	0.013 (0.034)	0.008 (0.041)	0.002 (0.030)
Performance incentives	0.029 (0.035)	0.042 (0.032)	0.045 (0.032)	0.038 (0.033)	0.024 (0.035)	0.056* (0.033)
Contract or volunteer teachers	−0.020 (0.022)	−0.040 (0.026)	−0.024 (0.031)	−0.037 (0.027)	−0.027 (0.025)	−0.022 (0.033)
School management, supervision	−0.010 (0.028)	−0.024 (0.027)	−0.022 (0.032)	−0.018 (0.029)	−0.018 (0.026)	−0.021 (0.030)
Months of treatment exposure at follow-up		−0.001 (0.001)	−0.001 (0.001)	−0.002 (0.001)	−0.002* (0.001)	−0.003* (0.001)
Follow-up conducted >1 month after treatment?		−0.025 (0.020)	−0.011 (0.020)	−0.028 (0.022)	−0.030 (0.022)	−0.009 (0.022)
Government implementer		−0.011 (0.031)	−0.024 (0.041)	−0.023 (0.035)	−0.002 (0.041)	−0.007 (0.048)
University/researcher implementer		−0.076*** (0.028)	−0.031 (0.027)	−0.076** (0.029)	−0.063* (0.036)	−0.038 (0.033)
Post-1990s			−0.016 (0.040)			−0.039 (0.043)
Latin America and Caribbean			0.108** (0.047)			0.125** (0.061)
East Asia and Pacific			0.051 (0.034)			0.054 (0.039)

(continued)

TABLE 5 (continued)

	1	2	3	4	5	6
South Asia			0.017 (0.038)			0.005 (0.042)
Log(gross domestic product per capita in baseline year, US$ (2000))			−0.070*** (0.019)			−0.087*** (0.029)
Grades 5–8 only			−0.022 (0.026)			−0.016 (0.031)
Both Grades 1–4 and 5–8			−0.035 (0.025)			−0.030 (0.026)
Uncertain grades			0.025 (0.042)			0.025 (0.041)
Convenience sample			−0.041** (0.019)			−0.051* (0.026)
Composite test score				0.019 (0.021)		−0.021 (0.035)
Math test score				−0.012 (0.015)		−0.007 (0.015)
Commercial/international test				−0.013 (0.027)		−0.001 (0.030)
Government or school test				−0.014 (0.026)		−0.027 (0.033)
Uncertain test				0.005 (0.029)		0.005 (0.043)
Experiment is published					0.020 (0.024)	−0.018 (0.025)
Alternating list assignment					0.024 (0.038)	0.037 (0.038)
Attrition at follow-up					0.085 (0.087)	0.018 (0.108)
Missing indicator					0.013 (0.026)	−0.032 (0.025)
Absolute value of differential attrition					−0.722** (0.277)	−0.513 (0.311)
Missing indicator					0.004 (0.018)	0.026 (0.018)
Constant	0.056** (0.026)	0.088*** (0.026)	0.571*** (0.129)	0.104*** (0.036)	0.069** (0.031)	0.730*** (0.213)

Note. The sample for each regression includes 259 effect sizes from 76 experiments clustered within 57 studies. Robust standard errors are clustered by study. See the text for details of sample, weights, and estimation.
*$p < .1$. **$p < .05$. ***$p < .01$.

Why are some categories apparently less effective after controlling for moderators? One conjecture is that their effects are explained by effective (and overlapping) treatment components such as teacher training, computer-assisted instruction, and class size reduction and small-group instruction. A related hypothesis is that effects are attributable not to either category in isolation but rather to the interaction between two or more treatment components. That is, instructional materials might be effective when combined with a complementary treatment component such as teacher training but training or materials alone would ineffective. In a metaregression, one might include interaction terms between treatment components, but the modest sample sizes in Table 5 do not allow convincing tests. (In specifications with added interactions, the standard errors are larger and coefficient estimates fluctuate substantially depending on the specification, suggesting "micronumerosity" and multicollinearity.)

There are surprisingly few experiments with fully factorial designs that allow for strong experimental tests of these hypotheses (e.g., three treatment arms consisting of training, materials, and their combination). I am only aware of three fully factorial designs that evaluate instructional or incentive interventions (Brooker et al., 2010; He et al., 2008; Nguyen, 2008). Many experiments in the sample use incomplete factorial designs (Duflo, Dupas, et al., 2012; Pradhan et al., 2011).

Among remaining moderators, there are few consistent correlates of effect sizes. Two are country related, including the Latin America region (relative to Africa) and the real GDP per capita in the baseline year. For a 10% increase in GDP per capita, all else equal, effect sizes decrease by approximately 0.01. The result has no evident causal interpretation, since country income may be proxying household incomes of school children, the quality of schooling in control groups, or other variables. Experiments using a convenience sample have lower effect sizes, on average, contrary to the intuition that purposively chosen samples of schools or students may also be the most likely to benefit from treatments. It could indicate that random experimental samples reflect careful research planning that is also likely to be correlated with well-designed treatments. Finally, there is a large but imprecisely estimated coefficient on the absolute value of differential attrition, suggesting that larger differences are associated with lower effects.

The results in Table 5 are helpful in testing two common methodological concerns. First, there is no evidence that the use of quasi-random assignment, such as alternating selection from ordered lists, is associated with lower or higher effects. Second, one might be concerned that publication bias leads journals to prefer studies with positive effects. There is a small and not statistically significant coefficient on a variable indicating published papers, versus working papers or reports. Despite the large number of unpublished experiments included in the sample, a lingering concern is that experiments with zero or negative effects, especially imprecise ones, are never reported. As a partial assessment, Figure S11 (see online appendix) illustrates a roughly symmetrical funnel plot of effect sizes against their standard errors. Given the difficulty of visually assessing asymmetry, I conducted a trim-and-fill analysis (Duval & Tweedie, 2000). In the full sample of 259 effect sizes, the random effects mean is 0.072. After including 14 "filled" effect sizes, the updated mean of 0.067 yields substantively similar conclusions.

Similarly effective treatments may vary widely in their costs, and treatments with smaller effects may nonetheless have relatively lower costs (Levin & McEwan, 2001; McEwan, 2012). In either case, it is misleading to use effect size as the sole criterion for ranking of investments. To facilitate cost comparisons, evaluations should ideally report the incremental cost of all resources (e.g., personnel, facilities, and materials) incurred by all stakeholders (e.g., schools and governments, NGOs, and clients) during the treatment's application. In the meta-analytic sample, I found that 56% of treatments reported no details on incremental costs, while most of the rest reported minimal details. For example, studies usually did not report sources of data, and some appeared to omit categories of resources. Most did not report the exchange rates used to convert estimates to a common currency (US$) or how cost estimates were adjusted for inflation.

Confronting similar issues, Kremer et al. (2013) report auxiliary cost estimates for a subset of experiments (Abeberese, Kumler, & Linden, 2012; Banerjee et al., 2007; Duflo, Dupas, et al., 2012; Duflo et al., 2011; Duflo, Hanna, & Ryan, 2012; Glewwe et al., 2009; Glewwe et al., 2010; Kremer et al., 2009; Nguyen, 2008; Pradhan et al., 2011). In addition to gathering quantity and price data for consistent categories of resources, the authors applied consistent assumptions regarding the discount rate (10%), inflation, and exchange rates (see www.povertyactionlab.org/doc/cea-data-full-workbook). Using these data, I calculate the social cost per student in 15 treatment arms of the meta-analytic sample, converted to US$ with purchasing power parity exchange rates. This article's estimates differ from Kremer et al. (2013) in three ways. First, I impute a deadweight loss from taxation as 20% of costs (Auriol & Warlters, 2012). Second, I do not include transfer payments in the cost estimates, although I do include the deadweight loss associated with transfers. Third, I recalculate costs in a few treatment arms so that they exactly correspond to the treatment-control contrast coded in the meta-analytic sample.

The diamonds in Figure 4 illustrate the social cost per student of increasing the effect size by 0.2 (hollow diamonds indicate that the underlying effect size is not significant at 10%). The *x*-axis uses a log scale to facilitate interpretations, since most CERs are below $100, but a few are far higher. The most cost-effective alternatives are the provision of information to students about economic returns in Madagascar (Nguyen, 2008), "linkages" between school committees and village leadership in Indonesia (Pradhan et al., 2011), and ability tracking in Kenya (Duflo et al., 2011). The least cost-effective alternatives include computer-assisted instruction in India (Banerjee et al., 2012), the provision of textbooks in Kenya (Glewwe et al., 2009), and class size reduction in Kenya (Duflo, Dupas, et al., 2012).

Despite suggestive findings, the results are subject to strong caveats. First, the cost-effectiveness analysis (CEA) is based on a subset of the meta-analytic sample. Ideally, one would develop valid estimates of comparable costs and CERs in all impact evaluations. This would facilitate broader analyses that compare cost-effectiveness across broader classes of interventions. Metaregressions could further be used to examine the moderators of CERs, although this has yet to be pursued in the literature on education CEA.

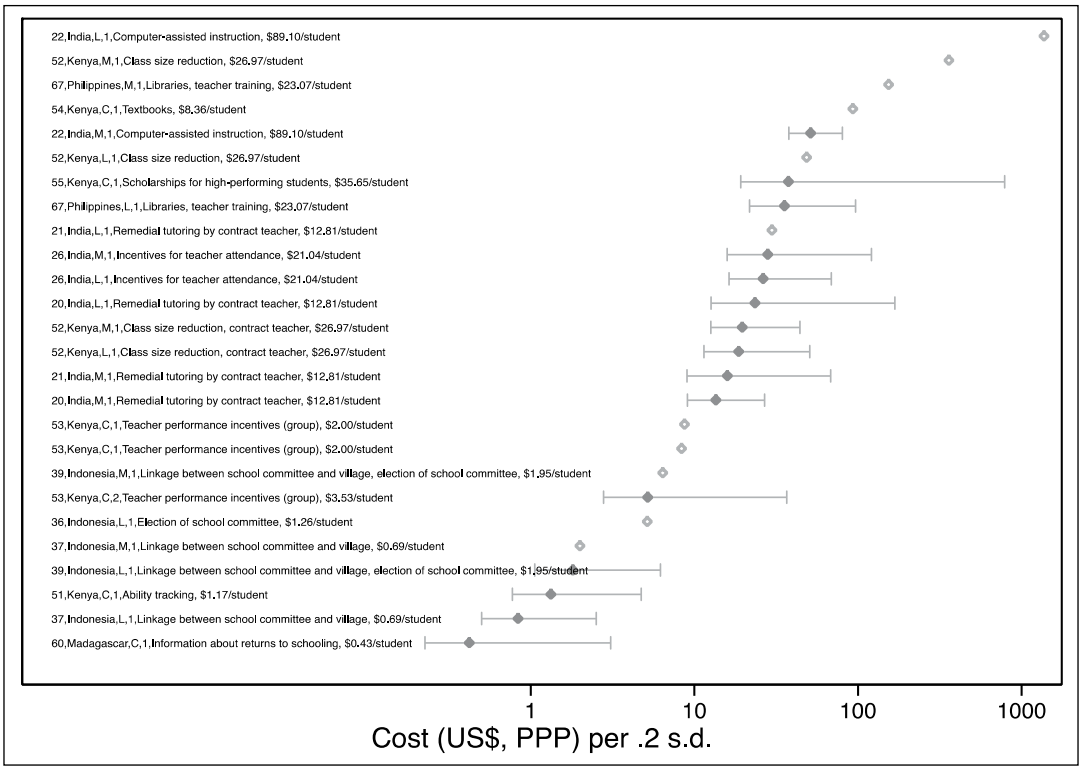

FIGURE 4. *Cost-effectiveness ratios in 15 treatment arms with comparable cost data.* Labels indicate the following: (a) experiment codes listed in the online appendix; (b) the country; (c) whether the outcome is language or reading, mathematics, or a composite; (d) the number of follow-up, from 1 to 3; (e) a treatment description; and (f) the cost per student used to calculate the cost-effectiveness ratio. The x-axis indicates the cost per 0.2 units of the effect size (on a log scale). Diamonds and brackets plot cost-effectiveness ratios and 90% confidence intervals, respectively (assuming no uncertainty in the cost estimate; see the text for details). Hollow diamonds indicate that the effect size used to calculate the ratio is not statistically significant at 10%. Cost data are from Kremer, Brannen, and Glennerster (2013), with modifications described in the text.

Second, it is tempting to allocate resources in ascending order of the CERs, but this would be short-sighted. Figure 4 reports 90% confidence intervals for CERs under the strong assumption that cost estimates—unlike effects—are not subject to estimation error.[8] The many overlapping intervals suggest that precise rankings are not feasible, at least in the middle of the CER distribution. Intervals might be even wider if incremental costs are estimated with error. Future research could address this by collecting student- and school-specific data on costs in the treatment and control groups. These data, in concert with standard methods from statistics and health research (Briggs et al., 2002), can be used to construct more credible confidence intervals for CERs.

Third, cost-effectiveness rankings such as Figure 4 also do not reflect the heterogeneous objectives of most school-based interventions but especially those in health and nutrition that have been shown to affect child health, enrollment, and attainment (Miguel & Kremer, 2004; Petrosino et al., 2012). In such cases, it would be less misleading to complement CEA with a full cost-benefit analysis that attaches monetary benefits to a wider range of child and adult outcomes (McEwan, 2012).

Discussion

Effects on Learning

The treatments can be divided into four broad groups. In the first, school grants and deworming treatments do not affect test scores, on average, and these effects are robust to controls for moderators. The second group includes nutritional, informational, and management-related treatments. On average, nutritional interventions have small effects on learning that are of a similar magnitude after controlling for moderators, albeit not statistically significant. Informational treatments also have relatively smaller effects that are not robust to controls for moderators. Finally, the average effects of management and supervision treatments are small, and not robust to moderators.

The third group includes interventions that are more effective, on average, but not robust to controls for moderators. These include instructional materials and contract and volunteer teachers. In each category, one can point to highly effective treatments that are, nonetheless, often accompanied by teacher training, computers or technology, class size reduction, or other interventions. Contract teacher interventions—especially those not relying on pure volunteers—are consistently effective but are often implemented at the same time as class size reduction (Bold et al., 2012; Duflo, Dupas, et al., 2012; Muralidharan & Sundararaman, 2010a) and other interventions conducted in small-group settings (Banerjee et al., 2007). Duflo, Dupas, et al. (2012) show that class size reduction by itself is minimally effective in Kenyan classrooms, but it is still not clear whether smaller classes are a necessary condition for the effectiveness of contract teachers.

The fourth group includes treatments that are effective, on average, even with a full set of moderator controls. These include teacher training, computers and technology, treatments that modify the size and composition of learning groups, and performance incentives. The results on teacher training merit some caution, since the degree of overlap with other interventions is substantial, though it is

telling that almost all successful instructional interventions in our sample include at least a minimal attempt to develop teachers' capacity to deliver effective classroom instruction. Treatments that reduce the size of classes or learning groups are effective when combined with complementary treatments but perhaps not alone (Duflo, Dupas, et al., 2012). As with computers and instructional materials, the broader lesson seems to be that reducing the size of learning groups can be effective, as long as there is a clear strategy—whether instructional or incentive based—for ensuring that additional instructional time is spent wisely.

There are three experiments in which teacher performance incentives have been shown to increase student learning when provided to individuals or groups, although a Kenyan one sounds a cautionary note about its potential effects on strategic behavior (Glewwe et al., 2010). The most consistent lesson to date is that teachers are indeed responsive to financial or in-kind incentives. The fundamental challenge, yet to be explored across many experiments, is whether teacher incentives can be designed that consistently maximize learning while minimizing strategic responses and whether these incentives can potentially enhance the positive results from effective instructional interventions.

Enrollment and Attainment

The review found zero or small learning effects of health interventions such as deworming and school meals. However, an oft-cited Kenyan experiment finds short- and longer run impacts of deworming treatments on school participation and attainment (Baird et al., 2012; Miguel & Kremer, 2004). In-kind transfers like school feeding programs have often been found to increase short-run measures of school participation such as enrollment and attendance (Petrosino et al., 2012; Vermeersch & Kremer, 2004). In a similar vein, conditional cash transfers (CCTs) have been shown to consistently improve school enrollment and attainment (Fiszbein & Schady, 2009; Galiani & McEwan, 2013), but with mixed effects on learning that may depend on the local context of school quality (Barham, Macours, & Maluccio, 2012; Behrman, Parker, & Todd, 2009).

Viewed together, the results suggest that attending school is a necessary but not sufficient condition for improving learning. Future experiments could profitably combine access-based interventions—such as school meals and CCTs—with instructional interventions in schools. In Honduras, for example, one of the earliest studies of CCTs developed a factorial design to separately evaluate block grants to schools, CCTs, and the combination of the two (Galiani & McEwan, 2013). Unfortunately, the block grant intervention was minimally implemented.

Sample Size and Power Analysis

Smaller experiments are less costly to conduct, but they also yield less precise estimates of treatment effects. This is worrisome when confidence intervals are so wide that researchers cannot statistically distinguish effect size estimates—even quite large ones—from zero. Consider a cluster-randomized experiment with 100 schools divided evenly between treatment and control groups, 50 students per school, and an intraclass correlation (ICC) of .2. The ICC measures the proportion of variance in the outcome that lies between clusters. It may vary by outcome and/or by country. In U.S. evaluations, the recommended ICCs usually fall between

.15 and .2 for test score outcomes (Hedges & Hedberg, 2007; Schochet, 2005; What Works Clearinghouse, 2011). The MDES of this design is 0.26, which is substantially larger than mean effect sizes for all treatment categories.[9]

Using the 2011 PIRLS (Progress in International Reading Literacy Study) assessment of fourth-grade reading achievement, I calculated ICCs for Botswana (0.36), Colombia (0.44), Honduras (0.35), Indonesia (0.41), Morocco (0.42), and Trinidad and Tobago (0.33). Zopluoglu (2012) estimates ICCs using all countries that participated in the fourth-grade PIRLS assessment, and the eighth-grade TIMSS mathematics assessment, finding many estimates that exceed the common assumption of 0.2. Assuming a more plausible ICC of .4 increases the MDES to 0.37. Only 12% and 7%, respectively, of this study's effect sizes exceed those values.

What options remain for researchers? With 200 schools and an ICC of .2, the MDES falls to 0.19. One could also stratify the experimental sample prior to randomization and/or gather baseline data such as a pretest. Assuming that baseline variables explain 35% of the variation in the test score outcome, the MDES further declines to 0.15 (or 0.21 under the larger ICC). Quadrupling the number of students per cluster changes the MDES by less than 0.01, implying little justification for gathering data on all students in a particular cluster.

The results highlight the challenges facing budget-constrained researchers who are evaluating interventions with modest effects on student outcomes. They imply several guidelines. First, researchers should conduct and report a realistic power analysis, based on country-specific assumptions about ICCs and a review of effect size estimates from similar treatments. Second, researchers should stratify school samples by pretest, poverty, or other plausible determinants of final test scores, and control for strata in their analyses. If variables are numerous, then researchers can apply pairwise matching and use all of them (Bruhn & McKenzie, 2009). Third, researchers should collect baseline data on students and schools, ideally a pretest, and report estimates that control for the pretest. Even when student-level data cannot be collected, a cluster-level pretest can increase power in cluster-randomized experiments (Bloom, Richburg-Hayes, & Black, 2007).

Student-level randomization is far more common in evaluations of individually administered health treatments (see Table 1), although it has been usefully applied to an instructional treatment (Mo et al., 2012). It can be a cost-effective way to conduct an experiment, since data collection and treatments occur in many fewer schools. The MDES of a student-level randomized experiment with 300 students, evenly allocated to treatment and control groups, is 0.29 (assuming that school effects are fixed, instead of random). Doubling the number of students, and assuming that baseline variables explain 35% of variation in the outcome, lowers the MDES to 0.16.

Despite its potential benefit, student-level randomization has three drawbacks. First, it is not well suited for instructional interventions that are given, by design, to entire classrooms or schools. Second, the proximity of students in treatment and control groups introduces the risk that control group students receive treatment benefits (Miguel & Kremer, 2004). The plausibility of spillovers may depend on the nature of the treatment and parallel efforts by the researcher to prevent them from occurring (Mo et al., 2012), suggesting no simple recipes. Third, most such experiments are conducted in small samples of schools,

potentially limiting external validity even beyond the convenience samples of schools in cluster-randomized experiments.

External Validity

A common critique of randomized experiments is that their results have limited generalizability to different populations of students, contexts, outcome measures, and variations in the treatment. The meta-analytic results suggest that some of these concerns are exaggerated. Metaregressions showed that several types of treatments were consistently associated with larger effect sizes, on average, even after controlling for moderators related to the context, outcome measures, and study quality. However, these results must be cautiously interpreted. The meta-analytic sample is still small, and it lacks variation in moderators of particular interest to policymakers, such as the implementing agency—whether government, NGO, or researcher—and the scale of the intervention. Even more caution is warranted in generalizing results to richer countries not included in the meta-analysis. The metaregressions suggested that effect sizes decline by approximately 0.01 with each 10% increase in GDP per capita, all else equal, which could indicate that control group schools have relatively higher levels of teacher quality, school resources, and household incomes. It is plausible, but cannot be directly tested with this study's data, that effect sizes for similar types of treatments would be smaller in higher income countries.

In the meantime, how can school-based experiments provide more generalizable knowledge to policymakers? First, experiments should use samples that are representative of well-defined, policy-relevant populations of schools and students, when possible (Muralidharan & Sundararaman, 2011). Whatever the sample, experiments can report estimates of treatment effects within policy-relevant subgroups of schools and students. Even so, Deaton (2010) cautions against situations in which evaluators "hunt" for statistically significant effects in arbitrarily chosen subgroups, and then construct ex post theories to explain these effects. To guard against this likelihood, it is desirable to define subgroups before data collection.

Second, experiments should measure and report a wide set of outcomes. Treatments with a narrow focus on language or reading may cause instructional time to be diverted from mathematics or other subjects, and vice versa. Also, treatments might lead to strategic behavior that increases one outcome but not another. Teacher performance pay in Kenya produced gains in the incentivized test score but not a lower stakes exam, which the authors attributed to test coaching behavior rather than instructional improvements (Glewwe et al., 2010).

Third, experiments can provide data to assess causal mechanisms. Policy effects may encompass multiple direct and indirect effects on learning. Researchers could experimentally manipulate one of those of mechanisms, but this is costly. In the best alternative, authors specify the hypothesized mechanisms in a theory of change; gather appropriate data on inputs, processes, and intermediate outcomes; and construct plausible narratives about what rendered a treatment more or less effective. These analyses inevitably depart from the experimental ideal—and rest on weaker causal evidence—but they provide useful context for policymakers to judge whether treatment effects are driven by specific features of the treatment

or context (and hence whether some or all it might be fruitfully transferred elsewhere).

Notes

I am grateful to the Quality Education in Developing Countries initiative of the William and Flora Hewlett Foundation for financial support (Grant/Award No. 2011-7042). David Evans, Adrienne Lucas, Chloe O'Gara, Maria Perez, Pat Scheid, Dana Schmidt, Rebecca Thornton, three anonymous referees, and seminar participants at the Inter-American Development Bank provided helpful comments; however, they are not responsible for errors or interpretations. Kate Kemmerer and Poppy Tian provided excellent support in the design and coding of the database. The data and statistical code used in this article are available at www.patrickmcewan.net/meta.

[1]Low-income countries in the sample include the Gambia, Kenya, Madagascar, Mali, Nepal, Tanzania, and Uganda. Lower middle–income countries include Guatemala, India, Indonesia, Nicaragua, Pakistan, Philippines, and Sri Lanka. Upper middle–income countries include Chile, China, Colombia, Ecuador, Jamaica, Mexico, Peru, South Africa, and Thailand.

[2]School personnel could use information to diagnose student weaknesses and efficiently allocate instructional resources to the neediest students. Parents could use information in a similar fashion to allocate resources within households, to exert direct pressure on school personnel or students who are judged to be low-performing, or to inform different choices about schools and teachers, at least when local institutions facilitate such choices (Bruns et al., 2011). Finally, students who receive information about the relationship between their current performance and future earnings may have improved incentives to exert effort in the short run.

[3]Contract teachers and volunteers may have stronger incentives to attend class regularly and deliver effective instruction, since they can be terminated for nonperformance. Even so, the typical experimental design makes it challenging to separate incentive effects from those of concomitant reductions in class size (Bold et al., 2012; Duflo, Dupas, et al., 2012; Muralidharan & Sundararaman, 2010a), instructional materials and training (Banerjee et al., 2007, 2012; Cabezas, Cuesta, & Gallego, 2012; He et al., 2008), or simply from pretreatment differences in the capacities of regular and contract teachers.

[4]Pollitt et al. (1989) and Sungthong et al. (2004) found that iron supplementation did not have statistically significant effects on achievement of Thai children, while Vazir et al. (2006) found no significant effects of a micronutrient-fortified beverage on school exam scores in India.

[5]School feeding programs did not have statistically significant impacts on math or literacy scores in two experiments conducted in Uganda (Adelman et al., 2008) and Peru (E. Jacoby, Cueto, & Pollitt, 1996). A Kenyan experiment did not report sufficient data to estimate effect sizes at each follow-up but reports average annual growth of test scores in treatments groups versus a control (Whaley et al., 2003). It found no significant effects of any treatment on a verbal test. On a math test, it found effects of 0.11 standard deviations per year (meat), 0.15 (energy-based diet), and 0.02 and zero or not statistically significant (milk supplement).

[6]An excluded study did not include sufficient data to estimate effect sizes, but it reported no statistically significant effects of a deworming treatment on school exams (Nga et al., 2011).

[7]At the suggestion of a referee, I reestimated the regressions in Table 5 with controls for the inverse standard error of the effect size and the natural log of the number of randomized units (whether students, schools, or villages). The results, available from the author, are substantively similar to those reported in Table 5.

[8]In the literature on health CEA, it is increasingly common to collect patient- or cluster-specific data on incremental treatment costs. With these data, it is possible to estimate the standard error of incremental costs and—using the bootstrap or Fieller's theorem—the standard error and confidence interval of a CER (Briggs, O'Brien, & Blackhouse, 2002). When the standard error of costs is assumed to be zero (as in Figure 4), the application of Fieller's theorem is equivalent to simply rescaling the 90% confidence interval of the effect size by the "known" cost parameter.

[9]This further assumes a significance level of .05, and a power of 0.8. The former is the probability of committing a Type I error (i.e., rejecting a null hypothesis of no effect when it is true). Power is $1 - \beta$, where β is the probability of committing a Type II error (i.e., failing to reject a null hypothesis of no effect when it is false).

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Abeberese, A. B., Kumler, T. J., & Linden, L. (2012). *Improving reading skills by encouraging children to read: A randomized evaluation of the Sa Aklat Siskat reading program in the Philippines*. Unpublished manuscript, Columbia University, New York, NY.

Adelman, S., Alderman, H., Gilligan, D. O., & Lehrer, K. (2008). *The impact of alternative food for education programs on learning achievement and cognitive development in northern Uganda*. Unpublished manuscript, University of Maryland, College Park.

*Andrabi, T., Das, J., & Khwaja, A. I. (2009). *Report cards: The impact of providing school and child test-scores on educational markets*. Unpublished manuscript, Pomona College, Claremont, CA.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, *92*, 1535–1558. doi:10.1257/000282802762024629

Auriol, E., & Warlters, M. (2012). The marginal cost of public funds and tax reform in Africa. *Journal of Development Economics*, *97*, 58–72. doi:10.1016/j.jdeveco.2011.01.003

*Baird, S., Hicks, J. H., Kremer, M., & Miguel, E. (2012). *Worms at work: Long-run impact of child health gains*. Unpublished manuscript, George Washington University, Washington, DC.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy*, *2*, 1–30. doi:10.1257/pol.2.1.1

*Banerjee, A., Banerji, R., Duflo, E., & Walton, M. (2012). *Effective pedagogies and a resistant education system: Experimental evidence on interventions to improve basic skills in rural India*. Unpublished manuscript, MIT, Cambridge.

*Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics*, *122*, 1235–1264. doi:10.1162/qjec.122.3.1235

Barham, T., Macours, K., & Maluccio, J. A. (2012). *More schooling and more learning? Effects of a three-year conditional cash transfer program in Nicaragua after 10 years* (Working Paper No. IDB-WP-432). Washington, DC: Inter-American Development Bank.

*Barr, A., Mugisha, F., Serneels, P., & Zeitlin, A. (2012). *Information and collective action in the community monitoring of schools: Field and lab experimental evidence*

from Uganda. Unpublished manuscript, University of Nottingham, Nottingham, England.

*Barrera-Osorio, F., & Linden, L. L. (2009). *The use and misuse of computers in education: Evidence from a randomized controlled trial of a language arts program*. Unpublished manuscript, Columbia University, New York, NY.

Behrman, J. R., Parker, S. W., & Todd, P. E. (2009). Medium-term impacts of the Oportunidades conditional cash transfer program on rural youth in Mexico. In S. Klasen & F. Nowak-Lehmann (Eds.), *Poverty, inequality, and policy in Latin America* (pp. 219–270). Cambridge, UK: MIT Press.

Berry, J. (2012). *Child control in education decisions: An evaluation of targeted incentives to learn in India*. Unpublished manuscript, Cornell University, Ithaca, NY.

Best, C., Neufingerl, N., Del Rosso, J. M., Transler, C., van den Briel, T., & Osendarp, S. (2011). Can multi-micronutrient food fortification improve the micronutrient status, growth, and cognition of schoolchildren? A systematic review. *Nutrition Reviews*, *69*, 186–204. doi:10.1111/j.1753-4887.2011.00378.x

Beuermann, D. W., Cristia, J. P., Cruz-Aguayo, Y., Cueto, S., & Malamud, O. (2012). *Home computers and child outcomes: Short-term impacts from a randomized experiment in Peru* (Working Paper No. IDB-WP-382). Washington, DC: Inter-American Development Bank.

Blimpo, M. P. (2010). *Team incentives for education in developing countries: A randomized field experiment in Benin*. Unpublished manuscript, Stanford University.

*Blimpo, M. P., & Evans, D. K. (2011). *School-based management and educational outcomes: Lessons from a randomized field experiment*. Unpublished manuscript, Stanford University, Stanford, CA.

Bloom, H. W., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*, 30–59. doi:10.3102/0162373707299550

*Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2012). *Interventions and institutions: Experimental evidence on scaling up education reforms in Kenya*. Unpublished manuscript, Stockholm University, Stockholm, Sweden.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). New York, NY: Russell Sage Foundation.

*Borkum, E., He, F., & Linden, L. L. (2012). *School libraries and language skills in Indian primary schools: A randomized evaluation of the Akshara library program* (Working Paper No. 18183). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w18183

Briggs, A. H., O'Brien, B. J., & Blackhouse, G. (2002). Thinking outside the box: Recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. *Annual Review of Public Health*, *23*, 377–401. doi:10.1146/annurev.publhealth.23.100901.140534

Brooker, S., Okello, G., Njagi, K., Dubeck, M. M., Halliday, K. E., Inyega, H., & Jukes, M. C. H. (2010). Improving educational achievement and anaemia of school children: Design of a cluster randomised trial of school-based malaria prevention and enhanced literacy instruction in Kenya. *Trials*, *11*(93), 1–14. doi:10.1186/1745-6215-11-93

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, *1*, 200–232. doi:10.1257/app.1.4.200

Bruns, B., Filmer, D., & Patrinos, H. A. (2011). *Making schools work: New evidence on accountability reforms*. Washington, DC: World Bank.

*Cabezas, V., Cuesta, J. I., & Gallego, F. A. (2011). *Effects of short-term tutoring on cognitive and non-cognitive skills: Evidence from a randomized evaluation in Chile*. Unpublished manuscript, Pontificia Universidad Católica de Chile, Santiago.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, *90*, 414–427. doi:10.1162/rest.90.3.414

*Carillo, P., Onofa, M., & Ponce, J. (2010). *Information technology and student achievement: Evidence from a randomized experiment in Ecuador* (Working Paper No. IDB-WP-223). Washington, DC: Inter-American Development Bank.

Chandler, A. K., Walker, S. P., Connolly, K., & Grantham-McGregor, S. M. (1995). School breakfast improves verbal fluency in undernourished Jamaican children. *Journal of Nutrition*, *125*, 894–900.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review*, *95*, 1237–1258. doi:10.1257/0002828054825529

Clarke, S. E., Jukes, M. C. H., Njagi, J. K., Khasakhala, L., Cundill, B., Otido, J., . . . Brooker, S. (2008). Effect of intermittent preventive treatment of malaria on health and education in schoolchildren: A cluster-randomised, double-blind, placebo-controlled trial. *Lancet*, *372*, 127–138. doi:10.1016/S0140-6736(08)61034-X

*Cristia, J. P., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2012). *Technology and child development: Evidence from the One Laptop per Child program* (Working Paper No. IDB-WP-304). Washington, DC: Inter-American Development Bank.

*Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2011). *School inputs, household substitution, and test scores* (Working Paper No. 16830). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w16830

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*, 424–455. doi:10.1257/jel.48.2.424

Dickson, R., Awasthi, S., Williamson, P., Demellweek, C., & Garner, P. (2000). Effects of treatment for intestinal helminth infection on growth and cognitive performance in children: Systematic review of randomised trials. *British Medical Journal*, *320*, 1697–1701. doi:10.1136/bmj.320.7251.1697

Dong, N., & Maynard, R. (2013). PowerUp! A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*, 24–67. doi:10.1080/19345747.2012.673143

*Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*, 1739–1774. doi:10.1257/aer.101.5.1739

*Duflo, E., Dupas, P., & Kremer, M. (2012). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools (Working Paper 17939). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w17939

Duflo, E., Glennerster, R., & Kremer, M. (2008). Using randomization in development economics research: A toolkit. In T. P. Schultz & J. Strauss (Eds.), *Handbook of development economics* (Vol. 4, pp. 3895–3062). Amsterdam, Netherlands: Elsevier. doi:10.1016/S1573-4471(07)04061-2

*Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, *102*, 1241–1278. doi:10.1257/aer.102.4.1241

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.

Eilander, A., Gera, T., Sachdev, H. S., Transler, C., van der Knapp, H. C. M., Kok, F. J., & Osendarp, S. J. M. (2010). Multiple micronutrient supplementation for improving cognitive performance in children: Systematic review of randomized controlled trials. *American Journal of Clinical Nutrition*, *91*, 115–130. doi:10.3945/ajcn.2009.28376

Evans, D. K., & Ghosh, A. (2008). *Prioritizing educational investments in children in the developing world* (Working Paper No. WR-587). Santa Monica, CA: RAND.

Evans, D., Kremer, M., & Ngatia, M. (2009). *The impact of distributing school uniforms on children's education in Kenya*. Unpublished manuscript, Harvard University, Cambridge, MA.

Falkingham, M., Abdelhamid, A., Curtis, P., Fairweather-Tait, S., Dye, L., & Hooper, L. (2011). The effects of oral iron supplementation on cognition in older children and adults: A systematic review and meta-analysis. *Nutrition Journal*, *9*, 4. doi:10.1186/1475-2891-9-4

*Fernando, D., de Silva, D., Carter, R., Mendis, K. N., & Wickremasinghe, R. (2006). A randomized, double-blind, placebo-controlled, clinical trial of the impact of malaria prevention on the educational attainment of school children. *American Journal of Tropical Medicine and Hygiene*, *74*, 386–393.

Fernando, S. D., Rodrigo, C., & Rajapakse, S. (2010). The "hidden" burden of malaria: Cognitive impairment following infection. *Malaria Journal*, *9*, 1–11. doi:10.1186/1475-2875-9-366

Fiszbein, A., & Schady, N. (2009). *Conditional cash transfers: Reducing present and future poverty*. Washington, DC: World Bank.

*Friedman, W., Gerard, F., & Ralaingita, W. (2010). *International independent evaluation of the effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) program in Mali: Mid-term report*. Research Triangle Park, NC: RTI International.

Friend, J., Searle, B., & Suppes, P. (Eds.). (1980). *Radio mathematics in Nicaragua*. Stanford, CA: Institute for Mathematical Studies in the Social Sciences, Stanford University.

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. *Review of Educational Research*, *64*, 119–157. doi:10.3102/00346543064001119

Galiani, S., & McEwan, P. J. (2013). The heterogeneous impact of conditional cash transfers. *Journal of Public Economics*, *103*, 85–96. doi:10.1016/j.jpubeco.2013.04.004

*Gardner, J. M., Grantham-McGregor, S., & Baddeley, A. (1996). Trichuris trichiura infection and cognitive function in Jamaican school children. *Annals of Tropical Medicine & Parasitology*, *90*, 55–63.

Glewwe, P. (2002). Schools and skills in developing countries: Education policies and socioeconomic outcomes. *Journal of Economic Literature*, *40*, 436–482. doi:10.1257/002205102320161258

Glewwe, P., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2011). *School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010* (Working Paper No. 17554). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w17554

*Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2*, 205–227. doi:10.1257/app.2.3.205

Glewwe, P., & Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 945–1017). Amsterdam, Netherlands: Elsevier. doi:10.1016/S1574-0692(06)02016-2

*Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics*, *1*, 112–135. doi:10.1257/app.1.1.112

*Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: The case of flip charts in Kenya. *Journal of Development Economics*, *74*, 251–268. doi:10.1016/j.jdeveco.2003.12.010

*Glewwe, P., & Maïga, E. (2011). *The impacts of school management reforms in Madagascar: Do the impacts vary by teacher type?* Unpublished manuscript, University of Minnesota, Minneapolis.

Glewwe, P., & Miguel, E. A. (2008). The impact of child health and nutrition on education in less developed countries. In T. P. Schultz & J. Strauss (Eds.), *Handbook of development economics* (*Vol. 4*, pp. 3561–3606). Amsterdam, Netherlands: Elsevier. doi:10.1016/S1573-4471(07)04056-9

*Glewwe, P., Park, A., & Zhao, M. (2011). *Visualizing development: Eyeglasses and academic performance in rural primary schools in China*. Unpublished manuscript, University of Minnesota, Minneapolis.

Grantham-McGregor, S., & Ani, C. (2001). A review of studies on the effect of iron deficiency on cognitive development in children. *Journal of Nutrition*, *131*(2 Suppl. 2), 649S–668S.

*Grigorenko, E. L., Sternberg, R. J., Jukes, M., Alcock, K., Lambo, J., Ngorosho, D., . . .Bundy, D. A. (2006). Effects of antiparasitic treatment on dynamically and statically tested cognitive skills over time. *Journal of Applied Developmental Psychology*, *27*, 499–526. doi:10.1016/j.appdev.2006.08.005

Hanushek, E. A. (1995). Interpreting recent research on schooling in developing countries. *World Bank Research Observer*, *10*, 227–246. doi:10.1093/wbro/10.2.227

*He, F., Linden, L. L., & MacLeod, M. (2008). *How to teach English in India: Testing the relative productivity of instruction methods with Pratham English Language Education Program*. Unpublished manuscript, Columbia University, New York, NY.

He, F., Linden, L. L., & MacLeod, M. (2009). *A better way to teach children to read? Evidence from a randomized controlled trial*. Unpublished manuscript, Columbia University, New York, NY.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. doi:10.3102/0162373707299706

Hermoso, M., Vucic, V., Vollhardt, C., Arsic, A., Roman-Viñas, B., Iglesia-Altaba, I., . . . Koletzko, B. (2011). The effect of iron on cognitive development and function in infants, children and adolescents: A systematic review. *Annals of Nutrition & Metabolism*, *59*, 154–165. doi:10.1159/000334490

Heyneman, S. P., Jamison, D. T., & Montenegro, X. (1984). Textbooks in the Philippines: Evaluation of a nationwide investment. *Educational Evaluation and Policy Analysis*, *6*, 139–150. doi:10.3102/01623737006002139

Inamdar, P. (2004). Computer skills development by children using "hole in the wall" facilities in rural India. *Australasian Journal of Educational Technology*, *20*, 337–350.

Jacoby, E., Cueto, S., & Pollitt, E. (1996). Benefits of a school breakfast programme among Andean children in Huaraz, Peru. *Food and Nutrition Bulletin*, *17*, 54–64.

Jacoby, H. G. (2002). Is there an intrahousehold "flypaper effect"? Evidence from a school feeding programme. *Economic Journal*, *112*, 196–221. doi:10.1111/1468-0297.0j679

*Jamison, D. T., Searle, B., Galda, K., & Heyneman, S. P. (1981). Improving elementary mathematics education in Nicaragua: An experimental study of the impact of textbooks and radio education. *Journal of Educational Psychology*, *73*, 556–567. doi:10.1037/0022-0663.73.4.556

*Jinabhai, C. C., Taylor, M., Coutsoudis, A., Coovadia, H. M., Tomkins, A. M., & Sullivan, K. R. (2001). A randomized controlled trial of the effect of antihelminthic treatment and micronutrient fortification on health status and school performance of rural primary school children. *Annals of Topical Paediatrics*, *21*, 319–333.

Jomaa, L. H., McDonnell, E., & Probart, C. (2010). School feeding programs in developing countries: Impacts on children's health and educational outcomes. *Nutrition Reviews*, *69*, 83–98. doi:10.1111/j.1753-4887.2010.00369.x

Jukes, M. C., Pinder, M., Grigorenko, E. L., Smith, H. B., Walraven, G., Bariau, E. M., . . . Bundy, D. A. P. (2006). Long-term impact of malaria chemoprophylaxis on cognitive abilities and educational attainment: Follow-up of a controlled trial. *PLoS Clinical Trials*, *1*, e19. doi:10.1371/journal.pctr.0010019

Kazianga, H., de Walque, D., & Alderman, H. (2012). Educational and child labor impacts of two food-for-education schemes: Evidence from a randomised trial in rural Burkina Faso. *Journal of African Economies*, *21*, 723–760. doi:10.1093/jae/ejs010

*Kleiman-Weiner, M., Luo, R., Zhang, L., Shi, Y., Medina, A., & Rozelle, S. (2013). Eggs versus chewable vitamins: Which intervention can increase nutrition and test scores in rural China? *China Economic Review*, *24*, 165–176. doi:10.1016/j.chieco.2012.12.005

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*, 297–300. doi:10.1126/science.1235350

Kremer, M., & Holla, A. (2009). Improving education in the developing world: What have we learned from randomized evaluations? *Annual Review of Economics*, *1*, 513–542. doi:10.1146/annurev.economics.050708.143323

*Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *Review of Economics and Statistics*, *91*, 437–456. doi:10.1162/rest.91.3.437

*Kremer, M., Moulin, S., & Namunyu, R. (2003). *Decentralization: A cautionary tale* (Poverty Action Lab Paper No. 10). Cambridge, MA: Poverty Action Lab.

Kristjansson, B., Robinson, V., Petticrew, M., MacDonald, B., Krasevec, J., Janzen, L., . . .Tugwell, P. (2006). School feeding for improving the physical and psychosocial health of disadvantaged students. *Cochrane Database of Systematic Reviews*, (14). doi:10.4073/csr.2006.14

Kvalsvig, J. D., Cooppan, R. M., & Connolly, K. J. (1991). The effects of parasite infections on cognitive processes in children. *Annals of Tropical Medicine and Parasitology*, *85*, 551–568.

*Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. (2011). *Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing* (Working Paper No. 228). Stanford, CA: Rural Education Action Project.

Lai, F., Zhang, L., Hu, X., Qu, Q., Shi, Y., Boswell, M., & Rozelle, S. (2012). *Computer assisted learning as extracurricular tutor? Evidence from a randomized experiment in rural boarding schools in Shaanxi* (Working Paper No. 235). Stanford, CA: Rural Education Action Project.

*Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M., & Rozelle, S. (2012). *Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in public schools in rural minority areas in Qinghai* (Working Paper No. 237). Stanford, CA: Rural Education Action Project.

*Lassibille, G., Tan, J.-P., Jesse, C., & Nguyen, T. V. (2010). Managing for results in primary education in Madagascar: Evaluating the impact of selected workflow interventions. *World Bank Economic Review*, *24*, 303–329. doi:10.1093/wber/lhq009

Lee, D. S., & Lemieux, T. (2009). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281–355. doi:10.1257/jel.48.2.281

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

*Li, T., Han, L., Rozelle, S., & Zhang, L. (2010). *Cash incentives, peer tutoring, and parental involvement: A study of three educational inputs in a randomized field experiment in China* (Working Paper No. 221). Stanford, CA: Rural Education Action Project.

Lien, D. T. K., Nhung, B. T., Khan, N. C., Hop, L. T., Nga, N. T. Q., Hung, N. T., . . . te Biesebeke, R. (2009). Impact of milk consumption on performance and health of primary school children in rural Vietnam. *Asia Pacific Journal of Clinical Nutrition*, *18*, 326–334.

*Linden, L. L. (2008). *Complement or substitute? The effect of technology on student achievement in India*. Unpublished manuscript, Columbia University, New York, NY.

Liu, C., Yi, H., Luo, R., Bai, Y., Zhang, L., Shi, Y., . . .Rozelle, S. (2013). *The effect of early commitment of financial aid on matriculation to senior high school among poor junior high students in rural China* (Working Paper No. 254). Stanford, CA: Rural Education Action Project.

Lockheed, M. E., & Verspoor, A. M. (1991). *Improving primary education in developing countries*. Oxford, England: Oxford University Press.

*Loyalka, P., Liu, C., Song, Y., Yi, H., Huang, X., Wei, J., . . . Rozelle, S. (2013). *Can information and counseling help students from poor rural areas go to high school? Evidence from China* (Working Paper No. 241). Stanford, CA: Rural Education Action Project.

*Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda. *Journal of Policy Analysis and Management, 33*, 950–976

*Luo, R., Shi, Y., Zhang, L., Liu, C., Rozelle, S., Sharbono, B., . . . Martorell, R. (2012). Nutrition and educational performance in rural China's elementary schools: Results of a randomized control trial in Shaanxi Province. *Economic Development and Cultural Change*, *60*, 735–772. doi:10.1086/665606

*Manger, M. S., McKenzie, J. E., Winichagoon, P., Gray, A., Chavasit, V., Pongcharoen, T., . . . Gibson, R. (2008). A micronutrient-fortified seasoning powder reduces morbidity and improves short-term cognitive function, but has no effect on anthropometric measures in primary school children in northeast Thailand: A randomized controlled trial. *American Journal of Clinical Nutrition*, *87*, 1715–1722.

McEwan, P. J. (2012). Cost-effectiveness analysis of education and health interventions in developing countries. *Journal of Development Effectiveness*, *4*, 189–213. doi:10.1080/19439342.2011.649044

McEwan, P. J. (2013). The impact of Chile's school feeding program on education outcomes. *Economics of Education Review*, *32*, 122–139. doi:10.1016/j.econedurev.2012.08.006

*Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*, 159–217. doi:10.1111/j.1468-0262.2004.00481.x

*Mo, D., Swinnen, J., Zhang, L., Yi, H., Qu, Q., Boswell, M., & Rozelle, S. (2012). *Can One Laptop per Child reduce the digital divide and educational gap? Evidence from a randomized experiment in migrant schools in Bejing* (Working Paper No. 233). Stanford, CA: Rural Education Action Project.

*Mo, D., Zhang, L., Lui, R., Qu, Q., Huang, W., Wang, J., . . . Rozelle, S. (2013). *Integrating computer assisted learning into a regular curriculum: Evidence from a randomized experiment in rural schools in Shaanxi* (Working Paper No. 248). Stanford, CA: Rural Education Action Project.

*Muralidharan, K., & Sundararaman, V. (2010a). *Contract teachers: Experimental evidence from India*. Unpublished manuscript, University of California, San Diego.

*Muralidharan, K., & Sundararaman, V. (2010b). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India. *Economic Journal*, *120*, F187–F203. doi:10.1111/j.1468-0297.2010.02373.x

*Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*, 39–77. doi:10.1086/659655

Newman, J., Pradhan, M., Rawlings, L. B., Ridder, G., Coa, R., & Evia, J. L. (2002). An impact evaluation of education, health, and water supply investments by the Bolivian Social Investment Fund. *World Bank Economic Review*, *16*, 241–274. doi:10.1093/wber/16.2.241

Nga, T. T., Winichagoon, P., Dijkhuizen, M. A., Khan, N. C., Wasantwisut, E., & Wieringa, F. T. (2011). Decreased parasite load and improved cognitive outcomes caused by deworming and consumption of multi-micronutrient fortified biscuits in rural Vietnamese schoolchildren. *American Journal of Tropical Medicine and Hygiene*, *85*, 333–340. doi:10.4269/ajtmh.2011.10-0651

*Nguyen, T. (2008). *Information, role models, and perceived returns to education: Experimental evidence from Madagascar*. Unpublished manuscript, MIT, Cambridge, MA.

Nitsaisook, M., & Anderson, L. W. (1989). An experimental investigation of the effectiveness of inservice teacher education in Thailand. *Teaching and Teacher Education*, *5*, 287–302. doi:10.1016/0742-051X(89)90027-9

*Nokes, C., Grantham-McGregor, S. M., Sawyer, A. W., Cooper, E. S., Robinson, B. A., & Bundy, D. A. P. (1992). Moderate to heavy infections of trichuris trichiura affect cognitive function in Jamaican school children. *Parasitology*, *104*, 539–547. doi:10.1017/S0031182000063800

*Osendarp, S. J. M., Baghurst, K. I., Bryan, J., Calvaresi, E., Hughes, D., Hussaini, M., . . . Wilson, C. (2007). Effect of a 12-mo micronutrient intervention on learning and memory in well-nourished and marginally nourished school-aged children: 2 parallel, randomized, placebo-controlled studies in Australia and Indonesia. *American Journal of Clinical Nutrition*, *86*, 1082–1093.

*Oster, E., & Thornton, R. (2009). *Menstruation and education in Nepal* (Working Paper No. 14853). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w14853

Pandey, P., Goyal, S., & Sundararaman, V. (2009). Community participation in public schools: Impact of information campaigns in three Indian states. *Education Economics*, *17*, 355–375. doi:10.1080/09645290903157484

Petrosino, A., Morgan, C., Fronius, T. A., Tanner-Smith, E. E., & Boruch, R. F. (2012). Interventions in developing nations for improving primary and secondary school enrollment of children: A systematic review. *Cochrane Database of Systematic Reviews*, (19). doi:10.4073/csr.2012.19

Piper, B., & Korda, M. (2011). *EGRA Plus: Liberia: program evaluation report*. Research Triangle Park, NC: RTI International.

Pollitt, E., Hathirat, P., Kotchabhakdi, N. J., Missell, L., & Valyasevi, A. (1989). Iron deficiency and educational achievement in Thailand. *American Journal of Clinical Nutrition, 50*, 687–697.

*Powell, C. A., Walker, S. P., Chang, S. M., & Grantham-McGregor, S. M. (1998). Nutrition and education: A randomized trial of the effects of breakfast in rural primary school children. *American Journal of Clinical Nutrition*, *68*, 873–879.

*Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Alishjabana, A., Gaduh, A., & Artha, R. P. (2011). *Improving educational quality through enhancing community participation: Results from a randomized field experiment in Indonesia* (Policy Research Working Paper No. 5795). Washington, DC: World Bank.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–333). New York, NY: Russell Sage Foundation.

*Rico, J. A., Kordas, K., López, P., Rosado, J. L., García Vargas, G., Ronquillo, D., & Stolzfus, R. J. (2006). Efficacy of iron and/or zinc supplementation on cognitive performance of lead-exposed Mexican schoolchildren: A randomized, placebo-controlled trial. *Pediatrics*, *117*, e518–e527. doi:10.1542/peds.2005-1172

Ringquist, E. J. (2013). *Meta-analysis for public management and policy*. San Francisco, CA: Jossey-Bass.

Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., Flores, P., . . . Salinas, M. (2003). Beyond Nintendo: Design and assessment of educational video games for first and second grade students. *Computers & Education*, *40*, 71–94. doi:10.1016/S0360-1315(02)00099-4

Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs* (MPR Reference No. 6046-310). Princeton, NJ: Mathematica Policy Research.

Seshadri, S., & Gopaldas, T. (1989). Impact of iron supplementation on cognitive functions in preschool and school-aged children: The Indian experience. *American Journal of Clinical Nutrition*, *50*, 675–686.

*Sharma, D. (2010). *The impact of financial incentives on academic achievement and household behavior: Evidence from a randomized trial*. Unpublished manuscript, The Ohio State University, Columbus.

*Simeon, D. T., Grantham-McGregor, S. M., Callender, J. E., & Wong, M. S. (1995). Treatment of trichuris trichiura infections improves growth, spelling scores and school attendance in some children. *Journal of Nutrition*, *125*, 1875–1883.

*Simeon, D. T., Grantham-McGregor, S. M., & Wong, M. S. (1995). Trichuris trichiura infection and cognition in children: Results of a randomized clinical trial. *Parasitology*, *110*, 457–464. doi:10.1017/S0031182000064799

*Soemantri, A. G. (1989). Preliminary findings on iron supplementation and learning achievement of rural Indonesian children. *American Journal of Clinical Nutrition*, *50*, 698–702.

*Soemantri, A. G., Pollitt, E., & Kim, I. (1985). Iron deficiency anemia and educational achievement. *American Journal of Clinical Nutrition*, *42*, 1221–1228.

*Solon, F. S., Sarol, J. N., Jr., Bernardo, A. B. I., Solon, J. A. A., Mehansho, H., Sanchez-Fermin, L. E., . . . Juhlin, K. D. (2003). Effect of a multiple-micronutrient-fortified fruit powder beverage on the nutrition status, physical fitness, and cognitive performance of schoolchildren in the Philippines. *Food and Nutrition Bulletin, 24*(4 Suppl.), S129–S140.

Sungthong, R., Mo-suwan, L., Chongsuvivtwong, V., & Geater, A. F. (2004). Once-weekly and 5-days a week iron supplementation differentially affect cognitive function but not school performance in Thai children. *Journal of Nutrition*, *134*, 2349–2354.

*Sylvia, S., Luo, R., Zhang, L., Shi, Y., Medina, A., & Rozelle, R. (2012). *Do you get what you pay for with school-based health programs? Evidence from a child nutrition experiment in rural China* (Working Paper No. 246). Stanford, CA: Rural Education Action Project.

*Tan, J.-P., Lane, J., & Lassibille, G. (1999). Student outcomes in Philippine elementary schools: An evaluation of four experiments. *World Bank Economic Review*, *13*, 493–508. doi:10.1093/wber/13.3.493

Taylor-Robinson, D. C., Maayan, N., Soares-Weiser, K., Donegan, S., & Garner, P. (2012). Deworming drugs for soil-transmitted intestinal worms in children: Effects on nutritional indicators, haemoglobin and school performance (review). *Cochrane Database of Systematic Reviews*, (7). doi:10.1002/14651858.CD000371.pub5

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *113*, F3–F33. doi:10.1111/1468-0297.00097

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural Bolivia. *Review of Economics and Statistics*, *88*, 171–177. doi:10.1162/rest.2006.88.1.171

*Van Stuijvenberg, M. E., Kvalsig, J. D., Faber, M., Kruger, M., Kenoyer, D. G., & Spinnler Benadé, A. J. (1999). Effect of iron-, iodine-, and β-carotene-fortified biscuits on the micronutrient status of primary school children: A randomized controlled trial. *American Journal of Clinical Nutrition*, *69*, 497–503.

Vazir, S., Nagalla, B., Thangiah, V., Kamasamudram, V., & Bhattiprolu, S. (2006). Effect of micronutrient supplement on health and nutritional status of schoolchildren: Mental function. *Nutrition, 22*(1 Suppl.), S26–S32. doi:10.1016/j.nut.2004.07.021

Vegas, E., & Petrow, J. (2008). *Raising student learning in Latin America: The challenge for the 21st century*. Washington, DC: World Bank.

Velez, E., Schiefelbein, E., & Valenzuela, J. (1993). *Factors affecting achievement in primary school: A review of the literature for Latin America and the Caribbean* (HRO Working Paper No. 2). Washington, DC: World Bank.

Vermeersch, C., & Kremer, M. (2004). *School meals, educational achievement, and school competition: Evidence from a randomized evaluation* (Policy Research Working Paper No. 3523). Washington, DC: World Bank.

*Watkins, W. E., Cruz, J. R., & Pollitt, E. (1996). The effects of deworming on indicators of school performance in Guatemala. *Transactions of the Royal Society of Tropical Medicine & Hygiene*, *90*, 156–161. doi:10.1016/S0035-9203(96)90121-2

Whaley, S. E., Sigman, M., Neumann, C., Bwibo, N., Guthrie, D., Weiss, R. E., . . . Murphy, S. P. (2003). The impact of dietary intervention on the cognitive development of Kenyan school children. *Journal of Nutrition*, *133*(11 Suppl. 2), 3965S–3971S.

What Works Clearinghouse. (2011). *Procedures and standard handbook* (Version 2.1). Washington, DC: Author. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

*Yi, H., Song, Y., Liu, C., Huang, X., Zhang, L., Bai, Y., . . . Rozelle, S. (2012). *Giving kids a head start: The impact of early commitment of financial aid on poor seventh grade students in rural China* (Working Paper No. 247). Stanford, CA: Rural Education Action Project.

*Zhang, L., Lai, F., Pang, X., Yi, H., & Rozelle, R. (2012). *The impact of teacher training on teacher and student outcomes: A randomized experiment in Bejing migrant schools* (Working Paper No. 236). Stanford, CA: Rural Education Action Project.

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*, 242–278.

Author

PATRICK J. McEWAN is a professor of Economics, Department of Economics, Wellesley College, 106 Central Street, Wellesley, MA 02481; e-mail: *pmcewan@wellesley.edu*. His research interests include the economics of education, Latin American education policy, and the impact and cost evaluation of education, health, and welfare policies in developing countries. For more information on his research, visit www.patrickmcewan.net.