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

Patrick J. McEwan*
Wellesley College

August 2013

Abstract: I identified and coded 76 randomized experiments conducted in developing-country primary schools from the mid-1970s to 2013. The experiments evaluated the impact of 110 school-based treatments on language and mathematics test scores, as compared with “business-as-usual” in the same settings. The treatments included instructional interventions, health interventions, and incentive-based interventions. On average, monetary grants and deworming had effects that were close to zero and statistically insignificant. Nutritional treatments, treatments that provided information to parents or students, and treatments that improved school management and 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 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); student and teacher performance incentives (0.10); and contract or volunteer teachers (0.10). Meta-regressions that controlled for treatment heterogeneity and other moderators suggested that the effects of materials and contract teachers, in particular, were partly accounted for by composite treatments that also included teacher training and class size reduction. A caveat is that interventions like deworming and school lunches often affected enrollment and attainment independently of learning, implying that student time is not always used productively in schools. There is insufficient data to gauge the relative cost-effectiveness of categories of interventions.

* pmcewan@wellesley.edu. I am grateful to the Quality Education in Developing Countries initiative of the William and Flora Hewlett Foundation for financial support. Adrienne Lucas, Chloe O’Gara, Maria Perez, Pat Scheid, Dana Schmidt, and Rebecca Thornton provided helpful advice and comments, although they are not responsible for any errors or interpretations. Maria Perez, Kate Kemmerer, and Poppy Tian provided excellent support in the design and coding of the database. The data used in this paper can be consulted at www.patrickmcewan.net/meta.
There is a vast non-experimental literature on school effectiveness in developing countries (for reviews, see Velez, Schiefelbein, & Valenzuela, 1993; Fuller & Clarke, 1994; Hanushek, 1995; Kremer, 1995; Glewwe, 2002). It uses regression analysis to identify the putative determinants of student learning, but it is hampered in two ways. First, it does not credibly distinguish between the causal effects of schools, and the confounding effects of the children and families that happen to attend those schools. Second, many datasets contain only simple proxies of school quality, such as teacher credentials and pupil-teacher ratios, that do not encompass the complex menu of investment choices available to policy-makers.

Experimental research addresses both issues. The use of random assignment of students or schools to school-based treatments improves the internal validity of causal inferences (Glewwe & Kremer, 2006; Duflo, Glennerster, & Kremer, 2008). Moreover, researchers have evaluated a rich variety of policy-relevant treatments that encompass (1) instructional interventions that combine teacher training, textbooks and other materials, class size reduction, and computer-assisted instruction; (2) school-based health and nutrition treatments, such as de-worming and micronutrient supplementation; and (3) interventions that modify stakeholder incentives to improve learning, such as school report cards, performance pay, and school-based management.

It is a propitious moment to survey this literature, and assess whether there are lessons for policy-makers and researchers.

I conducted a literature search in economics, education, and health, identifying 76 published and unpublished experiments that evaluate 110 treatments (see the Appendix). In each case, researchers randomly assigned children, schools, or entire villages to receive a school-based
treatment, versus “business-as-usual” in the same setting.¹ I coded effect sizes and their standard errors for outcome variables in language and mathematics. I further coded study attributes that describe the category of treatment, details on the country and experimental sample, outcome measures, and the study quality.

Two categories of interventions—monetary grants and school-based deworming—have mean effects that are close to zero and statistically insignificant (based on a random effects model). 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. 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); student and teacher performance incentives (0.10); and contract or volunteer teachers (0.10). The categories are not mutually exclusive, however, and meta-regressions 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 (Glewwe et al., 2004, 2009), and contract and volunteer teacher interventions usually overlap with class size reduction and/or instructional treatments (e.g., Banerjee et al., 2007; Bold et al., 2012).

¹ I initially sought to identify studies that used a regression-discontinuity design as well, since well-designed studies have strong internal validity (Lee & Lemieux, 2009; What Works Clearinghouse, 2011). I found only a few papers analyzing primary-school learning, including evaluations of after-school tutoring (Chay, McEwan, & Urquiola, 2005), class size reduction (Urquiola, 2006), and school meals (McEwan, 2013).
A challenge to the interpretation of learning effects is that some treatments, particularly deworming and school feeding programs, affect enrollment and attainment, despite weaker effects on learning (Miguel & Kremer, 2004; Baird et al., 2012; Petrosino et al., 2013). Even when attainment increases, it is plausible that student time is not being used productively in classrooms. This suggests that interventions with a primary focus on access should be combined—in future research and in practice—with interventions explicitly designed to increase learning.

Most papers 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 et al. (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, the inability to statistically distinguish between ranked cost-effectiveness ratios, and the evidence—cited above—that many treatments affect additional outcomes such as attainment.

The review also suggests methodological lessons for the conduct of future experiments. The overwhelming majority of instructional and incentive-based experiments use cluster-randomized assignment of schools (or groups of schools) to a treatment. The statistical power of these experiments is primarily determined by the number of clusters and the intraclass correlation of the outcome variable. The intraclass correlations for test scores are higher in developing-country

---

2 The earlier, non-experimental literature also provided little guidance on the relative cost-effectiveness of investments in learning, although several studies attempted to bridge this gap (Harbison & Hanushek, 1992; Fuller et al., 1994; Glewwe, 1999).
settings than typical U.S. standards (Zopluoglu, 2012). Further, this paper shows that “typical” effect sizes of some categories of treatments are smaller than commonly assumed. Viewed together, the evidence suggests many smaller cluster-randomized experiments—particularly those with fewer than 50 schools per treatment arm—are under-powered.

I further argue that experiments can enhance the potential external validity of their results by (1) adding treatment arms that manipulate key features of the treatment (such as the implementing agency); (2) experimenting within representative samples, and conducting subgroup analysis within policy-relevant and pre-specified subgroups of the full sample; (3) measuring a wider range of learning outcomes and not “cherry-picking” effects across those outcomes; (4) collecting process data that can be used to conduct (non-experimental) tests of the potential causal mechanisms of “black-box” experimental effects; and (5) complementing their findings with high-quality quasi-experimental research that evaluates scaled-up treatments using representative samples of schools. I further suggest that experimental reports, especially in economics, could benefit from a common reporting standard, along the lines of the CONSORT standards widely used in medicine.

The next section describes the conceptual framework that organizes the review, including a typology of school-based treatments used in the coding of studies. I then describe data and methods, including the literature search, the criteria for study inclusion and exclusion, the coding of experiments, and the statistical methods used to analyze effect sizes. The results section describes mean effect sizes by categories of treatment, as well as meta-regression models that analyze the correlates of those effects. The final section reviews lessons for policy and future experimental research.
School-Based Treatments to Improve Learning

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: (1) parent endowments, including parents’ schooling and abilities; (2) parent-provided education inputs like supplemental instruction; (3) child endowments, such as nutrition and health, but potentially including a wider range of cognitive or non-cognitive abilities; and (4) a variety of schools and teacher inputs.

School-provided inputs 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 pre-service or in-service training in pedagogy or content, by experience, or by teachers’ 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 determined by extrinsic motives (such as rewards or sanctions for poor performance) or intrinsic ones (such as 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 official proportion of time devoted to learning-related activities, and the local choices of many households and teachers.
In early research, researchers estimated the parameters of production functions using non-experimental data and regression analysis. Typically, they regressed measures of child learning on variables such as mother’s schooling, the availability of textbooks, and teachers’ years of experience and education credentials. Reviewers often noted potential biases in the estimates of key parameters (Kremer, 1995; Glewe, 2002; Glewe & 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 biased upward. Glewe et al. (2004) provide an example of such bias using Kenyan experimental and non-experimental data.

The random assignment of an education or health intervention strengthens the internal validity of inferences about causal effects, 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 (Todd & Wolpin, 2003; Glewe & Kremer, 2006).

To illustrate this in Figure 1, consider three examples. First, the in-school provision of micronutrients or calories may improve child nutrition or health, which may directly improve learning (Powell et al., 1998; Kazianga et al., 2013; McEwan, 2013). Since meals are an in-kind transfer to children, conditional upon regular school attendance, they may spur attendance. Thus,

---

3 The results are summarized in meta-analyses that rely on a vote-counting method—that is, the tabulation of positive and negative regression coefficients by category of inputs, and their further division by the coefficients’ statistical significance at conventional levels (Velez, Schiefelbein, & Valenzuela, 1993; Fuller & Clarke, 1994; Hanushek, 1995; Kremer, 1995).
lunches could affect learning by increasing the quantity of instructional time. But, households may react to the availability of free lunches by reallocating food within the household, perhaps towards needier siblings (Jacoby, 2002).

Second, monetary grants may finance the purchase of school-provided instructional inputs, which could directly affect learning (Blimpo & Evans, 2011; Das et al., 2011; Pradhan et al., 2011). When anticipated by parents, however, school grants create incentives for parents to reduce their own spending on household inputs (Das et al., 2011). Third, smaller class sizes potentially increase the quantity of instructional time by increasing the amount of time that teachers can spend with individual students (Betts & Shkolnik, 1999) or by reducing the probability of disruptions in large-group settings (Lazear, 2001). At the same, smaller classes are a non-pecuniary working condition that could improve teachers’ motivation to exert effort, though they may encourage rent-seeking behavior as teachers adjust effort downward (Duflo, Dupas, & Kremer, 2012). In all three examples, the policy effect of a school-based intervention could encompass multiple direct or indirect effects on learning. Most experiments cannot disentangle these mechanisms, at least with the same causal validity of the policy effect.

A Typology of School-Based Treatments

Reviewers use typologies to summarize the evidence on diverse treatments. Pragmatically, a typology should not be so general as to combine vastly different treatments into a single category (“school inputs”), nor so detailed that it leads to a nearly-saturated regression models when treatment-related independent variables are used to explain the magnitude of effect sizes (i.e., a specific brand of laptop or textbook series). The next paragraphs describe treatments within three categories, providing examples from the meta-analytic sample.
**Instructional treatments.** Instructional treatments improve the quantity or quality of classroom instruction by marshalling instructional materials, computers, and teacher training, and by manipulating the size and composition of learning groups. First, treatments provide instructional materials to teachers and schools, such as textbooks (Jamison et al., 1981; Glewwe et al., 2009), flipcharts (Glewwe et al., 2004), flashcards (He et al., 2009), classroom libraries (Abeberese et al., 2012); and teacher materials such as curriculum guides or lesson plans (Banerjee et al., 2007).

Second, treatments endow classrooms with computers, software, or other instructional technology. In Peru and China, respectively, laptops with pre-loaded software were delivered to schools (Cristia et al., 2012), or to parents after providing a minimal training session (Mo et al., 2012). But, in most cases, desktop computers are located in classrooms, and teachers monitor students’ use of tutoring software (Linden, 2008; Mo et al., 2013).

Third, higher-level education authorities disburse grants to school committees (Pradhan et al., 2011) and school personnel (Das et al., 2011; Blimpo & Evans, 2011), to support local purchases of instructional inputs.

Fourth, teachers receive in-service training designed to increase their capacity to deliver effective instruction. Ecuador’s *Más Tecnología* program endowed classrooms with computers and tutoring software, and trained primary school teachers in computer use (Carillo et al., 2010). Kenyan and Ugandan programs trained teachers in the Reading to Learn methodology, an instructional model designed to increase early-grade literacy (Lucas et al., 2013).

Fifth, treatments modify the size and ability composition of instructional groups. This includes common treatments such as class size reduction (Duflo et al., 2012; Krueger, 1999), as well as treatments that provide instruction to small groups of students within relatively larger
classes. For example, Chilean volunteers provided supplemental reading instruction during the school day to small groups of 5 or 6 students, but with no apparent ability grouping (Cabezas et al., 2011). In India, trained secondary-school graduates provided supplemental instruction—during school hours—to groups of 15 to 20 children who had been identified as low-achieving (Banerjee et al., 2007). Only one experiment “tracked” entire classrooms by ability, without modifying their size (Duflo et al., 2011).

**Health treatments.** The most common health-related treatments are nutritional interventions in schools, including iron and micronutrient supplements (Luo et al., 2012) and the in-school provision of food or beverages (Powell et al., 1998; Osendarp et al., 2007). A smaller group of experiments treats children for intestinal helminths (Miguel & Kremer, 2004; Grigorenko, 2006) or, less commonly, administers drugs to prevent malaria (Fernando et al., 2006). A residual category of health treatments provides vision screening and eyeglasses (Glewwe, Park, & Zhao, 2009) and menstrual cups (Oster & Thornton, 2009).

**Incentive treatments.** Four categories of treatments modify incentives for students, parents, or school personnel to improve student achievement. First, treatments disseminate information on student performance to teachers or school officials (Andrabi, Das, & Khwaja, 2009; Muralidharan & Sundararaman, 2010b; Glewwe & Maïga, 2011), to school management committees or parents (Andrabi et al., 2009; Barr et al., 2012; Glewwe & Maïga, 2011), or directly to students (Nguyen, 2008; Loyalka, 2013). The information is usually summarized in

---

---

4 There are several potential causal mechanisms. 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
“report cards,” sometimes devised in consultation with local stakeholders (Andrabi et al., 2009; Barr et al., 2012). The report cards sometimes include information on children’s health (Luo et al., 2012; Sylvia et al., 2012).

Second, treatments explicitly link rewards to objective performance measures. In some cases, school personnel receive monetary or in-kind rewards for higher student test scores (Muralidharan & Sundararaman, 2011), attendance (Duflo, Hanna, & Ryan, 2012), and student health (Sylvia et al., 2012). In others, high-performing students are offered cash rewards (Li et al., 2010) or future scholarships (Kremer, Miguel & Thornton, 2009; Yi et al., 2012).

Third, treatments encourage the recruitment and hiring of teachers with flexible labor contracts, often locally-hired “contract” teachers outside the civil service (Muralidharan & Sundararaman, 2010a; Bold et al., 2012; Duflo et al., 2012). In other cases, teachers are hired and trained by NGOs (Banerjee et al., 2007; He et al., 2008) or work as volunteers (Cabezas et al., 2012; Banerjee et al., 2012). Contract teachers and volunteers might be supposed to have stronger incentives to attend class regularly and deliver effective instruction, since they can be terminated for non-performance. Even so, the typical experimental design makes it challenging to separate incentive effects from those of concomitant reductions in class size (Muralidharan & Sundararaman, 2010a; Bold et al., 2012; Duflo et al., 2012), instructional materials and training (Banerjee et al., 2007, 2012; He et al., 2008; Cabezas et al., 2012), or simply from pre-treatment differences in the capacities of regular and contract teachers.

Fourth, a diffuse category of treatments attempts to improve the management and supervision of schools. For example, school personnel and school committee members received training in school management (Blimpo & Evans, 2011; Pradhan et al., 2011) and in the hiring, monitoring, about the relationship between their current performance and future earnings may have improved incentives to exert effort in the short-run.
and assessment of teacher performance (Duflo et al., 2012). Besides training, an Indonesian experiment organized elections of local school committee members, and increased linkages between school committees and local politicians (Pradhan et al., 2011). Finally, an experiment in Madagascar provided management training to subdistrict and district education officials, in addition to local school personnel (Lassibille et al., 2010; Glewwe & Maïga, 2011).

Data and Methods

Literature Search

I conducted a literature search between August 2012 and February 2013. I first examined the references of two meta-analyses of randomized and non-randomized evaluations of education interventions in developing countries (Glewwe et al., 2011; Petrosino et al., 2012). Each review employed a well-documented keyword search of scholarly databases such as Econlit, Eric, and Medline, although Petrosino et al. (2012) focuses 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, usually conducted by economists (Fuller & Clarke, 1994; Glewwe, 2002; Glewwe & Kremer, 2006; Evans & Ghosh, 2008; Glewwe & Miguel, 2008). Two of the most recent emphasize randomized impact evaluations (Kremer & Holla, 2009; Kremer, Brannen, & Glennerster, 2013).

To increase coverage of the literature on school-based health and nutrition interventions, I consulted meta-analyses and narrative reviews in nutrition and medicine. The treatments included de-worming medications (Dickson et al., 2000; Taylor et al., 2012); iron supplementation (Grantham-McGregor & Ani, 2001; Falkingham et al., 2011; Hermoso et al.,
2011), multiple micronutrient supplementation (Eilander et al., 2010; Best et al., 2011); malaria medications (Fernando, Rodrigo, & Rajapakse, 2010); and school feeding programs (Kristjansson et al., 2006; Jomaa, McDonnell, & Probart, 2010). Finally, I consulted recent World Bank reports on learning in developing countries (Vegas & Petrow, 2008; Bruns, Filmer, & Patrinos, 2011).


Criteria for Study Inclusion and Exclusion

I included studies if they were: (1) conducted in a lower to upper-middle-income country, as defined by the World Bank in 2013; (2) conducted in primary schools, broadly defined to include grades 1 to 8 (or ages 6 to 14, if the grades were not reported); (3) 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; (4) reported results for at least one continuously-measured learning outcome in language or reading, mathematics, or a composite assessment including either outcome; and (5) 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 (Jukes et al., 2006; He, Linden, & MacLeod, 2009) or secondary grades (Angrist et al., 2002; Angrist, Bettinger, & Kremer, 2006; Blimpo, 2010; Liu et al., 2013). Some studies did not randomly assign units to treatment and control groups (Nitsaisook & Anderson, 1989; Rosas et al., 2003; Inamdar, 2004; Heyneman, Jamison, & Montenegro, 1984). Berry (2012) compared two incentive treatments, but was unable to include a “business-as-usual” control group. Banerjee et al. (2010) reports results for binary indicators of learning outcomes, while the outcome measures in several papers did not include language or mathematics (Seshadri & Gopaldas, 1989; Kvalsig, Cooppan, & Connolly, 1991; Newman et al., 2002; Clarke et al., 2008; Lien et al., 2009; Beuermann et al., 2012).

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 et al., 1989; Whaley et al., 2003; Sunthong et al., 2004), in one case because statistically insignificant results were not reported in the paper (Nga et al., 2011). In other cases, I could not

6 In several instances, authors used a method of quasi-random assignment, such as alternating treatment-control assignment from an alphabetized list of clusters (e.g., Miguel & Kremer, 2004). I included these studies, but coded the study attribute for subsequent analysis.

7 Jacoby, Cueto, and Pollitt (1996) report sufficient descriptive statistics to calculate an effect size. Their Table 6 further reports that treatment effects from a model adjusting for baseline scores are not statistically significant at conventional levels, but no standard error is reported.
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 (Vazir et al., 2006; Adelman et al., 2008; Pandey, Goyal, & Sundararaman, 2009; Kazianga, de Walque, & Alderman, 2012).

Finally, I excluded cluster-randomized experiments in which standard errors did not correctly account for the unit of assignment.\(^8\) Chandler et al. (1995) randomly assigned 7 of 14 classrooms to a school feeding intervention, but did not cluster standard errors. Lai, Zhang, Hu, et al. (2012) randomly assigned schools to a computer-assisted instruction intervention, but clustered standard errors at the level of classrooms in the full-sample results. Finally, Piper and Medina (2011) randomly assigned 45 pre-defined groups of school to two treatment arms and a control group, but clustered results at the level of approximately 180 schools within the groups. When relevant, I refer to excluded studies in the narrative discussion of meta-analytic results.

**Coding of Experiments**

**Experiments and papers.** The dataset includes six groups of variables that describe (1) experiments, (2) papers, (3) treatment arms, (4) follow-ups, (5) outcome measures, and (6) effects.\(^9\) For coding purposes, I define 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

---

\(^8\) The most common approach is to estimate Huber-White standard errors clustered at the level of randomization (Duflo, Glennerster, & Kremer, 2008).

\(^9\) The database is maintained in Filemaker Pro, using a separate table for each group of variables. A searchable version is available at www.patrickmcewan/meta.
(1996) randomly assigned 125 primary-grade students in Guatemala to receive de-worming medication, and 125 to receive a placebo.

Sometimes one experiment is reported in multiple papers. Muralidharan and Sundararaman (2011) randomly assigned 500 schools to four treatment arms and a control group. The cited paper includes results on two performance incentive treatments, while Muralidharan and Sundararaman (2010a) and Das et al. (2011) report evidence on a contract teacher treatment and a block grant treatment, respectively. In cases where results from a single experiment are replicated in more than one paper, I used a preferred set of estimates. 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 et al., 2007, 2012; He, Linden, & MacLeod, 2008). More peculiarly, 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 (see their Table 1). 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 (see their Table 4). I code this as seven “experiments,” given the varying composition of each control group.

---

10 A fourth paper 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 and 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 statistical analyses account for possible statistical dependencies across the effect sizes from the two experiments, given the shared samples.

11 For example, a Malagasy experiment assigned some school districts to receive school reports cards and management training, and others to receive none. 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 et al. (2010).
Treatment arms. The coded attributes of treatment arms included sample sizes, treatment duration, implementing agencies, and its location within the typology. Most experiments included one treatment arm, but an experiment from Madagascar included seven, consisting of teacher-provided information for parents on the economic returns to schooling; information provided by three kinds of role models; and the combination of information and the three role-model treatments (Nguyen, 2008).

Follow-ups and outcome measures. Each treatment-control contrast is accompanied by as many as three follow-up data collections. I coded variables on each follow-up, including the date of data collection and differential attrition between the treatment and control group at the time of follow-up. Most experiments measure 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 (Friedman, Gerard, & Ralaingita, 2010). In these cases, I only coded the main outcome measure, unless the paper only reports subtest results (Cabezas et al., 2011). I further coded whether the outcome measures language or reading achievement, mathematics, or a composite, as well as the source of assessment items (e.g., 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 additional variables that are plausibly unaffected by the treatment, such as a pre-test.

**Unconditional effect sizes.** The literature on meta-analysis emphasizes that effects for continuous variables (such as 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 outcomes—measured at the follow-up—between treatment and control groups, divided by the sample standard deviation of the pooled treatment and control samples:

$$d = \frac{\bar{Y}_T - \bar{Y}_C}{s_{pooled}}$$

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. The literature commonly applies a small-sample correction to $d$ and its standard error, resulting in Hedges’ $g$ (Borenstein, 2009). In practice, the correction makes no difference in the pattern of this paper’s results.

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

$$O_{ij} = \beta_0 + \beta_1 T_{ij} + \varepsilon_{ij}$$  \hspace{1cm} (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 (versus the control), and $\beta_1$ represents the mean difference between
treatment and control groups, also interpreted as the effect of the intention-to-treat.\footnote{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 variables (IV) estimates of the local average treatment effect on those whose treatment status was influenced by random assignment to the treatment group. See Duflo et al. (2008) for technical details, and Banerjee et al. (2007) for an example. I excluded one paper because it only reports IV estimates (Evans, Kremer, & Ngatia, 2009).} 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 paper.

It is common to stratify the units of assignment—whether students or schools—by pre-treatment characteristics of the sample such as location or poverty. Then, random assignment occurs within each stratum (Duflo et al., 2008; Bruhn & McKenzie, 2009). 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 pair-wise matching. A more suitable specification would control for dummy variables indicating strata or pairs:

\[ O_{ijk} = \beta_0 + \beta_1 T_{ijk} + \delta_k + \varepsilon_{ij} \]  

(2)

where the $\delta_k$ indicate dummy variables for each stratum or pair $k$.

In equations (1) or (2), authors typically calculate standard errors that take account of the unit of randomization. In the case of student assignment, authors report heteroskedasticity-consistent, Huber-White standard errors (Yi et al, 2012). In cluster-randomized experiments, researchers usually report cluster-adjusted Huber-White standard errors or an alternative, such as generalized least squares with a group random effect (Glewwe, Kremer, & Moulin, 2009).
In this paper, I code estimates and standard errors from equation (2) when the experiment employs stratified or pair-wise 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 $O_{ijk}$.\textsuperscript{13} For remaining cluster-randomized experiments, I use estimates and standard errors 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’ $g$ and its standard error, using group-specific means and sample sizes, as well as the pooled standard deviation.\textsuperscript{14}

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’ $g$ in the full experimental sample using descriptive statistics that are disaggregated by the nutritional status of children (Powell et al., 1998).\textsuperscript{15} 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 and consistent. In practice, authors also report estimates based on:

\[
O_{ijk} = \beta_0 + \beta_1 T_{ijk} + X_{ijk}\gamma + \varepsilon_{ij}
\]  

\textsuperscript{(3)}

\textsuperscript{13} Some authors standardize the dependent variables by the mean and standard deviation of of the control group, but do not report sufficient data to calculate the pooled standard deviation (e.g., Barrera-Osorio & Linden, 2009). In lieu of a better alternative, I simply code the regression coefficient and its standard error.

\textsuperscript{14} This sometimes required auxiliary calculations that are contained in an Excel spreadsheet.

\textsuperscript{15} See their Table 2. In the same paper, Table 3 reports regression estimates of the treatment effect, but only with a grade-level interaction term.
where \( X_{ijk} \) is a vector of control variables that usually includes a measure of student achievement obtained at the baseline.

The main rationale of controlling for additional 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 in randomized experiments, all else equal (Dong & Maynard, 2013). On the other hand, controlling for additional 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 non-random attrition from the treatment and/or control groups.\(^{16}\)

I define an experiment as a single control group and one or more treatment arms. Thus, in an experiment with two treatment arms, two post-treatment 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 relies on 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 one study if they have overlapping samples and/or identical instructional treatments. For example, three

---

\(^{16}\) 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.
experiments on de-worming medication used different samples of children within a common set of schools (Simeon et al., 1995; Simeon, Grantham-McGregor, & Wong, 1995; Gardner et al., 1995). He, Linden, and MacLeod (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 ($\theta_{ijk}$) of various treatments. One could estimate

$$\theta_{ijk} = \theta + u_{ijk}, \quad u_{ijk} \sim N(0, \sigma^2_\theta)$$  \hspace{1cm} (4)

where $\theta$ 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 $\theta_{ijk}$, such that

$$\hat{\theta}_{ijk} = \theta_{ijk} + e_{ijk}, \quad e_{ijk} \sim N(0, v_{ijk}).$$  \hspace{1cm} (5)

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

$$\hat{\theta}_{ijk} = \theta + e_{ijk} + u_{ijk}, \quad e_{ijk} + u_{ijk} \sim N(0, v_{ijk} + \sigma^2_\theta).$$  \hspace{1cm} (6)

One can efficiently estimate $\theta$ as a weighted average of $\hat{\theta}_{ijk}$, applying inverse variance weights of $\frac{1}{v_{ijk} + \sigma^2_\theta}$ (Ringquist, 2013). $v_{ijk}$ is the square of the standard error of each effect size estimate, and $\sigma^2_\theta$ 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 $\frac{1}{v_{ijk} + \sigma^2_\theta} \times \frac{1}{n_{ijk}}$, 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. This has two benefits. First, it facilitates the calculation of Huber-White standard errors that are clustered by groups of studies, as defined previously, to allow for statistical dependencies across effect sizes. Second, it permits an extended specification

\[ \hat{\theta}_{ijk} = \theta + X_{ijk}\lambda + e_{ijk} + u_{ijk} \]  

(7)

where \( X_{ijk} \) is a vector of variables that potentially explain variation in effect sizes, referred to as moderators. I use four categories of moderators that describe: (1) treatment heterogeneity, (2) country contexts and experimental samples, (3) outcome variables, and (4) the quality of experimental design and implementation.

**Results**

**Experiments, Treatments, Follow-ups, and Outcomes**

**Experiments.** Table 1 reports descriptive data on 76 experiments, divided by three categories of treatments. First, it confirms that the use of randomized experiments has grown rapidly in the past decade (for an illustration of this growth, see Figure 2). At least two-thirds of instructional and incentive experiments are still unpublished (in part because they are so new), and most of the rest are published in economics journals. Impressionistically, economists and economics journals have embraced experimentation, while scholars in education schools and the field of comparative education have been slow to do so.\(^{17}\) The majority of health-related experiments are published in medical and nutrition journals.

---

\(^{17}\) Despite recent imbalances, the earliest experiment on instructional inputs—including mathematics textbooks and radio math lessons in Nicaragua—was conducted by an interdisciplinary team of scholars including education experts (Friend, Searle, & Suppes, 1980; Jamison et al., 1981).
Second, the smallest proportion of instructional (13%) and incentive experiments (3%) occur in Latin America, a distinction that is more stark if one notes that no experiments were conducted in Brazil. One the one hand, this may reflect the preferences and funding of scholars affiliated with the most active research centers, such as MIT’s Abdul Latif Jameel Poverty Action Lab (especially in Kenya and India) and Stanford University’s Rural Education Action Program (in China). On the other hand, there may be additional constraints to conducting randomized experiments in the relatively higher-income developing countries of Latin America.

Third, the majority of experiments begin with non-random, convenience sample of schools, often chosen because of geographic convenience, high poverty, low achievement, or because school officials consented to participate. Even Table 1 may understate the degree to which experimental samples are non-representative 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 multi-arm Indian experiment that drew a representative sample of schools in the large state of Andhra Pradesh (Muralidharan & Sundararaman, 2011). Whether convenience or random samples of schools, almost all of them include children from early primary grades 1-4, 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. This might be viewed as disconcerting, since over 90% of these same experiments are cluster-randomized. Such experiments randomly assign schools, groups of schools, or villages, and they have larger minimum detectable effect sizes than, say, a student-level randomization with the same number of children. As emphasized in the methodological literature, minimum detectable effects are
sensitive to the intraclass correlation (ICC) of the dependent variable, or the proportion of variance that lies between clusters, usually schools (Duflo et al., 2008; Dong & Maynard, 2013). In the U.S. literature, researchers have estimated ICCs for a range of outcomes and datasets to guide experimental planning (Hedges & Hedberg, 2007), but there are fewer comparable resources in developing countries (Zopluoglu, 2012).

Fifth, fewer than 10% of experiments use what might be termed quasi-random assignment, such as the alternating selection of schools from alphabetized lists. Subsequently, I assess whether use of this method, criticized by Deaton (2010), is associated with the magnitude of effect sizes. Lastly, researchers employ stratification or pair-wise matching in the majority of experiments. But, in the sample of conditional effect sizes obtained from stratified experiments, only 43% appear to have included dummy variables for strata or pairs. The result is similar to Bruhn and McKenzie (2009) who reviewed a broader sample of randomized experiments in development economics.

**Treatment arms.** Table 2 summarizes attributes of treatment arms. NGOs implement the majority of instructional and incentive treatments, while university researchers are far more likely to administer health treatments. Experimentation in collaboration with governments is rarer in all cases. Regarding the content of treatments, several patterns emerge from Table 2 and complementary data. First, treatments that use a particular instructional inputs often do so in concert with other instructional inputs. For example, 45% of instructional treatments provide materials such as textbooks. Of these, 79% also included teacher in-service training. Twenty-eight percent of instructional treatments use technology, 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 (Glewwe et al., 2009; Cristia et al., 2012).
Second, health-related treatments usually have a nutritional component or provide de-worming drugs. Malaria-related treatments are rarer, with only one included in the table’s sample of studies (Fernando et al., 2006). Health-related treatments are rarely applied in concert with either instructional inputs or incentives.\textsuperscript{18} Third, treatments with informational components and performance incentives are much less likely to also include instructional inputs, in contrast to contract and volunteer teacher treatments. Of the latter, 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 67 clusters in each treatment arm and 51 to 75 in its control group contrast. Health treatments that use student-level randomization have an average of 171 students, with 156 in the control. I will return to the issue of whether these experiments have adequate power to detect “typical” effects, as judged by the meta-analysis.

\textbf{Follow-ups and outcomes.} The average follow-up is conducted after 9-13 months of exposure to the treatment, with the research periods usually mapping onto school calendars. Despite the evident appeal of estimating medium- to long-run effects of treatments, it is rare that follow-ups occur at least one month after the conclusion of treatments (see Table 3). This is also illustrated in Figure 2, in which solid circles indicate baselines, horizontal lines indicate treatment durations, and hollow circles indicates follow-up data collection. In a few cases, follow-ups occur a year after treatment (Duflo et al., 2012). In a notable exception, Baird et al. (2012) tracked an experimental cohort of Kenyan children who received de-worming medication about 9 years after the baseline.

\textsuperscript{18} For an exception, see Sylvia et al. (2012). An ongoing Kenyan experiment combines malaria prevention and improved literacy instruction in a factorial design (Brooker et al., 2010).
For each follow-up, I attempted to code global attrition, or the proportion of the full experimental sample not present in follow-up data, and the differential attrition between a particular treatment arm and the control group. As the sample sizes in Table 3 make clear, data on global attrition are missing for 17% to 34% of follow-ups, with the smallest percentage in the sample of health treatments. Even more data is missing on differential attrition.

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 cases, 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, but conduct follow-up testing in a different cohort, perhaps with partial overlap but without student identifiers that could be used to calculate attrition (Friedman et al., 2010). Both are potential threats to internal validity, since student enrollment and drop-out—after the randomization of schools, but before follow-up testing—may create imbalances in observed or unobserved variables. 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 scatterplot in Figure 3 summarizes global and differential attrition rates, in the sample for which both are available. The median global attrition is less than 15%. The median differential attrition (treatment minus control) is negative but close to zero. In this paper, I remain agnostic about whether there are threshold levels of attrition that are “too high,” or

---

19 Missing attrition data is less likely in journals outside of economics, which may recommend or enforce the use of CONSORT guidelines, including the presentation of an experimental flow diagram (see http://www.consort-statement.org). In addition, health-related treatments are far more likely to use student-level randomization, which implies that panel data are collected on students.
whether a study has adequately ruled out attrition as a threat to internal validity.\textsuperscript{20} Instead, I specify four variables as potential moderators of study quality: global attrition, the absolute value of differential attrition between a treatment arm and the control group, and dummy variables indicating missing data for each variable.

Table 3 also describes features of the outcome assessments. I initially sought to code a wider range of outcome-related variables, such as reliability (e.g., Cronbach’s alpha), the location of test administration (school vs. home), the type of items (multiple choice vs. open response), and whether tests were oral, written, or some combination. This proved challenging because of the varied amount of details that were reported on test instruments. Of the variables in Table 3, it was not possible to identify the source of assessment items in 18-24% of instructional and incentive experiments, and only 2% of health-related experiments. As with attrition data, the difference reflects a more standardized approach to the reporting of experimental findings in health fields, often following CONSORT guidelines.\textsuperscript{21} In general, instructional and incentive treatments were most likely to use tests designed by evaluators, NGOs, or governments, while a slim majority of health treatments used off-the-shelf assessments from a firm, university, or international agency.

\textsuperscript{20} In the United States, the What Works Clearinghouse has established threshold attrition standards, based on a model and simulations (What Works Clearinghouse, 2011). Attrition is not a threat to internal validity if it occurs completely at random. That seems unlikely in poor schools, where high student absence and drop-out rates, poverty, and test score outcomes are plausibly correlated. Still, the main concern is that attrition does not introduce correlations between treatment group membership and the observed or unobserved variables that determine outcomes. Many of the coded papers test whether attrition creates imbalance in observed variables, statistically control for observed variables in linear regressions like equation (3), and apply robustness checks such as bounding (see the discussion in Duflo et al., 2008).

\textsuperscript{21} For example, the CONSORT checklist requires authors to “completely [define] pre-specified primary and secondary outcome measures, including how and when they were assessed” (http://www.consort-statement.org).
Effect Sizes by Treatment

**Instructional inputs.** Figures 4 to 8 summarize effect sizes for each category of instructional inputs. Diamonds indicate the effect size, bracketed by a 95% confidence interval. The relative size of diamonds, within figures, is proportional to their weight in the category mean. (Recall that weights may be lower because effect sizes are less precisely estimated, or because multiple effect sizes are reported for a single treatment arm.) Vertical lines are placed at zero and at the mean effect size. On the left-hand side, the label indicates the numerical code of each experiment (referenced in the Appendix), the country, the type of outcome (whether language, mathematics, or a composite), the number of the follow-up (from 1 to 3), and a thumbnail description.

Each figure reports the category-specific mean, estimated with equation (6), and $p$-value for the null hypothesis that the mean is equal to zero. It is estimated with the wild cluster bootstrap-$t$, suitable for the small number of study clusters in most figures (Cameron, Gelbach, & Miller, 2008; Ringquist, 2013). The mean effect sizes are positive and statistically different from zero in all categories except school grants. The means range from 0.08 (materials) to 0.15 (computers and technology).²² Within each category of intervention, there is substantial variation in effect sizes, although it bears emphasis that only two estimates across all treatments are negative and statistically significant. In the two estimates, the treatments unexpectedly led to reallocations of school resources away from learning activities for some students.²³

²²An excluded Liberian experiment provided teachers with training, school-based coaching, and scripted lesson plans in literacy instruction (Piper & Korda, 2011). It yielded large effect sizes on a reading assessment, though its standard errors may have been under-stated, since it did not cluster by the unit of randomization.

²³In one case, in-school computer-assisted instruction appeared to substitute away from other types of instruction (He et al., 2008). In another case, schools received a grant to reduce the incidence of anemia, accompanied by health information and a performance incentive for principals. There was no negative effect is a treatment arm without the performance incentive,
The categories of each figure belie the heterogeneity of interventions. Instructional materials, computers, training, and small-group instruction are often provided in concert with each other, and with other treatments. Upon visual inspection, the clearest pattern is that instructional materials alone do not improve learning (Kremer et al., 2003, 2004; Glewwe et al., 2009), although complex treatments also appear at the bottom of Figures 4 and 7, such as an NGO-led treatment that provided different mixes of training, materials, and volunteers to public schools in India (Banerjee et al., 2012). Similarly, Figure 5 suggests that computers have smaller effects on learning when they are given to students with no accompanying instructional strategy or monitoring (Cristia et al., 2008) or when they substitute away from useful instructional time during school hours (He et al., 2008).

**Health inputs.** The mean effect size of providing nutritional interventions is 0.04 and statistically different from zero at 10% (Figure 9), but there is no evidence that de-worming drugs affect achievement, on average (Figure 10). Two caveats are warranted. First, many excluded studies were in the first category, although the available data from these studies shows statistically insignificant effects for micronutrients and zero to small effects for school meals. Similarly, one excluded study reported no statistically significant effects of a de-worming treatment on school exams (Nga et al., 2011).

---

24 Pollitt et al. (1989) and Sunghthong 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.

25 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 (Jacoby et al., 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 statistically insignificant (milk supplement).
Second, in-kind transfers like school feeding programs have often been found to increase short-run measures of school participation such as enrollment and attendance that are not included in this paper (Petrosino et al., 2012; McEwan, 2013). An oft-cited Kenyan experiment—included in Figure 10—finds short- and longer-run impacts of de-worming treatments on school participation and attainment (Miguel & Kremer, 2004; Baird et al., 2012), despite the lack of test score effects. Viewed together, the results suggest that attending school is a necessary but not sufficient condition for improving learning.\footnote{In a similar vein, conditional cash transfers program have been shown to consistently improve school enrollment and attainment (see Galiani & McEwan, 2013 and the citations therein). In contrast to the literature on school meals, however, there are far fewer evaluations that measure the cognitive or learning outcomes of children in the short-run or long-run, and the effects are mixed (Behrman, Parker, & Todd, 2009; Barham, Macours, & Maluccio, 2012).} But, further complicating interpretations, Baird et al. (2012) find that de-worming increases wages, perhaps suggesting that we are not measuring the correct learning outcomes (or, more dismally, that attainment is associated with long-run success in labor markets, but not learning).

**Incentives.** Figures 11 to 14 summarize mean effect sizes for the categories of information (0.05), performance incentives (0.10), contract or volunteer teachers (0.10), and school management or supervision (0.06). The first and last of these are not statistically different from zero. Performance incentives and informational treatments have the least overlap with other categories of interventions, permitting an easier comparison between “consequential” accountability and softer policies. School report cards are sometimes effective, at least when designed to be useful to parents (Andrabi et al., 2009; Barr et al., 2012), but many effects in Figure 11 are small and imprecisely estimated. Information for students on the economic returns to schooling improves achievement in an oft-cited Magagasc 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. Recent experiments involving student information in China appear disproportionately near the bottom.

The sparse evidence on teacher performance incentives is more encouraging, particularly in India, but a Kenyan experiment found that effects were focused on the tests used in performance formulas (Glewwe et al., 2010; Muralidharan & Sundararaman, 2011). Student performance incentives are mixed, with positive effects in a well-known Kenyan experiment (Kremer et al., 2009), but no evidence that cash incentives raise students’ achievement in Chinese classrooms, unless combined with peer tutoring (Li et al., 2010).

Treatments with contract and volunteer teachers are heterogeneous and overlapping. One might conclude that volunteers alone are insufficient, even when combined with additional inputs (Cabezas et al., 2011; Banerjee et al., 2012). The effective use of contract teachers is often accompanied by smaller class sizes (Muralidharan & Sundararaman, 2010a; Bold et al., 2012; Duflo et al., 2012). Using an additional treatment arm, Duflo et al. (2012) determined that class size reduction has small and imprecisely estimated effects when implemented with civil-service teachers (noting its presence at the bottom of Figure 8). 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 treatments in Figure 14 cautions against simple conclusions about the effects of “school-based management.” Duflo 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 Magagascar 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**

I next analyze moderators of effect sizes in a multivariate framework. The regressions in Table 4 control for treatment heterogeneity in addition to variables drawn from Tables 1 to 3 that describe country contexts and experimental samples, outcome variables, and the quality of the experimental design and implementation. The sample includes 256 effect sizes in 75 experiments, grouped in 58 study clusters. Given the constraints of sample size, particularly the number of study clusters, I pool effect sizes across all categories of interventions.

Column (1) controls for a series of dummy variables indicating treatment categories, while additional columns include controls for separate categories of 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. Technology and performance incentives have consistently positive effects, which are larger and significant only when controlling for country context and sample (notably the GDP per capita of countries). Nutritional treatments present a special case. The coefficient in column (6) is statistically insignificant, but its positive sign (0.04) and large standard error (0.05) do not warrant strong conclusions. Finally, materials, grants, deworming, information, contract and volunteer teachers, and management and supervision have anywhere from a negative to statistically insignificant effects.

---

27 The 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 (Fernando et al., 2006).
Why are some categories less effective after controlling for moderators? A plausible explanation is that their effects are largely explained by effective (and overlapping) components such as teacher training, computer-assisted instruction, and class size reduction and small-group instruction. A related hypothesis is that effects are not attributable 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 that training or materials alone would ineffective. 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 the combination thereof). In a meta-regression, one might include interaction terms between treatment components, but the modest sample sizes in Table 4 do not allow convincing tests.

Among remaining moderators, there are surprisingly few consistent correlates of effect sizes. Two are country-related, including the Latin America region (relative to Africa) and 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 capturing household incomes of school children, the quality of schooling in “business-as-usual” 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

---

28 I am only aware of three fully factorial designs that evaluate instructional and/or incentive interventions (He et al., 2008; Nguyen, 2008; Brooker et al., 2010), noting that final results from Brooker et al. (2010) are not available. Many experiments in the sample are multi-arm experiments with mutually exclusive treatments (e.g., Muralidharan & Sundararaman, 2011) or incomplete factorial designs (e.g., Dupas et al., 2012; Pradhan et al., 2011).

29 In specifications with added interactions, the standard errors are larger and coefficient estimates fluctuate substantially depending on the specification, suggesting problems of “micronumerosity” and multicollinearity.
random experimental samples reflect careful research planning that is also likely to be correlated with the choice of 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 are helpful in testing two common 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. Table 4 provides no evidence of publication bias, given the small and insignificant coefficient on a variable indicating published papers, versus working papers or reports. A lingering concern is that experiments with zero or negative effects, especially imprecise ones, are simply placed in the “file drawer” and never circulated as working papers. The funnel plot in Figure 15 graphs standard errors against effect sizes for the sample used in Table 4. It is roughly symmetrical, with more precise studies (with smaller standard errors) more tightly clustered around the average. There is no tell-tale “notch” removed from the lower-left side, which might imply that negative effects, especially imprecise ones, are less likely to be circulated.

**Stratification and Baseline Controls**

Most experiments randomize within groups defined by location, gender, poverty, test scores, and other variables (Table 1). The aim is to ensure that treatments and control groups are balanced along stratifying variables, thus increasing precision of estimated effect sizes (Duflo et al., 2008; Bruhn & McKenzie, 2009). As noted above, it is considerably less common that researchers explicitly control for strata dummy variables, even though doing so may improve
precision (Bruhn and McKenzie, 2009). In addition to stratification, it is common to control for baseline variables such as a pre-test.

Table 5 presents exploratory regressions that explore whether these methodological choices influence the precision of effect size estimates. The sample in column (1) includes a stacked sample of 348 standard errors that correspond to all unconditional and conditional effect sizes. Based on prior terminology, an unconditional effect reflects, at most, controls for strata dummy variables, while a conditional effect reflects controls for strata dummies and other baseline variables. Column (1) shows that the use of stratified randomization—in the absence of strata controls—lowers the standard error by 0.006, on average, though the coefficient is not statistically significant. The reduction is larger when strata controls are included (0.032) and/or when other baseline controls are included (0.027). Either reduction is at least one-quarter of a “typical” effect size of 0.1. The results suggest that researchers risk overly conservative inference by failing to control for strata dummies (Bruhn & McKenzie, 2009). Column (2) adds fixed effects for each pair of unconditional and conditional estimates reported for a particular experimental outcome. It also suggests that standard errors are substantially reduced by including baseline controls (for the other variables, there is no within-pair variation).

Columns (3) and (4) substitute effect sizes as the dependent variable. A cynic might predict a positive coefficient on the variable indicating baseline controls, perhaps implying that researchers strategically choose baseline controls so as to increase effect sizes. But, the results are consistent with the notion that neither strata nor baseline controls are associated with consistently larger or smaller effects. Finally, columns (5) and (6) repeat the exercise with the smaller sample of effects for which an $R^2$ is reported. The inclusion of strata controls and other baseline controls are, not surprisingly, associated with a higher proportion of explained variance.
in student outcomes. Baseline controls increase the R² by 0.17 to 0.2, on average. The literature on power analysis for cluster-randomized experiments emphasizes that including controls such as school-level or student pre-test can increase the power of designs to detect effects of a given size (Bloom, Richburg-Hayes, & Black, 2007; Dong & Maynard, 2013).

Discussion

Policy Lessons from Experiments

Effects on learning. The treatments can be broadly divided into four groups, discussed in ascending order of their average effects on learning. In the first group, school grants and deworming treatments do not affect test scores, on average, and these disappointing effects are robust to controls for moderators. Across a diverse group of studies, there is only one instance of a statistically significant effect of grants on learning (Das et al., 2011).

The second group include nutritional, information, and management and supervision treatments. On average, nutritional interventions have small effects on learning that are of a similar magnitude after controlling for moderators (albeit statistically insignificant). Figure 9 suggests that micronutrient interventions have the greatest potential to increase learning, at least in Indonesia (Soemantri et al., 1985; Soemantri, 1989) and China (Kleiman-Weiner et al., 2013; Luo et al., 2012). Kleiman-Weiner et al. (2013) is particularly relevant, because it finds no effects of a food-based intervention in a second treatment arm. Informational treatments also have relatively smaller effects that are not robust to controls for moderators. The most promising results suggest that the availability of “school report cards” can affect learning in the poor contexts of Pakistan (Andrabi et al., 2009) and Uganda (Barr et al., 2012), especially when they are designed to fit the needs of parents and communities. Despite encouraging results on the
provision of information to students about economics returns to schooling in one study (Nguyen, 2008), the results are not robust across other treatment arms or emerging studies from China (Figure 11). Finally, the average effects of management and supervision treatments are small, and not robust to moderators. In some cases, school-based management is accompanied by class size reduction and contract teachers (Duflo et al., 2012). Across multiple treatment groups, an Indonesian experiment found that “linkages” between school committees and villages improves language but not mathematics scores (Pradhan et al., 2011).

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, computer and technology, class size reduction, or other interventions. In Figure 4, the simple provision of materials is often ineffective (Glewwe et al., 2004, 2009), but is more effective when combined with training and a well-defined instructional model (Friedman et al., 2010; Lucas et al., 2013). 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 (Muralidharan & Sundararaman, 2010a; Bold et al., 2012; Duflo et al., 2012) and other interventions conducted in small-group settings (Banerjee et al., 2007). Though Duflo et al. (2012) show that class size reduction by itself is minimally effective, 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.

Computer-assisted instruction is also uniformly effective in raising student learning, at least when combined with a systematic plan to encourage the use of tutoring software that is aligned with curricular objectives. The experiments are notable for their geographic reach, with encouraging results in China (Lai et al., 2012; Mo et al., 2013), Ecuador (Carillo et al., 2010), and India (Banerjee et al., 2007; He et al., 2008; Linden, 2008). On the other hand, computer-assisted instruction is less effective when it does not appear to emphasize tutoring (Barrera-Osorio & Linden, 2009), when it simply distributes computers to students without additional parent or student training (Cristia et al., 2012), or when it supplants valuable instruction time in classrooms (Linden, 2008).

Treatments that reduce class sizes are often effective, although many such effect sizes in Figure 8 are contributed by a single study that also finds that class size without complementary treatments is not effective (Duflo et al., 2012). An overlapping experiment on ability tracking found that separating students by ability has the counter-intuitive result of increasing learning for all students, perhaps because teachers can better match instruction to students’ levels. 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.

Lastly, there are three experiments in which teacher performance incentives have been shown to increase student learning when provided to individuals or groups (Figure 12), although a
Kenyan one sounds a cautionary note about its potential effects on strategic behavior by teachers (Glewwe et al., 2010). Despite encouraging results in a Kenyan experiment that provided student performance incentives (Kremer et al., 2009), recent experiments in China have less consistent findings (Figure 12). 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.** Although this review has focused entirely on learning outcomes, several categories of interventions have been shown to increase enrollment, attendance, and attainment. These include conditional cash transfers (Fiszbein & Schady, 2009; Galiani & McEwan, 2013), in-kind transfers such as school feeding programs (Vermeersch & Kremer, 2004; Petrosino et al., 2013), de-worming (Miguel & Kremer, 2004), and the construction of new school infrastructure where there are supply constraints (Burde & Linden, 2013; Kazianga et al., 2013).

Categories with more experimental research on learning, such as school feeding programs and de-worming, show little robust evidence of learning effects. The evidence is sparse and mixed for conditional cash transfers (Behrman et al., 2009; Barham et al., 2012). There is encouraging evidence that school construction can increase enrollment as well as test scores, although the counterfactual in each case includes many children who are exposed to no schooling whatsoever (in contrast to most school meal or de-worming experiments).

---

30 For a meta-analysis of these outcomes, see Petrosino et al. (2012). For recent narrative reviews, see Kremer and Holla (2009) and Kremer et al. (2013)
The evidence suggests that providing incentives, reducing constraints, and improving health are useful tools for increasing schooling access, but the mixed effects on learning suggest that children’s time is not always used productively once they are attending. These gaps in our knowledge could be filled by future experiments that combine access-based interventions with supply-side interventions in the instructional quality of schools. In Honduras, for example, one of the earliest studies of conditional cash transfers developed a factorial design to separately evaluate block grants to schools, cash transfers, and the combination of the two (Galiani & McEwan, 2013). Unfortunately, the block grant intervention was minimally implemented.

**Costs and cost-effectiveness.** It is important to consider the costs of treatments alongside their effects (Levin & McEwan, 2001; McEwan, 2012). The overarching concern is that similarly effective treatments may vary widely in their costs, or that treatments with smaller effects may nonetheless have relatively lower costs. In either case, it is potentially inefficient to use effect size as the sole criterion for ranking of investments.

To facilitate comparisons of costs across studies, 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 key resources. Most did not report the exchange rates used to convert estimates to a common currency (usually 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 (Banerjee et al., 2007; Nguyen, 2008; Glewwe et al., 2009; Kremer et al., 2009;
Glewwe et al., 2010; Abeberese et al., 2012; Duflo, Dupas, & Kremer, 2011, 2012; Pradhan et al., 2011; Duflo, Hanna, & Ryan, 2012). 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. 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.

Figure 16 illustrates the effect sizes in these treatment arms, as well as 95% confidence intervals (the cost per student is reported in each row). The figure highlights the importance of cost-effectiveness comparisons: a few relatively effective interventions, such as computer-assisted instruction, have among the higher costs. Figure 17 next illustrates the cost per student of increasing the effect size by 0.2. The 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 cost-effectiveness ratios are below $100, but a few are far higher.

The most cost-effective alternatives in Figure 17 are the provision of information to students about economic returns in Madagascar (Ngyuen, 2008), “linkages” between school committees and village leadership in Indonesia (Pradhan et al., 2012), 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, & Kremer, 2012).

---

31 The disaggregated cost data are reported in a spreadsheet available at www.povertyactionlab.org/doc/cea-data-full-workbook (downloaded July 15, 2013).
32 This paper’s cost estimates are different from Kremer et al. (2013) in three regards. 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 re-calculate costs in a few treatment arms so that they exactly correspond to the treatment-control contrast recorded in the meta-analytic sample. For the updated spreadsheet, see www.patrickmcewan.net/meta.
Despite these suggestive findings, the interpretation and use of cost-effectiveness ratios faces challenges. First, it is not obvious how to rank cost-effectiveness in the presence of multiple outcomes. For example, two experiments reported lower effects in mathematics, not least because they focused on the acquisition of reading skills (Abeberese et al., 2012; Banerjee et al., 2007). Other experiments may report stronger effects in both, or may simply omit one from the evaluation. It seems clear that available learning outcomes should not be omitted from a cost-effectiveness analysis, raising the possibility of single, weighted average. The weights connote the “utility” derived by decision-makers, although the methods for deriving those weights are sometimes ad hoc (Levin & McEwan, 2001).

The challenge of multiple outcomes is more acute in the case of longer-run (and usually unobserved) outcomes such as attainment and earnings. Ultimately, we should aim to identify interventions with the highest ratio of lifetime, monetary benefits to costs. Childhood learning gains plausibly affect benefits through multiple channels, including direct or indirect effects on labor market earnings. Many school-based experiments are too recent for long-run follow-ups, despite encouraging findings in at least one case (Baird et al., 2012). In the absence of such evidence, the options are to (1) focus on identifying and measuring learning and other short-run outcomes that validly predict lifetime outcomes such as earnings, and (2) use short-run findings to impute long-run benefits, aided by assumptions about the effects of childhood test-score gains on adult earnings (McEwan, 2012). Both options suggest a greater need for empirical studies that assess the causal relationship between childhood and adult outcomes in developing countries.

Second, it is tempting to allocate resources in ascending order of the cost-effectiveness ratios (CERs). But, as Figure 16 made clear, effect sizes that are statistically different from zero often cannot be statistically distinguished from the magnitude of relatively smaller or larger effects.
Figure 17 did not estimate confidence intervals for CERs, although best-case and worst-case CERs can be easily constructed using the right-hand and left-hand sides, respectively, of the effect size interval. Those results emphasize that uncertainty about effect sizes can seriously affect our ability to precisely rank the cost-effectiveness of treatments. A constructive response would be to incorporate cost-effectiveness considerations in the study design and data collection. First, it is clear that cost estimates are themselves estimates of an uncertain parameter, implying that confidence intervals should be constructed when feasible. Second, power analyses used to guide sample size should analyze costs as well as learning outcomes. Third, future work should endeavor to estimate the standard errors and confidence intervals of cost-effectiveness ratios, and incorporate uncertainty into the ranking of investment options.

Third, the interpretation of Figure 17 warrants caution because it is based on a subset of the meta-analytic sample. Ideally, one might develop valid estimates of comparable costs and cost-effectiveness ratios in all impact evaluations. This would facilitate broader analyses that compare cost-effectiveness across broader classes of interventions. Meta-regressions could further be used to examine correlates of CERs, such as features of samples and contexts that are potential moderators of cost-effectiveness.

**Methodological Lessons from Experiments**

**Sample size and power analysis.** Smaller experiments are less costly to conduct, but they also yield less precise estimates of treatment effects (illustrated in Figures 4 to 14 by wider confidence intervals around effect size estimates). This is particularly worrisome when confidence intervals are so wide that researchers cannot confidently conclude that an effect size—even a relatively larger one—is statistically different from zero. The meta-analytic sample
provides clues as to whether researchers are making judicious tradeoffs between experimental cost and statistical power.

The average cluster-randomized experiment of an instructional treatment includes approximately 100 schools, divided evenly between treatment and control groups (Table 2). A common approach to conducting a power analysis is to calculate a minimum detectable effect size (MDES) under a set of researcher-supplied assumptions about significance and power levels, sample sizes, and the intraclass correlation (Schochet, 2005; Dong & Maynard, 2013).\(^{33}\) In a cluster-randomized experiment that evenly allocates clusters across groups, the MDES is smaller when the number of clusters is larger, or when the intracluster correlation (ICC) is smaller. It is typically less sensitive to the number of students within each cluster (Schochet, 2005). The ICC measures the proportion of variance in the outcome that lies between clusters. It may vary by outcome and, certainly, by country. In U.S. evaluations that randomize schools, the recommended ICC’s usually fall between 0.15 to 0.2 for test score outcomes (Schochet, 2005; Hedges & Hedberg, 2007; What Works Clearinghouse, 2011). Using the 2011 PIRLS assessment of fourth grade reading achievement, I calculated larger ICC’s of 0.33 to 0.44 in six developing countries.\(^{34}\) Zopluoglu (2012) estimates ICCs using all countries that participated in the 4th grade PIRLS assessment, and the 8th grade TIMSS mathematics assessment, finding many estimates that exceed the standard U.S. assumptions. For example, Chile’s mathematics ICC is 0.51, perhaps not surprising given the implementation of full school choice since the early 1980s.

\(^{33}\) It is common to assume a significance level of 0.05, and a power of 0.8. The former is the probability of committing a Type I error (that is, rejecting a null hypothesis of no effect when it is true). Power is \(1 - \beta\), where \(\beta\) is the probability of committing a Type II error (that is, failing to reject a null hypothesis of no effect when it is false).

\(^{34}\) The countries included Botswana (0.36), Colombia (0.44), Honduras (0.35), Indonesia (0.41), Morocco (0.42), and Trinidad and Tobago (0.33).
The MDES of a cluster-randomized experiment is 0.26, assuming 100 schools, 50 students per school, and an ICC of 0.2. Quadrupling the number of students per cluster changes the MDES by less than 0.01, implying few benefits that might justify the costs of gathering data on all treated students in a particular cluster.\textsuperscript{35} Assuming a larger ICC of 0.4 increases the MDES to 0.37. Only 12\% and 7\%, respectively, of this study’s estimates exceed those values.

To decrease the MDES, one can increase the number of clusters. With 200 schools and an ICC of 0.2, the MDES falls to 0.19. One could also stratify the experimental sample prior to randomization, and gather baseline data (both, ideally, using variables that are highly correlated with test score outcomes, such as pre-test). Based on evidence from Table 5, I assume that controlling for strata indicators and other baseline variables explains 35\% of the variation in the test score outcome. This further reduces the MDES to 0.15 (or 0.21 under the larger ICC).

The results highlight the substantial challenges facing budget-constrained researchers who are evaluating interventions that might have relatively modest effects on student outcomes. They imply several guidelines for future experiments. 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 pre-test, poverty, or other plausible determinants of final test scores, and control for strata in their analyses. If variables are numerous, then researchers can apply pair-wise matching and use all of them (Bruhn & McKenzie, 2009). Third, researchers should collect baseline data on students and schools, ideally a pre-test, and report estimates that control for the pre-test. Even when student-level data cannot be collected, a cluster-level pre-test can increase power in cluster-randomized experiments (Bloom et al., 2007). Fourth, it is clear that very small numbers

\textsuperscript{35} I use the spreadsheet developed by Dong and Maynard (2013), which implements standard formulas described by Schochet (2005) and others.
of schools—say, experiments with less than 50 per treatment arm—are not well-suited to identifying the relatively small effects on learning observed in many categories of education and health treatments.

Similar recommendations can be applied to the student-level experiments (though ICCs are not relevant in this case). As Table 1 showed, student-level randomization is far more common in evaluations of individually-administered health treatments, such as medications and nutritional supplements, although it has been usefully applied to a technology treatment (Mo et al., 2012). Student-level randomization can be a cost-effective way to conduct an experiment, since data collection and treatments can occur in many fewer schools. Despite this benefit, student-level randomization has three potential drawbacks in a school setting. 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 raises 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 for the use of student randomization. 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.

**Experiments and external validity.** One of the most common critiques of randomized experiments is that their results have limited generalizability, at least beyond the small-scale and perhaps highly-controlled settings in which they are conducted. Our meta-analytic results suggest

---

36 The MDES of a student-level randomized experiment with 300 students, even split across treatment and control groups, is 0.29. Doubling the number of students, and assuming that baseline variables explain 35% of variation in the outcome, lowers it to 0.16. This assumes that school effects are treated as fixed, rather than random (Schochet, 2005).
that at least some of these concerns are exaggerated. Meta-regressions showed that several types of treatments were consistently associated with larger effect sizes, on average, even after controlling for moderators related to the treatment, context, and outcome measures. Yet, these results must be cautiously interpreted. The meta-analytic sample is still small and has insufficient data on some moderators, such as outcomes. On others, it lacks sufficient variation in moderators of particular interest, such as the implementing agency and the scale of the intervention.

In the meantime, how can experiments provide more generalizable knowledge to policy-makers? First, experiments should directly test hypotheses about external validity by randomly assigning schools and students to variations of the treatment. For example, a common concern is that interventions will be less successful when implemented partially or entirely by a government agency rather than an NGO. There are several instances of successful partnerships between public schools and NGOs, in which NGO-led training and monitoring led to increased learning in public school classrooms (e.g., Lucas et al., 2013; Friedman et al., 2010). But, an Indian experiment found a severe breakdown in the quality of program implementation, when responsibility for monitoring was partially handed over to public education officials (Banejee et al., 2012). Bold et al. (2012) considered these issues in the context of a single experiment, randomly assigning schools to a contract teacher intervention run by an NGO or to a similar program run by the government, finding weaker effects in the latter. If it is not feasible to use multiple treatment groups in a single experiment, it may be possible to conduct parallel experiments that implicitly hold constant most aspects of the treatment, context, and outcomes, but selectively vary others (e.g., He et al., 2008; Linden, 2008).

Second, experiments should use samples that are representative of well-defined populations of schools and students (despite the logistical challenges that this might entail). They should also
test whether effect sizes vary by pre-defined attributes of schools and students. An evaluation of teacher performance pay used a random sample of schools in a very large Indian state (Muralidharan and Sundararaman, 2011). This facilitated the estimation of average treatment effects that could be generalized to the “typical” primary student in the state—a question of considerable policy interest. Further subgroup analysis suggested that treatment effects were similar among multiple subgroups, allaying concerns that effects were driven partially or entirely by a “special” group of students or schools.

There are drawbacks to subgroup analysis. One is that it may limit statistical power by relying on ever-smaller samples of schools or students, requiring careful planning by researchers. Relatedly, Deaton (2010) cautions against in which evaluators “hunt” for statistically significant effects in arbitrarily chosen groups of data, and then construct ex post theories to “explain” these effects. To guard against this likelihood, it would be desirable to define subgroups in the planning phase, effectively tying researchers’ hands. A related option is to define subgroups using experimental strata, perhaps defined by poverty, pre-tests, or gender (Galiani & McEwan, 2013).

Third, experiments should measure and report a wider set of outcomes. There are three arguments in favor of doing so. One is that treatments with a tight focus on language or reading may cause instructional time to be diverted from mathematics or other subjects, and vice-versa. It is also plausible that treatments unexpectedly improve a wider set of outcomes, perhaps by developing teacher capacity more broadly. In either case, it argues in favor of over-measuring rather than under-measuring student competencies. But, in the spirit of Deaton (2010), it implies that researchers should comprehensively report treatment effects for all outcomes, and avoid “cherry-picking” the statistically significant results.
A second issue is that treatments themselves might lead to strategic behavior that increases one outcome, but not another. Teacher performance pay in Kenyan produced gains in the incentivized test score, but not a lower-stakes exam, which the authors attributed to test coaching behavior rather than real instructional improvements (Glewwe et al., 2010). The implication, especially in performance pay experiments, is that a broad set of outcomes should be measured.

A third issue is that assessments—especially those designed by evaluators themselves—mainly emphasize competencies that are the explicit focus of a short-term treatment (e.g., simple math operations or letter-sound relationships, rather than a broader ability to analyze complex mathematical problems or comprehend text across multiple subjects). There is clearly an interest in assessing whether a tightly-focused treatment succeeds in delivering students a narrow set of important skills, especially in early primary grades. But, most school systems promote more ambitious curricular objectives. Basic decoding and math skills are usually a necessary but not sufficient condition for achieving these objectives, which are often measured in grade-specific government assessments that are aligned with the curriculum. To be fair, the meta-regression found no evidence that assessment type was clearly associated with the size of effects, but this would be far easier to determine if experiments were to include a broader set of outcome measures.

Fourth, experiments can provide more data to assess the causal mechanisms that undergird “black-box” causal effects. Policy effects encompass multiple direct and indirect effects that, alone or in concert, could explain effects on learning. Researchers could experimentally manipulate one of those of mechanisms, but the constraints of time and money lead to compromises. In the best cases, authors explicitly lay out the hypothesized mechanisms in a theory of change and use it to guide data collection during the evaluation for testing secondary
hypothetical hypotheses about mechanisms (Brooker et al., 2010). Then, they report descriptive analyses of inputs, processes, and intermediate outcomes to 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 policy-makers 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).

Fifth, researchers should complement experiments with thoughtfully conceived quasi-experimental research that evaluates scaled-up, policy-relevant treatments in representative samples of schools and students. The regression-discontinuity design (RDD) is one of the most promising research designs, since it can deliver effect sizes with internal validity that is comparable to randomized experiments, depending on features of the design (McEwan & Shapiro, 2008; Lee & Lemieux, 2010). In a typical education RDD, treatments are assigned to schools or students on the basis of a single assignment variable (e.g., poverty or test scores) and an assignment cutoff. For example, Chile’s Ministry of Education assigned a remedial after-school tutoring program to public schools with mean achievement that fell below a cutoff, and trained aides provided tutoring to lower-achieving students within the school (Chay et al., 2005). The intervention was implemented on a large scale by a government agency, and evaluated with a national exam given to the population of 4th graders. The treatment increased test scores in language and mathematics, although these effects are most generalizable to students attending schools near the cutoff (which are, potentially, the “best of the worst” given the assignment rule). This highlights a common critique of the RDD: that it lacks external validity because effects are not representative of all participating students.
But, consider a related experiment in the meta-analytic sample. Cabezas et al. (2011) evaluated a more recent tutoring program in a smaller convenience sample of Chilean schools, chosen in part because the schools had at least 90 students per grade. It found no consistent effects in the full sample. One explanation for the difference with the RDD is that the tutoring program was less effective among relatively larger (and more urban) primary schools. Another explanation is that the treatment was different. Run by an NGO, it used volunteers rather than trained aides, targeted all fourth-graders, and occurred during school hours rather than after school. In this context, it is clear that neither evaluation has the sole claim to external validity, and that both evaluations should contribute to policy conversations in Chile.

**Standardized reporting of experiments.** Experimental reports vary widely in their content and detail, although there is less apparent variation when researchers adhere to a reporting standard such as the CONSORT guidelines referenced earlier. Based on these guidelines, and my own observations during coding, there are four low-cost methods of enhancing the clarity and credibility of experiments.

First, researchers should use a flow diagram to describe the size of the experimental sample—including schools and students—from recruitment to randomization to baseline data collection to follow-up(s). The primary goal is to provide details on whether the original experimental sample, as randomized, is intact by the time follow-ups are conducted, since non-random attrition is a fundamental threat to internal validity. In the special case of cluster-randomized experiments, this implies that authors should identify the baseline sample of students enrolled in schools, before general knowledge of treatment conditions can affect the sorting of students across schools. Non-random sorting—which could include drop-outs, re-enrollments, or even defections from untreated to treated schools—has the potential to create imbalance across
treatment and control groups in student attributes. In lieu of this basic information, the causal claims of randomized experiments must be regarded cautiously. Accompanying the flow diagram, researchers should (at least) report the means, standard deviations, and sample sizes of all dependent and independent variables, separately by treatment and control groups.

Second, researchers should report both unconditional and conditional effects, as defined earlier in this paper. When stratified random assignment or pair-wise matching is used, both estimates should control for strata or pair dummy variables, in order to maximize precision. Baseline controls in the conditional effects should be limited to pre-specified covariates, such as a pre-test or gender, in order to minimize the likelihood of specification searching for statistically significant effects. The standard errors of the estimates should correctly account for the unit of randomization (most commonly, and agnostically, by reporting Huber-White standard errors clustered at the unit of randomization).

Third, researchers should complement the flow diagram’s analysis of attrition by reporting descriptive statistics—using baseline variables—within four groups defined by treatment status and presence in the follow-up data. The purpose is to demonstrate that attrition is plausibly random (if attritors are observably similar to non-attritors in both groups) or, barring that, that non-random attrition does not create imbalance in observed variables across remaining members of treatment and control groups. It is possible that high or imbalanced attrition increases the risk of selection bias, recalling Paul Holland’s comment that “a randomized controlled experiment is just a quasi-experiment waiting to happen” (cited in Briggs, 2008). In these cases, especially, researchers should present additional robustness checks such as bounding (Duflo et al., 2008).

Fourth, researchers should provide a complete description of the treatment and outcome measures. This point may seem obvious, but it is common that experimental reports conflate—
most likely unintentionally—a general label with a more complex treatment or outcome construct. What, specifically, were the educational materials provided to children? Who conducted teacher training, and did it follow a pre-established curriculum or instructional method that is described elsewhere? During the intervention itself, did someone monitor or coach the teachers? At the follow-up, what “math” skills were assessed and why? Is there any evidence that the math test is associated with more distant outcomes of greater interest, such as labor market success? Without belaboring the obvious, vague treatment and outcome descriptions hamper the ability of policy-makers to interpret or replicate the results (which is, after all, the nominal purpose of experimentation).

Fifth, researchers should report detailed estimates of the incremental cost of the treatment relative to the control group, at least in an appendix (McEwan, 2012). Without this evidence, the utility of the experimental findings for policy decisions is sharply diminished, and it forces reviewers to cobble together cost estimates from documents and the fading memories of implementers.

**Looking Forward**

The collected evidence suggests several priorities for future experimentation. First, as Figure 2 made clear, there is now a large pipeline of experiments that measured test score impacts over a single school year, and far fewer experiments that conduct longer-run follow-ups. One of the latter suggested that students treated with de-worming medication had improved attainment and labor market success, despite smaller learning gains (Kremer & Miguel, 2004). The results are reminiscent of long-run experiments in the U.S., such as the Tennesee STAR class size experiment. It found that early test score gains among treated students faded out by later grades,
but re-emerged in the form of higher levels of participation in tertiary education (Schanzenbach, 2007). The existence and explanation of these dynamics should be analyzed in developing countries, particularly in large and well-documented experiments with novel treatments, substantial effect sizes, and modest attrition rates (e.g., Duflo et al., 2012; Muralidharan & Sundararaman, 2011).

Second, treatments that include instructional interventions have become increasingly complex, evolving from basic textbook distribution to well-integrated instructional approaches with complementary inputs. The experimental literature should continue to pursue novel instructional treatments, ideally with fully factorial designs that provide greater insight into the causal mechanisms of layered treatments.

Third, experiments should contemplate factorial designs that combine instructional interventions with health and incentive treatments. To date, health experiments have not included instructional treatments, and vice-versa, although there are compelling reasons to do so. Even when health and nutrition treatments do not lead to large gains in test scores—as with deworming and school meals—other evidence suggests that they improve attainment. Their effects on learning could be magnified to the extent that children’s time in school is used more productively. In a similar vein, conditional cash transfer experiments consistently find large effects on enrollment and attainment, but evidence on test score effects is sparse and mixed. These demand-side treatments could also be combined with supply-side investments in school quality.

The literature has also shown promising results from incentive-based treatments. When combined with instructional interventions, as in contract teacher interventions, it is difficult to separately identify the contributions of instructional inputs (like class size reduction), existing
teacher capacity, and modified teacher incentives. There is still no evaluation that combines teacher performance pay with a novel instructional intervention, despite mounting evidence that teacher capacity and incentives both matter. As above, a factorial design would allow researchers to identify the separate impact of each, and the interaction between the two.
References

References with an code in brackets were included in the meta-analytic sample.


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. [27-28]

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


Table 1: Characteristics of experiments

<table>
<thead>
<tr>
<th></th>
<th>Instructional inputs (N=39)</th>
<th>Health inputs (N=22)</th>
<th>Incentives (N=33)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of treatment arms per experiment</td>
<td>1.462 (0.85)</td>
<td>1.409 (0.80)</td>
<td>1.727 (1.28)</td>
</tr>
<tr>
<td>Number of follow-ups per experiment</td>
<td>1.359 (0.58)</td>
<td>1.273 (0.55)</td>
<td>1.364 (0.60)</td>
</tr>
<tr>
<td>Number of outcomes per experiment</td>
<td>2.051 (1.08)</td>
<td>2.091 (1.02)</td>
<td>2.030 (1.16)</td>
</tr>
<tr>
<td>Year:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-1990s</td>
<td>0.026</td>
<td>0.091</td>
<td>0</td>
</tr>
<tr>
<td>1990s</td>
<td>0.103</td>
<td>0.591</td>
<td>0.061</td>
</tr>
<tr>
<td>Post-1990s</td>
<td>0.872</td>
<td>0.318</td>
<td>0.939</td>
</tr>
<tr>
<td>Region:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>0.282</td>
<td>0.182</td>
<td>0.242</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>0.128</td>
<td>0.318</td>
<td>0.030</td>
</tr>
<tr>
<td>East Asia and Pacific</td>
<td>0.282</td>
<td>0.409</td>
<td>0.394</td>
</tr>
<tr>
<td>South Asia</td>
<td>0.308</td>
<td>0.091</td>
<td>0.333</td>
</tr>
<tr>
<td>GDP per capita in baseline year, US$ (2000)</td>
<td>1,284 (1265)</td>
<td>2,012 (1491)</td>
<td>1,113 (1220)</td>
</tr>
<tr>
<td>Grades included at baseline:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grades 1-4 only</td>
<td>0.564</td>
<td>0.318</td>
<td>0.576</td>
</tr>
<tr>
<td>Grades 5-8 only</td>
<td>0.077</td>
<td>0.091</td>
<td>0.091</td>
</tr>
<tr>
<td>Both</td>
<td>0.359</td>
<td>0.409</td>
<td>0.303</td>
</tr>
<tr>
<td>Uncertain</td>
<td>0</td>
<td>0.182</td>
<td>0.030</td>
</tr>
<tr>
<td>Published in:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Economics journal</td>
<td>0.205</td>
<td>0.182</td>
<td>0.303</td>
</tr>
<tr>
<td>Medical or nutrition journal</td>
<td>0</td>
<td>0.682</td>
<td>0</td>
</tr>
<tr>
<td>Psychology journal</td>
<td>0.026</td>
<td>0.046</td>
<td>0</td>
</tr>
<tr>
<td>Unpublished (working paper, report, etc.)</td>
<td>0.769</td>
<td>0.091</td>
<td>0.697</td>
</tr>
<tr>
<td>Convenience sample (vs. random sample)</td>
<td>0.769</td>
<td>0.864</td>
<td>0.697</td>
</tr>
<tr>
<td>Power analysis reported</td>
<td>0.179</td>
<td>0.409</td>
<td>0.212</td>
</tr>
<tr>
<td>Alternating list assignment</td>
<td>0.103</td>
<td>0.091</td>
<td>0.061</td>
</tr>
<tr>
<td>Cluster randomized</td>
<td>0.949</td>
<td>0.227</td>
<td>0.909</td>
</tr>
<tr>
<td>Stratification or pair-wise matching</td>
<td>0.667</td>
<td>0.818</td>
<td>0.606</td>
</tr>
</tbody>
</table>

Note: Each experiment consists of a control group and one or more treatment arms. See text for details.
Table 2: Characteristics of treatment arms

<table>
<thead>
<tr>
<th></th>
<th>Instructional inputs (N=53)</th>
<th>Health inputs (N=28)</th>
<th>Incentives (N=50)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Implementer:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government</td>
<td>0.151</td>
<td>0.107</td>
<td>0.280</td>
</tr>
<tr>
<td>Non-governmental organization</td>
<td>0.623</td>
<td>0.037</td>
<td>0.500</td>
</tr>
<tr>
<td>University or researcher</td>
<td>0.226</td>
<td>0.857</td>
<td>0.220</td>
</tr>
<tr>
<td><strong>Instructional inputs:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Materials</td>
<td>0.453</td>
<td>0</td>
<td>0.180</td>
</tr>
<tr>
<td>Computer or technology</td>
<td>0.283</td>
<td>0</td>
<td>0.040</td>
</tr>
<tr>
<td>Grants</td>
<td>0.113</td>
<td>0</td>
<td>0.060</td>
</tr>
<tr>
<td>Teacher training</td>
<td>0.547</td>
<td>0</td>
<td>0.180</td>
</tr>
<tr>
<td>Class size, small-group instruction, tracking</td>
<td>0.226</td>
<td>0</td>
<td>0.200</td>
</tr>
<tr>
<td><strong>Health inputs:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food, beverage, and/or micronutrients</td>
<td>0</td>
<td>0.643</td>
<td>0.020</td>
</tr>
<tr>
<td>De-worming drugs</td>
<td>0</td>
<td>0.250</td>
<td>0</td>
</tr>
<tr>
<td>Malaria drugs</td>
<td>0</td>
<td>0.036</td>
<td>0</td>
</tr>
<tr>
<td>Other health inputs</td>
<td>0</td>
<td>0.071</td>
<td>0</td>
</tr>
<tr>
<td><strong>Incentives:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information for students, parents or schools</td>
<td>0.057</td>
<td>0</td>
<td>0.380</td>
</tr>
<tr>
<td>Performance incentives for students or schools</td>
<td>0.038</td>
<td>0</td>
<td>0.180</td>
</tr>
<tr>
<td>Contract or volunteer teachers</td>
<td>0.245</td>
<td>0</td>
<td>0.260</td>
</tr>
<tr>
<td>School management or supervision</td>
<td>0.094</td>
<td>0.036</td>
<td>0.280</td>
</tr>
<tr>
<td><strong>For cluster-randomized experiments:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of clusters in treatment arm</td>
<td>51</td>
<td>20</td>
<td>67</td>
</tr>
<tr>
<td>(36) (15) (42)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of clusters in control</td>
<td>51</td>
<td>21</td>
<td>75</td>
</tr>
<tr>
<td>(30) (10) (53)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>For student randomized experiments:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of students in treatment arm</td>
<td>295</td>
<td>171</td>
<td>461</td>
</tr>
<tr>
<td>(205) (117) (19)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of students in control</td>
<td>295</td>
<td>156</td>
<td>461</td>
</tr>
<tr>
<td>(205) (111) (19)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Characteristics of experimental follow-ups and outcome measures

<table>
<thead>
<tr>
<th></th>
<th>Instructional Inputs</th>
<th>Health inputs</th>
<th>Incentives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (s.d.)</td>
<td>Mean (s.d.)</td>
<td>Mean (s.d.)</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Months of treatment exposure at</td>
<td>12.9 (8.8)</td>
<td>9.2 (11.1)</td>
<td>10.5 (7.0)</td>
</tr>
<tr>
<td>follow-up</td>
<td>70</td>
<td>35</td>
<td>67</td>
</tr>
<tr>
<td>Follow-up conducted &gt;1 month</td>
<td>0.100 (0.17)</td>
<td>0.114 (0.10)</td>
<td>0.075 (0.07)</td>
</tr>
<tr>
<td>after treatment?</td>
<td>70</td>
<td>35</td>
<td>67</td>
</tr>
<tr>
<td>Attrition at follow-up (proportion)</td>
<td>0.199 (0.06)</td>
<td>0.088 (0.03)</td>
<td>0.136 (0.02)</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>29</td>
<td>44</td>
</tr>
<tr>
<td>Abs. value of differential attrition (proportion)</td>
<td>0.040 (0.06)</td>
<td>0.023 (0.03)</td>
<td>0.024 (0.02)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td>Content of outcome:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language or reading</td>
<td>0.513 (0.17)</td>
<td>0.614 (0.03)</td>
<td>0.404 (0.02)</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td>Mathematics</td>
<td>0.408 (0.06)</td>
<td>0.295 (0.03)</td>
<td>0.421 (0.02)</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td>Composite score</td>
<td>0.079 (0.06)</td>
<td>0.091 (0.03)</td>
<td>0.175 (0.02)</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td>Source of items:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluator or NGO</td>
<td>0.421 (0.06)</td>
<td>0.114 (0.03)</td>
<td>0.298 (0.02)</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td>Commercial or international</td>
<td>0.079 (0.06)</td>
<td>0.545 (0.03)</td>
<td>0.193 (0.02)</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td>Government or school test</td>
<td>0.263 (0.06)</td>
<td>0.318 (0.03)</td>
<td>0.333 (0.02)</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td>Uncertain</td>
<td>0.237 (0.06)</td>
<td>0.023 (0.03)</td>
<td>0.175 (0.02)</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td><strong>Materials</strong></td>
<td>-0.052*</td>
<td>-0.064***</td>
<td>-0.020</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.022)</td>
<td>(0.021)</td>
</tr>
<tr>
<td><strong>Computers or technology</strong></td>
<td>0.022</td>
<td>0.043</td>
<td>0.066*</td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
<td>(0.042)</td>
<td>(0.034)</td>
</tr>
<tr>
<td><strong>Grants</strong></td>
<td>-0.063</td>
<td>-0.055*</td>
<td>-0.034</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.028)</td>
<td>(0.024)</td>
</tr>
<tr>
<td><strong>Teacher training</strong></td>
<td>0.090***</td>
<td>0.104***</td>
<td>0.072***</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.023)</td>
<td>(0.023)</td>
</tr>
<tr>
<td><strong>Class size, small-group</strong></td>
<td>0.073***</td>
<td>0.088***</td>
<td>0.069**</td>
</tr>
<tr>
<td>instruction, tracking</td>
<td>(0.021)</td>
<td>(0.022)</td>
<td>(0.033)</td>
</tr>
<tr>
<td><strong>Food, beverage, micronutrients</strong></td>
<td>-0.018</td>
<td>0.037</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.042)</td>
<td>(0.037)</td>
</tr>
<tr>
<td><strong>De-worming</strong></td>
<td>-0.043</td>
<td>-0.001</td>
<td>-0.017</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.036)</td>
<td>(0.037)</td>
</tr>
<tr>
<td><strong>Other health</strong></td>
<td>0.017</td>
<td>0.016</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.044)</td>
<td>(0.062)</td>
</tr>
<tr>
<td><strong>Information</strong></td>
<td>-0.008</td>
<td>0.016</td>
<td>-0.020</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.037)</td>
<td>(0.031)</td>
</tr>
<tr>
<td><strong>Performance incentives</strong></td>
<td>0.033</td>
<td>0.043</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.033)</td>
<td>(0.032)</td>
</tr>
<tr>
<td><strong>Contract or volunteer teachers</strong></td>
<td>-0.020</td>
<td>-0.039</td>
<td>-0.023</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.027)</td>
<td>(0.032)</td>
</tr>
<tr>
<td><strong>School management, supervision</strong></td>
<td>-0.010</td>
<td>-0.024</td>
<td>-0.020</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.028)</td>
<td>(0.031)</td>
</tr>
<tr>
<td><strong>Months of treatment exposure at follow-up</strong></td>
<td>-0.001</td>
<td>-0.001</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td><strong>Follow-up conducted &gt;1 month after treatment?</strong></td>
<td>-0.025</td>
<td>-0.009</td>
<td>-0.028</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.021)</td>
<td>(0.022)</td>
</tr>
<tr>
<td><strong>Government implementer</strong></td>
<td>-0.010</td>
<td>-0.016</td>
<td>-0.022</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.042)</td>
<td>(0.035)</td>
</tr>
<tr>
<td><strong>University/researcher implementer</strong></td>
<td>-0.075**</td>
<td>-0.006</td>
<td>-0.076**</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.026)</td>
<td>(0.031)</td>
</tr>
<tr>
<td><strong>Post-1990s</strong></td>
<td>-0.010</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Latin America and Caribbean</strong></td>
<td></td>
<td>0.129***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.047)</td>
<td></td>
</tr>
<tr>
<td><strong>East Asia and Pacific</strong></td>
<td></td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.035)</td>
<td></td>
</tr>
<tr>
<td><strong>South Asia</strong></td>
<td></td>
<td>0.030</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.040)</td>
<td></td>
</tr>
<tr>
<td><strong>Log(GDP per capita in baseline year, US$ (2000))</strong></td>
<td>-0.085***</td>
<td></td>
<td>-0.096***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.021)</td>
<td></td>
</tr>
<tr>
<td><strong>Grades 5-8 only</strong></td>
<td>-0.010</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Both grades 1-4 and 5-8</strong></td>
<td>-0.041</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Uncertain grades</strong></td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Convenience sample</strong></td>
<td>-0.037*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Composite test score</strong></td>
<td></td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.034)</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Math test score</td>
<td>-0.012</td>
<td>-0.007</td>
<td></td>
</tr>
<tr>
<td>(0.015)</td>
<td>(0.015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial/international test</td>
<td>-0.013</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>(0.027)</td>
<td>(0.029)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government or school test</td>
<td>-0.014</td>
<td>-0.019</td>
<td></td>
</tr>
<tr>
<td>(0.028)</td>
<td>(0.033)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncertain test</td>
<td>0.005</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>(0.030)</td>
<td>(0.041)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment is published</td>
<td>0.020</td>
<td>-0.020</td>
<td></td>
</tr>
<tr>
<td>(0.024)</td>
<td>(0.027)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternating list assignment</td>
<td>0.023</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>(0.039)</td>
<td>(0.038)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attrition at follow-up</td>
<td>0.083</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>(0.090)</td>
<td>(0.108)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Missing indicator)</td>
<td>0.014</td>
<td>-0.026</td>
<td></td>
</tr>
<tr>
<td>(0.027)</td>
<td>(0.024)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abs. value of differential attrition</td>
<td>-0.723**</td>
<td>-0.505</td>
<td></td>
</tr>
<tr>
<td>(0.276)</td>
<td>(0.310)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Missing indicator)</td>
<td>0.003</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td>(0.018)</td>
<td>(0.018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.057**</td>
<td>0.089***</td>
<td>0.650***</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.026)</td>
<td>(0.138)</td>
</tr>
</tbody>
</table>

Note: In each column, the sample include 256 effect sizes from 75 experiments, clustered within 58 studies. Robust standard errors are clustered by study (see text for additional details of sample and estimation). ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.
Table 5: The importance of stratification, strata controls, and baseline controls

<table>
<thead>
<tr>
<th></th>
<th>Dependent variable: $\theta_{i/k}$</th>
<th>Dependent variable: $\theta_{i/k}$</th>
<th>Dependent variable: $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard error of $\theta_{i/k}$</td>
<td>Standard error of $\theta_{i/k}$</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Stratification or pair-wise matching?</td>
<td>-0.006</td>
<td>--</td>
<td>-0.018</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.065)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>Strata controls?</td>
<td>-0.032**</td>
<td>--</td>
<td>-0.035</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.038)</td>
<td>(0.052)</td>
</tr>
<tr>
<td>Baseline controls?</td>
<td>-0.027**</td>
<td>-0.019***</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.006)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>N</td>
<td>348</td>
<td>348</td>
<td>348</td>
</tr>
<tr>
<td>Fixed effects?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Note: The stacked samples include unconditional and conditional estimates, as described in the text. Fixed effects indicate experiment-by-treatment-arm-by-follow-up-by-outcome cells, including 1 to 2 observations per cell. Robust standard errors are clustered by study. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.
Figure 1: The direct and indirect determinants of child learning
Figure 2: Dates of treatment and data collection in 76 experiments

Note: Numerical codes identify experiments (see the Appendix). 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.
Figure 3: Global and differential attrition at follow-up in experiments

Note: Each circle indicates one follow-up of one treatment group. Dashed lines indicate the median of each variable. The figure excludes one outlier.
Figure 4: Treatments with instructional materials

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 5: Treatments with computers or technology

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 7: Treatments with teacher training

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 9: Treatments with food, beverages, and/or micronutrients

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 10: Treatments with de-worming drugs

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 11: Treatments with information

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 12: Treatments with performance incentives

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 13: Treatments with contract or volunteer teachers

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 14: Treatments that modify school management or supervision

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 15: Funnel plot, with pseudo 95% confidence intervals
Figure 16: Effect sizes in 15 treatment arms that have comparable cost data

Notes: Numerical codes identify experiments (see the Appendix). Diamonds and brackets indicate effect sizes and 95% confidence intervals. See text for details.
Figure 17: Cost-effectiveness ratios in 15 treatment arms that have comparable cost data

Note: The x-axis indicates the cost per 0.2 units of the effect size on a log scale. Hollow diamonds indicate that the underlying effect size, from Figure 16, is not statistically significant at 10%.
Appendix: Codes of experiments included in the meta-analysis

Chile

China


Colombia

Ecuador

Gambia

Guatemala

India


**Indonesia**


**Jamaica**


**Kenya**


**Madagascar**


**Mali**


Mexico

Nepal

Nicaragua

Pakistan

Peru

Philippines

South Africa

Sri Lanka

Tanzania

Thailand

Uganda