Improving Learning Outcomes in South Asia: Findings from a Decade of Impact Evaluations

Salman Asim, Robert S. Chase, Amit Dar, and Achim Schmitten

There have been various initiatives to improve learning outcomes in South Asia. Still, outcomes remain stubbornly resistant to improvements, at least when considered across the region. To collect and synthesize the insights about what actually works to improve learning outcomes, this paper conducts a systematic review and meta-analysis of 21 education-focused impact evaluations from South Asia, establishing a standard that includes randomized control trials and quasi-experimental designs. It finds that while there are impacts from interventions that seek to increase the demand for education in households and communities, those targeting teachers or schools and thus the supply side of the education sector are generally much more adept at improving learning outcomes. Systematic Review, Meta-Analysis, Impact Evaluations, Education Outcomes, South Asia. JEL codes: I25, I21, O15

Promoting learning for all and improving learning outcomes are vital objectives, both for individuals and for countries as a whole. Education is essential to living a fuller and more productive life. For people, regardless of their poverty status, education is an end in itself: having the capability to read, to compute, and to understand the world around us opens opportunities and life chances that otherwise would be closed, allowing people to live fuller lives. Quality education can empower people to imagine and achieve what they thought out of reach, contributing to their own welfare and that of society in ways that they previously had not imagined. Further, from the perspective of individual productivity, education generates economic returns, increasing the capacity of individuals to improve livelihoods, manage shocks, and generate new economic opportunities, as well as take

advantage of them. Quality education is particularly valuable for the nearly 507 million people living in extreme poverty, most of whom have few other assets beyond their human potential. Given the benefits of education and their aggregation, providing learning for a country's populace can have positive impacts for a country as whole. For all countries, economic growth depends on productivity that in turn depends on quality education and learning for all.

South Asian countries have made important progress increasing educational attainment. South Asia's net primary enrollment rate was 75 percent in 2000 and rose to 89 percent by 2010. The number of children between the ages of eight and 14 years who are out of school fell from 35 million to 13 million between 1999 and 2010. Sri Lanka and the Maldives have consistently enrolled almost all their children in primary schools. Bhutan and India have recently made significant progress by increasing enrollment rates steadily to about 90 percent of children aged six to 14 years. In Pakistan, the primary net enrollment rate jumped from 58 percent to 74 percent between 2000 and 2011. Moreover, according to data from the UNESCO Institute for Statistics, between 2000 and 2010 South Asia's lower secondary enrollment rate increased from about 44 percent to 58 percent.

While there has been progress in improving access to education in the region, access for all remains elusive, particularly in getting the disadvantaged and marginalized into school, especially into postbasic education. Still, the key challenging policy agenda is to enhance the quality of education and to make progress toward improving learning outcomes—the ultimate goal of any educational system and (as, e.g., shown by [Hanushek and Woessmann 2012](#) and [Hanushek 2013](#)) a stronger driver of economic growth than years of schooling. A recent report on challenges, opportunities, and policy priorities in school education in South Asia by [Dundar et al. \(2014\)](#) establishes that mean student achievement in mathematics, reading, and language is very low throughout the South Asia region, except for Sri Lanka. For example, India's National Council of Educational Research and Training finds that only a third of grade five students can compute the difference between two decimal numbers. Similarly, in rural Pakistan, only 37 percent of grade five students can divide three-digit numbers by a single-digit number, while in rural Bangladesh more than a third of fifth graders do not even have grade three competencies (cf. [World Bank 2012](#)).

This paper conducts a systematic review and meta-analysis to contribute to the vital education and development issue of what approaches are most effective in improving learning in schools in South Asia. More specifically, it evaluates a crucial question that has long been discussed in the area of education interventions: what is the relative effectiveness of interventions that address the supply of education versus the demand for it? Some experts argue that education outcomes primarily depend on having educational services supplied through the provision of physical facilities, teachers who have incentives to educate children, curricula and materials for children to interact with, and all of these provided in a way that can be sustained

over time. Correspondingly, a significant proportion of education interventions focus on the supply of education. At the same time, other experts are convinced that to achieve improved learning children need to want to attend school and learn, supported by their families' demand for education. In turn, it is conjectured that the relative value that people within the local neighborhood or community place on education vis-à-vis other important demands on their time and resources will influence students' enthusiasm for learning within the classroom. Based on such views, several interventions seek to support the demand for education, generating enthusiasm that allows children to benefit from the educational opportunities on offer.

Over the past two decades, researchers have increasingly analyzed innovations that seek to improve development outcomes using impact evaluations. Impact evaluations assess changes in the well-being of individuals, households, and communities or firms that can be attributed to a particular project, program, or policy (cf. Baker 2000). In practice, it is often challenging to reliably estimate the impacts attributable to an intervention because this necessitates an answer to the question of what would have happened to those receiving the intervention if they had not, in fact, received the program. At the same time, analytic techniques that fall under the broad category of experimental and quasi-experimental designs have made it possible to make a lot of progress on this challenge.

Randomized control trial (RCT) studies represent the highest standard of impact evaluation evidence in development. This research methodology derives from techniques used in the physical sciences. Overall, the approach is to isolate a specific innovation, working to ensure that the innovation is the only systematic difference between the general population and the "treatment" group that participates in the experiment. The central design feature of the RCT approach is that the treatment group is randomly selected so that, for the purposes of statistical analysis, treatment and control groups are indistinguishable, save for the fact that the treatment group is subjected to the innovation whose impact the study is evaluating. By comparing development outcomes in the treatment group versus the control group, research that uses experimental methods is able to rigorously isolate any differences in outcomes that can be attributed to the tested innovations.

Quasi-experimental designs are likewise highly rigorous approaches to gathering evidence on the impact of development innovations. While these methods do not randomly assign participation in treatment and control groups, they use econometric techniques to exploit features of the process of how and where innovations were implemented that introduced random variation. Taking advantage of this variation, researchers using quasi-experimental methods are able to isolate whether the tested innovations make an attributable difference. Quasi-experimental designs are sometimes seen as comparatively less rigorous than randomized control designs. At the same time, they offer a number of advantages. In particular, quasi-experimental designs are often cheaper to implement than RCTs and offer the possibility to evaluate

programs after their introduction using existing data. Further, there are sometimes fewer concerns about the external validity of results with quasi-experiments, as this type of design is frequently used to analyze interventions introduced on either the national level or at least in a large geographical area.

Experimental and quasi-experimental evidence has been applied with increasing prevalence in countries around the world. Within that growing set of experimental and quasi-experimental design impact evaluations, there is a relatively large number of impact evaluations that have tested and gathered high-quality evidence on education in South Asia. As presented below, our systematic review of the relevant literature identifies 21 distinct studies that document the results of impact evaluations of education interventions in South Asia on learning outcomes that reached our standard of rigor.

Based on this systematic review, a number of meta-analyses allow us to address the question of what works to improve learning outcomes in South Asia. The main result of the meta-analyses is that supply-side interventions have the potential to induce moderate to important improvements in learning outcomes. In fact, supply-side interventions hold promise for improving learning levels irrespective of whether these learning levels are measured in terms of native language, mathematics, or overall test scores. In contrast, demand-side interventions seem to have some impacts on learning outcomes but are comparatively less effective.

While South Asia has been at the forefront of the movement to rigorously evaluate education-related interventions, a substantial number of such impact evaluations have also been conducted in other regions. Recently, this body of literature has been reviewed by several literature reviews. Together with a group of more verbal or narrative reviews (for instance, by [Glewwe et al. 2011](#); [Kremer, Brannen, and Glennerster 2013](#); and Murnane and Ganimian forthcoming), four works stand out because—similar to this paper—they combine a systematic literature review with a rigorous meta-analysis of the available evidence to investigate what kind of interventions are most effective in improving education outcomes. One of these four reviews ([Petrosino et al. 2012](#)) centers primarily on school enrollment and attendance. The other three ([Krishnaratne, White, and Carpenter 2013](#); [Conn 2014](#); and [McEwan 2015](#)) are mainly concerned with students' learning outcomes. Another differentiating factor among the four reviews is that [Petrosino et al. \(2012\)](#), [Krishnaratne, White, and Carpenter \(2013\)](#) and [McEwan \(2015\)](#) considered evidence from all over the developing world while [Conn \(2014\)](#) concentrated on Sub-Saharan Africa, a region that shares many similarities with South Asia but also exhibits a number of important differences. Reassuringly, the results of our systematic review and meta-analysis are generally very consistent with the findings from the four other methodologically comparable works. What makes our paper stand out is the clear focus on South Asia as well as our assessment of the question of whether interventions that focus on the supply or on the demand side of education are more promising.¹

The rest of this paper is structured as follows. The Methodology section provides methodological background on both the systematic review of the literature and the meta-analysis of learning outcomes. The Results section then systematically summarizes the available evidence. The final section concludes.

Methodology

Systematic Review

Our systematic review strives to identify, appraise, and synthesize rigorous education-related impact evaluations for South Asia. It serves as the basis of the ensuing meta-analysis and covers all research that fulfills the following three primary criteria: (a) the research evaluates the impact of one or more clearly-defined education-related interventions, (b) it measures effects on mathematics, native language, or composite test scores, and (c) it uses data for at least one South Asian country. Additionally, in order to be included in the systematic review, a number of additional, strict quality criteria also have to be satisfied. In particular, all causal statements need to be based on evidence gained from an RCT or a credible quasi-experiment; a well-defined “business-as-usual” control group has to be present; and an intervention’s effects have to be reported in a way that is transparent, standardized, and comparable to effects reported by other studies.

In the systematic search, a two-step approach was used to identify the universe of studies that document rigorously evaluated education interventions in South Asia. First, a search was performed in a number of databases such as the American Economic Association’s Econlit database, the World Bank’s Impact Evaluations in Education (IE²) database, the database of randomized evaluations from the Abdul Latif Jameel Poverty Action Lab (J-PAL), and an internal database of impact evaluations compiled by the office of the World Bank’s Chief Economist for the South Asia region. Then, additional relevant articles were identified through the reference lists of the articles found in the four databases. This “snowballing” is enormously helpful for the identification of unpublished (“gray”) or otherwise obscure literature and for evidence that does not neatly fit the inclusion criteria defined by the different databases. For instance, the American Economic Association’s EconLit database only lists studies in the economics literature.²

A rigorous search and appraisal of the available evidence combined with these strict quality criteria results in a set of 21 distinct studies with 36 distinct treatment arms that document the results of rigorous impact evaluations of education interventions in South Asia. The specific interventions evaluated by these studies, as well as the main impacts, are listed in [table 1](#). Additionally, the table lists whether a specific intervention primarily addresses the demand for or supply of education.

Table 1. Details on Coverage of Systematic Review and on Meta-Analysis of Interventions' Impacts

Study	Intervention	Supply vs demand	Impacts on native language test scores			Impacts on math test scores			Impacts on composite test scores		
			ES	95% CI	Wgt.	ES	95% CI	Wgt.	ES	95% CI	Wgt.
Andrabi, Das, and Khwaja (2015)	Report cards	Demand	0.10	0.02 0.18	4.13	0.15	0.03 0.26	2.02	0.09	0.01 0.16	4.49
Aturupane et al. (2014)	School-based management	Demand				0.22	0.07 0.37	1.19			
	Report cards	Demand				0.03	-0.12 0.19	1.13			
Banerjee et al. (2007)	Balsakhi (year 1)	Supply	0.08	-0.03 0.19	2.21	0.18	0.09 0.27	3.33	0.14	0.05 0.23	2.93
	Balsakhi (year 2)	Supply	0.19	0.09 0.29	2.78	0.35	0.22 0.49	1.48	0.28	0.17 0.40	1.80
	Computer-assisted learning	Supply	-0.03	-0.19 0.14	1.03	0.39	0.25 0.54	1.29	0.19	0.03 0.35	0.94
Banerji, Berry, and Shotland (2013)	Classes for mothers	Demand	-0.00	-0.04 0.04	19.24	0.04	-0.00 0.07	19.52	0.02	-0.02 0.05	20.00
	Training for mothers	Demand	0.03	-0.01 0.07	17.36	0.05	0.01 0.08	19.52	0.04	0.00 0.07	20.00
	Classes and training for mothers	Demand	0.05	0.02 0.09	19.24	0.07	0.03 0.10	21.75	0.06	0.03 0.10	22.43
Barrera-Ororio and Raju (2010)	Group bonuses	Supply							-0.87	-2.13 0.39	0.02
	Sanctions to schools	Supply							0.66	0.21 1.12	0.12
Barrera-Ororio et al. (2013)	New private schools	Supply	0.64	0.38 0.89	0.41	0.66	0.40 0.91	0.41	0.67	0.41 0.93	0.36
Borkum, He, and Linden (2013)	Libraries	Supply	-0.05	-0.15 0.05	2.67	-0.05	-0.17 0.07	1.89			
Burde and Linden (2013)	New schools	Supply							0.66	0.49 0.84	0.80
Chaudhury and Parajuli (2010)	School management (grade 3)	Demand	-0.43	-1.31 0.45	0.03	-0.72	-1.66 0.22	0.03			
	School management (grade 5)	Demand				0.21	-0.67 1.09	0.03			

continued

Table 1. Continued

Study	Intervention	Supply vs demand	Impacts on native language test scores			Impacts on math test scores			Impacts on composite test scores		
			ES	95% CI	Wgt.	ES	95% CI	Wgt.	ES	95% CI	Wgt.
Das et al. (2013)	School grants	Supply	0.08	0.00 0.15	4.81	0.09	0.01 0.17	3.99	0.09	0.01 0.16	4.49
Duflo, Hanna, and Ryan (2012)	Attendance verification	Supply	0.14	-0.02 0.30	1.09	0.18	-0.08 0.44	0.42	0.15	-0.03 0.33	0.80
He, Linden, and MacLeod (2008)	English curriculum 1	Supply				0.05	-0.09 0.19	1.40	0.26	0.08 0.44	0.78
	English curriculum 2	Supply				0.32	0.12 0.52	0.68	0.35	0.04 0.65	0.27
	English curriculum 3	Supply				0.28	0.06 0.51	0.51	0.37	0.07 0.66	0.29
	English curriculum 4	Supply				0.39	0.19 0.59	0.66	0.37	0.05 0.68	0.25
He, Linden, and MacLeod (2009)	Literacy program 1	Supply	0.26	0.08 0.44	0.82						
	Literacy program 2	Supply	0.44	-0.43 1.32	0.03						
	Literacy program 3	Supply	0.55	0.34 0.76	0.58						
	Literacy program 4	Supply	0.70	0.54 0.86	1.03						
Lakshminarayana et al. (2013)	Remedial teaching, etc.	Supply	0.64	0.52 0.76	1.93	0.73	0.59 0.87	1.44	0.75	0.63 0.87	1.80
Linden (2008)	Computer-assisted learning 1	Supply	-0.28	-0.71 0.15	0.14	-0.57	-1.06 -0.07	0.11	-0.48	-0.96 -0.00	0.11
	Computer-assisted learning 2	Supply	0.18	-0.15 0.50	0.25	0.28	-0.06 0.62	0.24	0.25	-0.09 0.60	0.21
Muralidharan and Sundararaman (2010)	Feedback to teachers	Supply	0.02	-0.06 0.11	3.59	-0.02	-0.11 0.08	3.06	0.00	-0.09 0.09	3.20
Muralidharan and Sundararaman (2011)	Group bonuses	Supply	0.11	0.02 0.20	3.14	0.18	0.06 0.29	2.17	0.14	0.04 0.24	2.59
	Individual bonuses	Supply	0.13	0.04 0.22	3.43	0.18	0.07 0.30	2.02	0.16	0.06 0.25	2.59

continued

Table 1. Continued

Study	Intervention	Supply vs demand	Impacts on native language test scores			Impacts on math test scores			Impacts on composite test scores		
			ES	95% CI	Wgt.	ES	95% CI	Wgt.	ES	95% CI	Wgt.
Muralidharan and Sundararaman (2013)	Contract teachers	Supply	0.08	0.01 0.15	5.67	0.11	0.04 0.19	4.63	0.09	0.03 0.16	5.29
Muralidharan and Sundararaman (2015)	School-choice program	Demand	-0.08	-0.19 0.03	2.30	-0.05	-0.18 0.07	1.67	0.01	-0.11 0.13	1.74
Rao (2014)	Mixing wealthy and poor students	Supply	0.03	-0.10 0.15	1.75	0.04	-0.06 0.13	3.19	-0.02	-0.16 0.11	1.40
Sarr et al. (2010)	Grants to schools	Supply	0.07	-0.22 0.35	0.32	0.19	-0.17 0.56	0.21	0.17	-0.20 0.39	0.29
(a) Supply			0.14	0.11 0.7	37.70	0.16	0.13 0.19	33.13	0.18	0.15 0.21	31.34
(b) Demand			0.03	0.00 0.05	62.30	0.05	0.03 0.07	66.87	0.04	0.02 0.06	68.66
Overall			0.07	0.05 0.09	100	0.09	0.07 0.11	100	0.09	0.07 0.10	100

Note: CI stands for confidence interval. ES stands for Effect Size.

Source: World Bank staff calculations.

Some of the important features of the studies included in the systematic review and meta-analysis are visualized in [figure 1](#). The figure summarizes the number of studies or interventions by whether they address the supply of or demand for education, as well as by country, methodology, and extent of World Bank involvement. Since some of the studies included in the final sample of the systematic review document impacts of more than one intervention, at times targeting both the supply of and demand for education, the upper left panel of figure 1 lists the number of interventions instead of the number of studies. We find that 27 interventions targeted the supply side of education, and only nine interventions the demand side. The other three panels of figure 1 are aggregated on the level of studies instead of interventions. Among other things, they show that, in geographic terms, the large majority of rigorous education-related impact evaluations focus on India.³

[Figure 1](#)’s lower left panel demonstrates that 17 of the 21 studies included in the final sample of the systematic review involve RCTs. The other remaining

Source: World Bank staff calculations based on studies listed in [Table 1](#).

studies rely on quasi-experimental designs such as difference-in-differences, regression discontinuity, or a combination thereof.⁴ Finally, the lower right panel of figure 1 categorizes the 21 studies according to the proportion of their authors who indicated being affiliated with the World Bank at the time of the study’s publication. This is one measure of how prominent the World Bank is in commissioning and executing rigorous impact evaluations of education-related interventions in South Asia.

The quality of the selected studies was assessed independently by two reviewers according to a five-point scale, separately for RCTs and quasi-experiments using sets of relevant indicators for each. For RCT-based studies, the indicators included the following questions: (a) Was the study published in a peer-reviewed journal or completed within the last three years? (b) Was randomization appropriate and no threats to internal validity present? (c) Was information provided on the number and reasons of drop out (attrition) of study participants? (d) Was the study sufficiently powered with explicit power calculations reported, or was the size of treatment/control clusters 50 or more? (e) Was the study sufficiently representative of the state/region where the experiment was conducted? For non-RCT studies, all indicators were the same except for indicator (d), which was replaced by an indicator of robustness of methods: Were the study results robust to at least two sets of statistical methods?

Table 2 summarizes the result of the critical assessment of the available evidence. We find that two-thirds of both RCT and non-RCT studies were published or completed within the last three years. Further, as expected, RCTs performed well on endogeneity concerns, ensuring internal validity in 95 percent of cases. The percentage is relatively high, at 82 percent for quasi-experiments as well. Approximately 73 percent of studies of both types reveal no systematic attrition, while only 5 percent of studies showed representativeness. RCTs generally revealed

Table 2. Quality of Available Evidence by Quality Criterion and Method		
Criterion	Sample	
	RCTs only (%)	Both RCTs and quasi-experiments (%)
Published or recent	78	73
No endogeneity concerns	95	82
No systematic attrition	72	73
Arm size ≥ 50	78	
No sample selectivity	6	5
Effects robust across methods	–	50

Notes: See text for definitions of criteria. Arm-size criterion used for RCTs only, robustness criterion for quasi-experiments only.

Source: World Bank staff calculations based on studies listed in Table 1.

good power, with three out of every four studies having size of 50 or more for clusters (“arm size ≥ 50 ”). However, only half of the quasi-experiments were robust across methods.⁵

Meta-Analysis

A meta-analysis relies on statistical techniques to combine results from two or more individual interventions. The objective is to improve the precision of estimated treatment effects and to assess whether these treatment effects vary for different types of interventions (cf. Egger, Davey Smith, and Phillips 1997). Thus, a meta-analysis is an ideally suited tool to utilize the education interventions in South Asia—identified through the systematic review in the last section—to derive rigorous and robust policy recommendations.

In the context of a meta-analysis, four methodological issues are of paramount importance: (a) the selection of the underlying evidence, (b) the grouping of individual interventions, (c) the weighting of the evidence, and (d) the selection of adequate outcome measures.

- (a) With regard to the selection of the underlying evidence, the meta-analyses described here are based on the set of studies listed in [table 2](#). That is, only those 21 studies are included in the meta-analyses that evaluate the impact of a clearly defined education-related intervention on an equally well-defined learning outcome, contain data for at least one South Asian country, and satisfy strict quality and reporting criteria.
- (b) Interventions are grouped according to whether they address demand or supply of education. The exact mapping of interventions into these categories is also listed in [table 2](#).⁶
- (c) Standard errors of interventions’ impacts are used for weighting the evidence. In other words, interventions for which impacts have been more precisely estimated are given larger weights as compared to interventions for which the exact size of the impact is less clear.
- (d) Adequate outcome measures should be pertinent to the question under study, widely used in the relevant literature, and easily comparable across different interventions. For these reasons, three distinct meta-analyses are undertaken. The first two are concerned with interventions’ impacts on standardized test scores in children’s native language and mathematics. Additionally, a meta-analysis for effects on overall or composite test scores is done.

In addition to these four fundamental methodological issues, a meta-analysis of education interventions in South Asia posed a number of more practical challenges. Most critically, some studies report interventions’ impacts on learning outcomes in

a way that is not directly comparable to estimates from other studies. For instance, [Lakshminarayana et al. \(2013\)](#) report estimated impacts on test scores in terms of changes on a percentage scale. Whenever possible, every effort was made to transform estimated impacts into standardized effects. The aim was to construct a set of evidence underlying the meta-analyses that was comparable, comprehensive, and unbiased. For instance, because baseline standard deviations are also reported by [Lakshminarayana et al. \(2013\)](#), their estimated impacts in terms of changes on a percentage scale could be transformed into impacts in terms of changes in standard deviations comparable to what was reported by other studies.

Moreover, many studies report results for different groups, time periods, and specifications. This means a decision had to be made about which impact estimates to include in the meta-analysis. Here, this decision was made according to the general principle that each meta-analysis would include exactly one impact estimate for each distinct intervention. That is, two impact estimates were included in a meta-analysis for a study detailing two distinct interventions, but only one impact estimate was included in the meta-analysis for a distinct intervention discussed by more than one study. Besides, whenever available, a meta-analysis would include pooled impact estimates for boys and girls. Only in cases where these estimates were not available would the meta-analysis include separate impact estimates for boys and girls. Similarly, whenever impact estimates for different time horizons were available, a meta-analysis would include the one closest to interventions' impacts after 12 months.⁷ Further, whenever impact estimates for different specifications were available, a meta-analysis would include the one for the author's (or authors') preferred specification. Finally, whenever separate impact estimates for an intervention's intent-to-treat effect and treatment effect on the treated were available, a meta-analysis would include the impact estimate for the intent-to-treat effect.⁸

In this context, it should be pointed out that the number of distinct interventions used in the meta-analyses is higher than the number of distinct studies. This is mostly because a large number of RCTs estimate impacts for more than one treatment arm. These different treatment arms might be used to analyze the impacts of very different types of treatments (for instance, the employment of a *balsakhi* and the introduction of computer-assisted learning; [Banerjee et al. 2007](#)); different implementations of the same intervention (such as four distinct implementations of a literacy program; [He, Linden, and MacLeod 2009](#)); or even the same intervention's impacts in different samples (e.g., school-based management in grades three and five; [Chaudhury and Parajuli 2010](#)).⁹ Since it often appears arbitrary whether different treatment arms are analyzed in one publication or in a larger number of separate studies, the meta-analyses treat them as entirely independent.

Of course, meta-analyses with different specifications than the one reported here would be conceivable. In particular, there are studies that list interventions'

impacts on outcome variables other than children's test scores in mathematics, language, or a composite measure of learning. Examples of alternative outcome variables used in the literature include standardized test scores in English, school enrollment, and attendance. However, comparable evidence on how different interventions impact outcomes other than those used here is available for such a small number of studies that it seems unfeasible to include them in a meta-analysis. The same caveat applies to meta-analyses of interventions' heterogeneous effects on different subgroups (like boys and girls) and of the costs of these interventions.¹⁰

Results

Main Results

Our main meta-analyses of education interventions in South Asia are visualized in figures 2–4. These figures are so-called forest plots and contain two pieces of information: on the left-hand side, all interventions included in the corresponding meta-analysis are listed. These are identified by the author(s) and year of the underlying study, as well as by the specific intervention, and grouped by whether they address the supply of or demand for education. On the right-hand side, these interventions' impacts on the outcome variable are visualized. For each individual intervention, the horizontal line gives a 95 percent confidence interval for the intervention's impact on the relevant outcome variable. Further, the solid black diamond indicates the point estimate of each intervention's impact. Again, for each intervention, the size of the gray rectangle is determined by the relative precision of the respective estimate. It thus represents the weight that any particular data point is given in the meta-analysis.

For each of the two groups of interventions (i.e., those targeting supply of or demand for education), the row marked "Subtotal" lists and visualizes the results of the actual meta-analyses. The visualization of results again contains different elements: The center of the transparent diamond indicates the meta-analysis' best estimate of the impact of a particular group of interventions on the corresponding outcome variable. The spread between the diamond's left and right edge gives a 95 percent confidence interval for this impact. In case the diamond overlaps with the black vertical line originating at zero, the impact of a particular group of interventions on the outcome variable is not statistically significantly different from zero (at least, not on the 5 percent level of statistical significance). In contrast, if the diamond and the line do not overlap, the meta-analysis reveals that a particular group of interventions has a statistically significant impact on the outcome variable.

Figure 2 contains the results of our first meta-analysis. This meta-analysis synthesizes the impacts of interventions targeting education supply or demand on children’s native language test scores. Depending on the specific intervention, the exact native language used in the construction of the outcome variable varies. Examples of languages in which tests underlying the meta-analysis are administered include Bangla (Sarr et al. 2010), Telugu (Muralidharan and Sundararaman

Figure 2. Meta-Analysis of Interventions’ Impacts on Native Language Test Scores

Notes: Learning impacts measured in standard deviations. For individual interventions, solid diamonds indicate point estimates, black lines 95 percent confidence intervals, and gray rectangles estimates’ weights in meta-analysis. For subtotals, centers of transparent diamonds indicate point estimates, and the spreads between diamonds’ left and right edges 95 percent confidence intervals. The sample includes both RCTs and quasi-experiments.

Source: World Bank staff calculations based on studies listed in Table 1.

2010), and Urdu (Andrabi, Das, and Khwaja 2015). As figure 2 shows, the meta-analysis of interventions' impact on children's native language test scores contains 20 interventions in the supply-side category. On the other hand, six interventions address the demand for education.

In terms of results, both groups of interventions have a statistically significant impact on native language test scores: overall, supply-centric interventions increase average native language test scores by 0.14 standard deviations, while those addressing the demand side of education lead to a low increase of average native language test scores by 0.03 standard deviations. On the 5 percent level, the impact of both types of interventions is statistically significantly different from zero.

Among all interventions included in the meta-analysis of impacts on native language test scores, the most effective ones are the establishment of publicly funded private primary schools (Barrera-Ororio et al. 2013), which raises test scores by 0.64 standard deviations; the provision of supplementary remedial teaching by community volunteers, combined with the distribution of learning materials and additional material support for some girls, which leads to an increase in test scores of 0.64 standard deviations (Lakshminarayana et al. 2013); and an intervention that introduces a new English education curriculum, which boosts native language test scores by 0.70 standard deviations (He, Linden, and MacLeod 2009). All three interventions are supply-centric by design.

In figure 3, the results of a second meta-analysis are reported. This meta-analysis compiles the impacts of different interventions on children's mathematics test scores. As with the meta-analysis of impacts on native language test scores, this second meta-analysis is also based on 20 interventions addressing the supply of education. However, the number of interventions addressing the demand for education is larger for this category. This is because more impact evaluations of interventions in this group analyze impacts on math test scores. As a result, figure 3 lists nine distinct interventions centered on households or connected to right-to-information innovations, community involvement approaches, or similar programs.

Figure 3 shows that both groups of interventions have a statistically significant impact on average mathematics test scores: Interventions focusing on the supply of education have a relatively larger effect, raising average math test scores by 0.16 standard deviations. Interventions targeting households and communities also increase average math test scores in a statistically significant way, but the meta-analysis reveals that, at 0.05 standard deviations, their impact is relatively small.

As is the case for the meta-analysis of interventions' impacts on native language test scores, the most promising interventions with regard to raising math test scores are the establishment of publicly funded private primary schools (Barrera-Ororio et al. 2013) and the provision of supplementary remedial teaching by

Figure 3. Meta-Analysis of Interventions' Impacts on Math Test Scores

Notes: Learning impacts measured in standard deviations. For individual interventions, solid diamonds indicate point estimates, black lines 95 percent confidence intervals, and gray rectangles estimates' weights in meta-analysis. For subtotals, centers of transparent diamonds indicate point estimates, and the spreads between diamonds' left and right edges 95 percent confidence intervals. The sample includes both RCTs and quasi-experiments.

Source: World Bank staff calculations based on studies listed in [Table 1](#).

community volunteers ([Lakshminarayana et al. 2013](#)). The establishment of private primary schools increases average test scores in mathematics by 0.66 standard deviations, and the provision of supplementary remedial teaching induces an average increase in math test scores of 0.73 standard deviations. Both these effects are even higher than those recorded for the respective interventions' impacts on native language test scores (0.64 standard deviations for both respective interventions).

Lastly, the results of the third meta-analysis are depicted in [figure 4](#). This meta-analysis is concerned with different interventions' impacts on children's overall test scores. Overall test scores are defined as composite learning scores, that is, test scores that are constructed by combining standardized test scores from at least two different subjects. Depending on the intervention, the exact definition of composite learning scores varies. In particular, composite learning scores might or might not draw on native language and math scores used in one of the other two meta-analyses.

Figure 4. Meta-Analysis of Interventions' Impacts on Composite Test Scores

Notes: Learning impacts measured in standard deviations. For individual interventions, solid diamonds indicate point estimates, black lines 95 percent confidence intervals, and gray rectangles estimates' weights in meta-analysis. For subtotals, centers of transparent diamonds indicate point estimates, and the spreads between diamonds' left and right edges 95 percent confidence intervals. The sample includes both RCTs and quasi-experiments.

Source: World Bank staff calculations based on studies listed in [Table 1](#).

The meta-analysis visualized in [figure 4](#) draws on 27 interventions. Twenty-two of these interventions fall into the supply-side category, and five are centered on education demand. [Figure 4](#) demonstrates that both groups of interventions have a statistically significant impact on overall test scores. What is more, for at least one group, the impacts are relatively substantial: Interventions focusing on the supply of education raise average overall test scores by 0.18 standard deviations. In contrast, interventions directly targeting the demand for education raise composite test scores by a mere 0.04 standard deviations. This is a relatively weak effect, but one that is statistically significant.

Once more, the establishment of publicly funded private primary schools ([Barrera-Ororio et al. 2013](#)) and the provision of supplementary remedial teaching by community volunteers ([Lakshminarayana et al. 2013](#)) are among the most promising interventions. The first intervention raises average overall test scores by 0.67 standard deviations, and the second one increases them by 0.75 standard deviations. Two other interventions are also very effective. These are the establishment of a school in a village for the first time ([Burde and Linden 2013](#)) and the threatening of private schools with the withdrawal of public funds if they fail to achieve a minimum pass rate criterion ([Barrera-Ororio and Raju 2010](#)). Given the effects on overall learning scores of 0.66 standard deviations each, both interventions appear to be potent mechanisms for raising average overall test scores.

The main findings from the three meta-analyses of interventions’ impacts on native language, mathematics, and overall learning levels are summarized in [table 3](#). In the table, interventions are again grouped by whether they address the supply or demand side of education. The table also gives examples of especially promising interventions.

One of the main results of the meta-analyses is that supply-side interventions have the potential to induce moderate to important improvements in learning

Table 3. Summary of Results of the Meta-Analyses				
Actor	Effects			Examples of especially promising interventions
	Language	Math	Composite	
(a) Supply	+	++	++	Supplementary remedial teaching by community volunteers; publicly funded private primary schools; revised English education curriculum
(b) Demand	(+)	+	(+)	School management structures and training/support services for community members
<i>Notes:</i> (+) means effect smaller than 0.05 SD ; + implies effect between 0.05 and 0.15 SD ; and ++ stands for effect between 0.15 and 0.25 SD . A meta-analysis of community-focused interventions on overall learning outcomes is impossible because only one intervention falls into this category. <i>Source:</i> World Bank staff calculations based on studies listed in Table 1 .				

outcomes. As [table 3](#) makes clear, supply-side interventions hold promise for improving learning levels, irrespective of whether these learning levels are measured in terms of native language, mathematics, or overall test scores. In contrast, demand-side interventions seem to be less effective in improving learning outcomes.

The full set of quantitative results of the three meta-analyses is summarized in [table 1](#). Additionally, the table also contains the results of the three meta-analyses on learning effects of education interventions. These results are displayed in the last row of the table. They show that, overall, education interventions that have been the subject of rigorous impact evaluations in South Asia increase native language test scores by 0.07 standard deviations, mathematics test scores by 0.09 standard deviations, and composite test scores again by 0.09 standard deviations. While the policy implications of a meta-analysis of such a broad and diverse set of interventions might be less clear, one could certainly follow the interpretation of [Krishnaratne, White, and Carpenter \(2013, 42\)](#), who performed a similar exercise for education interventions from all over the world and concluded simply by saying that “interventions aimed at getting children into school work.”

Alternative Groupings

In order to better understand the process behind the findings of the meta-analyses of education-related interventions in South Asia, [figures 5](#) and [6](#) explore how results change when only impact estimates derived from RCTs are considered or whether conclusions differ systematically between those studies where at least one author was affiliated with the World Bank at the time of publication, as well as impact evaluations where this was not the case. Both figures focus on interventions’ impacts on composite test scores. This is because composite test scores are the most comprehensive outcome variable utilized in the meta-analyses of the Main Results section.

[Figure 5](#) groups interventions according to whether they address the supply of or demand for education. In contrast to the meta-analysis visualized in [figure 4](#), it only uses evidence derived from RCTs and not from quasi-experiments. As discussed earlier, RCTs are generally considered to be more rigorous than quasi-experiments. Therefore, the objective here is to determine whether a focus on more rigorous evaluations changes any substantial results or conclusions. Of course, dropping the evidence from quasi-experimental studies reduces the number of interventions that can be incorporated into the meta-analysis. However, the reduction in the number of interventions is relatively small: the meta-analysis visualized in [figure 5](#) draws on 23 interventions, that is, only four less than the meta-analysis described in [figure 4](#). Eighteen out of the 23 interventions that utilize

Figure 5. Meta-Analysis of Interventions' Impacts on Composite Test Scores (RCTs only)

Notes: Learning impacts measured in standard deviations. For individual interventions, solid diamonds indicate point estimates, black lines 95 percent confidence intervals, and gray rectangles estimates' weights in meta-analysis. For subtotals, centers of transparent diamonds indicate point estimates, and the spreads between diamonds' left and right edges 95 percent confidence intervals. The sample includes only RCTs.

RCTs fall into the supply-side category, while five belong to the demand-side group.

Figure 5 shows that the results from a meta-analysis that only uses evidence derived from RCTs are very similar to those derived for the full sample of rigorously evaluated education-related interventions in South Asia. As before, both groups of interventions have a statistically significant impact on composite test scores. Moreover, at 0.19 and 0.04 standard deviations, the effects of supply- and demand-centric interventions are similar or identical to the effects identified for the interventions in the sample that included both RCTs and quasi-experiments.

In [figure 6](#), we categorize interventions according to whether any of the author(s) of the corresponding study indicated being affiliated with the World Bank at the time of the study’s publication. The objective of this exercise is to determine whether impact evaluations with strong World Bank involvement are more or less likely to find positive impacts of education-related interventions. If interventions where at least one author was affiliated with the World Bank at the time of publication are more likely to find positive learning impacts, this could be a sign that the World Bank is more successful in identifying promising interventions than

Figure 6. Meta-Analysis of Interventions’ Impacts on Composite Test Scores (by Affiliation)

Notes: Learning impacts measured in standard deviations. For individual interventions, solid diamonds indicate point estimates, black lines 95 percent confidence intervals, and gray rectangles estimates’ weights in meta-analysis. For subtotals, centers of transparent diamonds indicate point estimates, and the spreads between diamonds’ left and right edges 95 percent confidence intervals. The sample includes both RCTs and quasi-experiments.

Source: World Bank staff calculations based on studies listed in [Table 1](#).

other institutions. Alternatively, it could also be interpreted as a worrying symptom of the internal World Bank political economy: impact evaluations performed by World Bank staff are often tied to operational projects, and, allegedly, there is pressure on staff to demonstrate that these operational projects are producing meaningful results on the ground.

As [figure 6](#) illustrates, it is futile to speculate what could explain differing results between impact evaluations with and without strong World Bank engagement. This is because overall impacts on composite test scores from both groups of interventions are extremely similar: the meta-analysis reveals that the 11 interventions that benefitted from the involvement of authors affiliated with the World Bank increased composite test scores by 0.09 standard deviations, while the 16 interventions without direct involvement of World Bank staff raised these test scores by a close 0.08 standard deviations. Statistically, the two impacts are not significantly different.

Conclusions

Promoting quality education and learning for all is a vital objective in South Asia, particularly at this point in the region's development trajectory. Recognizing the importance of this goal, South Asian countries have made genuine, impressive progress to improve education access for their people. South Asia's net primary enrollment rate stood at 75 percent in 2000 and had risen to 89 percent by 2010. Concurrently, the number of out-of-school children between the ages of eight and 14 years fell from 35 million in 1999 to 13 million in 2010. Progress has been made throughout the region, and there has been significant movement toward gender parity, as well as some success in drawing the more marginalized into schools.

While there has been progress to improve access to education in the region, access for all remains elusive, particularly in getting the most disadvantaged and marginalized into school. At the same time, the key challenge now is to enhance the quality of education. A recent report ([Dundar et al. 2014](#)) on opportunities, challenges, and policy priorities in school education in South Asia established that mean student achievements in mathematics, reading, and language are very low throughout the region, perhaps with the exception of Sri Lanka.

However, while awareness of these challenges is a vital precondition to addressing them, a large and unresolved policy question is how to improve learning outcomes in South Asia. This paper has argued that an answer to this policy question might be found with the help of an important trend that has been boosting our capacity to address development challenges: development researchers have increasingly used rigorous impact evaluations to analyze innovations that seek to

improve development outcomes both in education and other fields. Using innovative analytic techniques that fall under the broad category of experimental and quasi-experimental designs, impact evaluations have made it possible to analyze whether innovations really make the differences they intend to make. In particular, in the education sector, South Asia has been at the forefront of this movement, and, by now, a sufficient body of evidence exists to draw stringent conclusions with regard to the impacts of a large number and variety of policy innovations.

Among the policy-related lessons that emerge from a systematic review and meta-analysis of the available evidence is that demand-side interventions that target either individual households or whole communities can be somewhat effective in raising learning outcomes. However, interventions that target teachers or schools and, thus, the supply side of the education sector are generally much more adept at improving learning outcomes.

These findings from rigorous impact evaluations provide guidance on policy options that are worth pursuing in order to improve learning outcomes. However, even if the systematic review and meta-analysis shows that a specific group of interventions has positive effects on learning outcomes, these interventions should not be seen as one-size-fits-all solutions that work in every context. Instead, it is important to keep a few caveats in mind. In particular, evidence on many important topics is still relatively scarce—for example, on incentives to learn, increases in school choice, and the promotion of nonstate education. Further, even for groups of interventions where more evidence is available, this evidence does not always point in the same direction. While these are exactly the reasons for focusing on South Asia as a distinct region with some shared characteristics and for doing a rigorous meta-analysis—which is essentially a statistical tool used to combine results from two or more individual interventions—the current state of knowledge allows us to perform meaningful meta-analyses only on a rather high level of aggregation. This means that the meta-analysis allows us to draw firm conclusions related more to relatively abstract concepts (education supply and demand) than to specific interventions, such as, for instance, the engagement of volunteer teachers recruited from local communities or incentives to learn for children or their parents.

Besides, impact evaluations tend to analyze very specific interventions under tightly controlled conditions and, frequently, for a relatively small or pilot sample. This has led to criticism about whether results from impact evaluations are generalizable to different contexts or even to other subpopulations in a shared context. While this kind of critique does not apply exclusively to impact evaluations but is also common to empirical research generally, the issue of external validity certainly is critical, and every planned policy intervention that is being justified through results from relevant impact evaluations will have to answer questions such as: Are the conditions on the ground similar to the ones in the setting where

the impact evaluation(s) was/were conducted? Will costs or effects of the intervention change when it is scaled up; that is, are there positive or negative returns to scale? Will there be general equilibrium effects; that is, will the scaling up of the policy lead to spillovers or changes in behavior that did not happen when it was piloted on a much smaller scale? Only a thorough deliberation of these kinds of questions will ensure that a scaled-up version of a successful pilot program or even the replication of an intervention in a different context will have similarly positive impacts on education-related outcome variables.

Two other issues that have to be considered during a review of existing impact evaluations of education-related interventions related to the interventions' costs and objectives. As mentioned above, comparable evidence on the costs of different interventions is available for such a small number of studies that it seemed unfeasible to include them in our meta-analysis. A meta-analysis of the cost-effectiveness of supply- and demand-side interventions—an issue of high policy relevance—is therefore left for future research. Besides, for some interventions, an increase in learning outcomes might not be the primary objective, and so, even if these interventions fail to improve learning outcomes, they might still be worthwhile if they are an effective instrument to achieve other objectives. For instance, there is some evidence that demand-side interventions are more appropriate for widening access to education than for improving learning outcomes. At least in those circumstances, where access for all and, in particular, for marginalized or disadvantaged groups remains an issue, this might be a strong rationale for investments in demand-side interventions that might otherwise not show significant improvements in learning outcomes, such as improved test scores. Moreover, many demand-side interventions—such as conditional cash transfers targeting households—are motivated more by equity than by efficiency considerations, and there are strong indications that some of the demand-side interventions included in our meta-analysis that showed no or few significant effects on learning (like those evaluated by [Chaudhury and Parajuli 2010](#) or [Banerji, Barry, and Shotland 2013](#)) at the same time improved equity.¹¹ Again, this might be a rationale for investments in these kinds of interventions despite their somewhat underwhelming effects on learning outcomes.

These considerations illustrate that evidence from impact evaluations should not be blindly trusted. Rather, it should be seen as one building block of informed decision-making. At the same time, one should not overlook the huge potential offered by the relatively plentiful and high-quality education-related impact evaluations that have been conducted in South Asia. During the last 10 years, these impact evaluations have created more knowledge on the impacts of a diverse set of often innovative interventions than was available earlier. Moreover, South Asia is well ahead of other regions in terms of the number of education-related impact evaluations. As shown, the average quality of these impact evaluations is quite

impressive, with most of them paying close attention to common methodological pitfalls and biases. Today, the region enjoys a unique opportunity to benefit from the knowledge that has been created over the last 10 years and to harness this knowledge for evidence-based policy making. If used appropriately, the rigorous impact evaluations of education-related interventions could help South Asian economies in mastering solutions to formidable policy challenges that they currently face in terms of strengthening their education systems and improving learning outcomes.

Notes

Salman Asim, Robert S. Chase (email: rchase@worldbank.org), Amit Dar, and Achim Schmitten are, respectively, Economist, Adviser-Strategy, Director-Strategy and Operations, and Economist with the World Bank. The authors thank Martin Rama and Jesko Hentschel for guidance and Peter Lanjouw (the editor), Mohan Prasad Aryal, Harsha Aturupane, Tara Beteille, Shantayanan Devarajan, Sangeeta Goyal, Elisabeth King, Tobias Linden, Matthew Morton, Shinsaku Nomura, Florencia Pinto, Lant Pritchett, Dhushyanth Raju, Venkatesh Sundararaman, Huma Ali Waheed, other World Bank colleagues, and three anonymous referees for helpful comments and suggestions. Findings, interpretations, and conclusions expressed in this paper do not necessarily represent the views of the World Bank, its affiliated organizations, its Executive Directors, or the governments these represent.^a

1. In a recent review of systematic reviews and meta-analyses of education-related interventions, [Evans and Popova \(2015\)](#) document some overlap but also quite a bit of divergence with regard to the different reviews' and meta-analyses' findings. This makes it all the more important to investigate what kind of interventions in what kind of regional or otherwise peculiar settings are in fact effective at improving learning outcomes. With our clear focus on demand- versus supply-side interventions and on South Asia, we aim to contribute to this endeavor.

2. Most systematic reviews use "snowballing" in addition to or instead of relying exclusively on searching for a set of key words or strings in different databases. Conclusions and patterns identified through both approaches tend to be quite similar (cf. [Greenhalgh and Peacock, 2005](#); [Jalali and Wohlin 2012](#)).

3. No study contains data for more than one South Asian country. [Das et al. \(2013\)](#) evaluate evidence both from India and Zambia.

4. Difference-in-differences methods compare a treatment and comparison group (first difference) before and after a program (second difference). Unlike in an RCT, assignment to a treatment or comparison group is not random ([Baker 2000](#)). Regression discontinuity designs are possible for programs that assign a threshold separating the treatment and control groups by comparing data points on either side of the threshold. A number of other techniques, such as regressions controlling for fixed effects, instrumental variable regressions, and propensity score matching estimators, are sometimes included in a broader definition of quasi-experimental designs but not considered in our systematic review and the ensuing meta-analysis because they are arguably less rigorous than RCTs or quasi-experiments in the narrow sense of the term.

5. An aggregation of the findings from table 2 shows that, irrespective of which method is used and whether the supply side or the demand side is addressed, all average scores are at least three (out of five) and sometimes considerably higher on the quality scale. While this is far from a perfect result, it still appears remarkable that the relatively young literature that uses rigorous methods to evaluate education-related interventions in South Asia has already attained such a respectable level of quality. We also use funnel plots to test for publication bias in the selected studies. Publication bias refers to a higher probability of scientific journals publishing studies that reveal significant, positive impacts, as

compared to studies that find negative effects or effects that are not statistically significant. We find that with regard to studies that address both the supply of and the demand for education, the selected studies appear largely free of significant publication bias. Further details on this test for publication bias are provided in Appendix B.

6. Of course, groupings other than demand- versus supply-side interventions would have also been possible. For instance, a large literature differentiates between interventions that provide resources to education actors and those that seek to influence those actors' incentives (cf. e.g., Hanushek 1995; Kremer 1995; and Glewwe et al. 2011).

7. The impact estimates closest to the interventions' impacts after 12 months were selected because the median time span between intervention and measurement of impacts was 12 months. While this approach maximized the comparability of impact estimates from different studies, 12 months is a relatively short period of time. King and Behrman (2009) point out that in particular in the human development sector longer time spans might be more adequate for assessing cumulative effects. In fact, Muralidharan and Sundararaman (2011) and Muralidharan (2012) show that group and individual teacher performance pay programs implemented across the Indian state of Andhra Pradesh initially have very similar effects on learning outcomes. However, over the course of five years, there are consistently positive and significant impacts of the individual teacher incentive program, while the impacts of the group teacher program decrease over time and eventually become insignificant.

8. One of the interventions evaluated by Sarr et al. (2010) combines grants to schools and education allowances to students and, thus, addresses the supply of and demand for education at the same time. It is not included in the meta-analysis.

9. At the same time, at least two interventions are discussed by more than one study. The short-run impacts of both individual and group bonuses for teachers in Andhra Pradesh, India, are analyzed in Muralidharan and Sundararaman (2011), whereas the medium- and long-term impacts are evaluated in Muralidharan (2012).

10. It should be noted that meta-analyses presume a normal distribution of impacts—or at least one that is not skewed—and thus are sometimes misleading in the context of heterogeneity or bimodal distributions of impacts (cf. Kremer 1995).

11. In certain circumstances, a trade-off might even exist between the objectives of improving learning and improving equity. Chaudhury and Parajuli (2010) note that their finding of positive enrollment but insignificant average learning effects of schools with devolved school management might be due to the ability of these schools to both attract and retain children from marginalized households. More research on this topic is needed.

References

- Andrabi, T., J. Das, and A. I. Khwaja. 2008. "A Dime a Day: The Possibilities and Limits of Private Schooling in Pakistan." *Comparative Education Review* 52 : 329–55.
- . 2015. "Report Cards: The Impact of Providing Schools and Child Test Scores on Educational Markets." Policy Research Paper 7226. World Bank, Policy Research Department, Washington, DC.
- Aslam, M., and G. Kingdon. 2011. "Evaluating Public Per-Student Subsidies to Low-Cost Private Schools – Regression-Discontinuity Evidence from Pakistan." *Economics of Education Review* 30 : 559–74.
- Aturupane, H., P. Glewwe, T. Keeleghan, R. Ravina, U. Sonnadara, and S. Wisniewski. 2014. "An Assessment of the Impacts of Sri Lanka's Programme for School Improvement and School Report Card Programme on Students' Academic Progress." *Journal of Development Studies* 50 : 1647–69.

- Baker, J. 2000. "Evaluating the Impact of Development Projects on Poverty – A Handbook for Practitioners." Directions in Development. World Bank, Washington, DC.
- Banerjee, A., D. Cole, E. Duflo, and L. Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics* 122 : 1235–64.
- Banerjee, A., E. Duflo, R. Glennerster, and D. Kothari. 2010. "Improving Immunisation Coverage in Rural India: Clustered Randomised Controlled Evaluation of Immunisation Campaigns with and without Incentives." *British Medical Journal* 340 : c2220.
- Banerji, R., J. Berry, and M. Shotland. 2013. "The Impact of Mother Literacy and Participation Programs on Child Learning: Evidence from a Randomized Evaluation in India." Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (J-PAL).
- Barrera-Osorio, F., and D. Raju. 2010. "Short-run Learning Dynamics under a Test-based Accountability System." Policy Research Paper 4836. World Bank, Policy Research Department, Washington, DC.
- . 2015. "Evaluating the Impact of Public Student Subsidies on Low-Cost Private Schools in Pakistan." *Journal of Development Studies* 51 : 808–25.
- Barrera-Osorio, F., D. Blakeslee, M. Hoover, L. Linden, D. Raju, and S. Ryan. 2013. "Leveraging the Private Sector to Improve Primary School Enrolment: Evidence from a Randomized Controlled Trial in Pakistan." Cambridge, MA: Harvard Graduate School of Education (HGSE).
- Berry, J. 2015. "Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India." *Journal of Human Resources* 50 : 1,051–80.
- Borkum, E., F. He, and L. Linden. 2012. "The Effects of School Libraries on Language Skills: Evidence from a Randomized Controlled Trial in India." Working Paper 18183. National Bureau of Economic Research.
- Chaudhury, N., and D. Parajuli. 2010. "Giving it Back: Evaluating the Impact of Devolution of School Management to Communities in Nepal." Unpublished manuscript, World Bank, Washington, DC.
- Conn, K. 2014. "Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-analysis of Rigorous Impact Evaluations." New York: Columbia University.
- Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan, and V. Sundararaman. 2013. "School Inputs, Household Substitution, and Test Scores." *American Economic Journal: Applied Economics* 5 : 29–57.
- Duflo, E., R. Hanna, and S. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4) : 1,241–78.
- Dundar, H., T. Beteille, M. Riboud, and A. Deolalikar. 2014. "Student Learning in South Asia: Challenges, Opportunities, and Policy Priorities." World Bank, Washington, DC.
- Egger, M., G. Davey Smith, and A. N. Phillips. 1997. "Meta-analysis: Principles and Procedures." *British Medical Journal* 315 : 1,533.
- Egger, M., G. Davey Smith, M. Schneider, and C. Minder. 1997. "Bias in Meta-analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315 : 629–34.
- Evans, D., and A. Popova. 2015. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." Policy Research Paper 7203. World Bank, Policy Research Department, Washington, DC.
- Glewwe, P., E. Hanushek, S. Humpage, and R. Ravina. 2011. "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010." In P. Glewwe, ed., *Education Policy in Developing Countries*, 13–64. Chicago: University of Chicago Press.
- Greenhalgh, T., and R. Peacock. 2005. "Effectiveness and Efficiency of Search Methods in Systematic Reviews of Complex Evidence: Audit of Primary Sources." *British Medical Journal* 331 : 1,064–65.

- Hanushek, E. 1995. "Interpreting Recent Research on Schooling in Developing Countries." *World Bank Research Observer* 10 : 227–46.
- . 2013. "Economic Growth in Developing Countries: The Role of Human Capital." *Economics of Education Review* 37 : 204–12.
- Hanushek, E., and L. Woessmann. 2012. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17 : 267–321.
- He, F., L. Linden, and M. MacLeod. 2008. "How to Teach English in India: Testing the Relative Productivity of Instruction Methods within the Pratham English Language Education Program." New York: Columbia University. Mimeographed Document.
- . 2009. "A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trial." Unpublished Manuscript. New York: Columbia University.
- IEG (Independent Evaluation Group). 2011. "Do Conditional Cash Transfers Lead to Medium-Term Impacts? Evidence from a Female School Stipend Program in Pakistan." World Bank, Washington, DC.
- Jalali, S., and C. Wohlin. 2012. "Systematic Literature Studies: Database Searches vs. Backward Snowballing." International Conference on Empirical Software Engineering and Measurement, September 19–20, 2012, Lund, Sweden.
- Jueni, P., F. Holenstein, J. A. C. Sterne, C. Bartlett, and M. Egger. 2002. "Direction and Impact of Language Bias in Meta-Analysis of Controlled Trials: Empirical Study." *International Journal of Epidemiology* 31 : 115–23.
- King, E. M., and J. R. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *World Bank Research Observer* 24 : 55–82.
- Kingdon, G. 1996. "The Quality and Efficiency of Private and Public Education: A Case-Study of Urban India." *Oxford Bulletin of Economics and Statistics* 58 : 57–82.
- Kremer, M. 1995. "Research on Schooling: What We Know and What We Don't: A Comment on Hanushek." *World Bank Research Observer* 10 : 247–54.
- Kremer, M., C. Brannen, and R. Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340 : 297–99.
- Krishnaratne, S., H. White, and E. Carpenter. 2013. "Quality Education for all Children? What Works in Education in Developing Countries." Working Paper 20. International Initiative for Impact Evaluation (3ie).
- Lakshminarayana, R., A. Eble, P. Bhakta, C. Frost, P. Boone, D. Elbourne, and V. Mann. 2013. "The Support to Rural India's Public Education System (STRIPES) Trial: A Cluster Randomised Controlled Trial of Supplementary Teaching, Learning Material and Material Support." *PLOS ONE* 8 (7).
- Linden, L. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India." Working Paper 17. World Bank Infodiv.
- McEwan, P. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-analysis of Randomized Experiments." *Review of Educational Research*. 85 : 353–94.
- Muralidharan, K. 2012. "Long-term Effects of Teacher Performance Pay: Experimental Evidence from India." Society For Research On Educational Effectiveness.
- Muralidharan, K., and V. Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *Economic Journal* 120 : F187–203.
- . 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 : 39–77.
- . 2013 "Contract Teachers: Experimental Evidence from India". Working Paper 19440. National Bureau of Economic Research.

- . 2015. “The Aggregate Effect of School Choice: Evidence from a Two-stage Experiment in India.” *Quarterly Journal of Economics* 130 : 1,011–66.
- Murnane, R., and A. Ganimian. Forthcoming. “Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations.” *Review of Educational Research*.
- Petrosino, A., C. Morgan, T. A. Fronius, E. E. Tanner-Smith, and R. F. Boruch 2012. “Interventions in Developing Nations for Improving Primary and Secondary School Enrollment of Children: A Systematic Review.” *Campbell Systematic Reviews* 2012 : 19.
- Rao, G. 2014. “*Familiarity Does Not Breed Contempt: Diversity, Discrimination and Generosity in Delhi Schools*.” UC Berkeley Mimeo.
- Sarr, L. R., H. A. Dang, N. Chaudhary, D. Parajuli, and N. Asadullah. 2010. “*Reaching Out-of-School Children (ROSC) Project Evaluation Report*.” World Bank, Washington, DC.
- Shamsuddin, M. 2015. “Labour Market Effects of a Female Stipend Programme in Bangladesh.” *Oxford Development Studies* 43 : 425–47.
- Sterne, J., and R. Harbord. 2004. “Funnel Plots in Meta-Analysis.” *Stata Journal* 4 : 127–41.
- World Bank. 2012. “Bangladesh Education Quality Note.” *Draft background paper prepared for the Bangladesh Education Sector Review*. World Bank, South Asia Human Development Unit, Washington, DC.

Appendix: Critical Assessment of Available Evidence

A way to test for publication bias in the body of literature that evaluates the effectiveness of education-related interventions in South Asia is to rely on so-called funnel plots. Funnel plots are scatterplots where the treatment effects estimated from individual impact evaluations are plotted on the horizontal axis and the standard errors of these treatment effects are charted on the vertical axis. The underlying idea is that in the absence of biases results from studies with a small sample size will tend to be imprecise and will scatter widely at the bottom of the graph. In contrast, larger studies evaluating the same or similar interventions will produce point estimates that will hardly differ from each other and will also be very precisely measured.

Funnel plots for education-related interventions’ impacts on native language, mathematics, and composite test scores are visualized in Figures A1 to A3. Following the methodology introduced in the Methodology section, interventions are grouped according to whether they target the supply of or demand for education.

As noted by [Sterne and Harbord \(2004, 131\)](#), “Funnel plots were first proposed as a means of detecting a specific form of bias—publication bias.” Publication bias exists if among all studies that test the effectiveness of different interventions those that show significant positive effects are more likely to be published in scientific journals or similar outlets than those that find negative or statistically insignificant effects. This particular form of bias might, for instance, be caused by journal editors who are most interested in publishing results of startlingly effective interventions or by funding or implementing agencies that enjoy bragging about their

achievements but would rather not admit the failure of some of the interventions they sponsor. Publication bias is a severe threat to the validity of any meta-analysis because it can lead to upwardly biased estimates about the effectiveness of certain interventions. It can partly be mitigated by a thorough search strategy of the systematic review underlying a meta-analysis that incorporates not only studies in peer-reviewed scientific journals but also unpublished results or “gray” literature like working papers or preliminary reports. Such a search strategy is likely to help in identifying studies that rigorously document the impacts of various interventions but fail to be officially published because they do not demonstrate that a certain intervention works in improving outcomes. The search strategy that was the part of the systematic review underlying our meta-analysis follows exactly such guidelines and aims to be as comprehensive as possible.

Once a body of literature is identified through a systematic review, funnel plots can provide further evidence about whether publication bias is present. If this is the case, funnel plots will have asymmetrical appearances. More precisely, there will be a gap in the left bottom side of the graphs (the place where one would expect to see studies that find that interventions have low, statistically insignificant or even negative effects). As again noted by [Sterne and Harbord \(2004\)](#), the interpretation of funnel plots is further facilitated by the inclusion of vertical lines for a group of interventions’ treatment effects as estimated with the help of a meta-analysis and of diagonal lines representing the (pseudo) confidence intervals around these effects. Such diagonal lines—by convention drawn here for 95 percent confidence intervals—show the expected distribution of studies in the absence of publication biases.

It should be noted, however, that according to [Egger et al. \(1997\)](#) publication bias is not the only possible explanation for any small study effects such as asymmetries in funnel plots or data points lying outside the 95 percent confidence limits. First, distortions caused by the failure to include all relevant estimates and especially those based on small samples in a meta-analysis (“selection bias”) might be due to factors other than publication bias. For instance, [Jueni et al. \(2002\)](#) documented that studies reporting the results of clinical trials that are published in languages other than English tend to include relatively small samples and to be overlooked by many meta-analyses. In addition to different forms of selection bias, asymmetries in funnel plots might also be caused by the heterogeneous effects of individual interventions, data and methodological irregularities (which might be most severe in small studies), artifacts, or simply chance. The heterogeneous effects of individual interventions, in particular, might also be behind studies that fall far outside the 95 percent confidence intervals, as is arguably the case for one of the supply-centric interventions in each of Figures A1 to A3 (in fact, all three data points relate to the intervention providing remedial teaching by community volunteers and other support analyzed by [Lakshminarayan et al. 2013](#)) and some of the demand-centric interventions.

Figure A1. Funnel Plots for Impacts on Native Language Test Scores

Notes: Vertical lines denote point estimates from meta-analyses, and diagonal lines 95 percent pseudo-confidence intervals around these estimates. Asymmetrical appearances and gaps in the left bottom sides of graphs are indications of small sample effects.

Source: World Bank staff calculations based on studies listed in [Table 1](#).

Figure A2. Funnel Plots for Impacts on Math Test Scores

Notes: Vertical lines denote point estimates from meta-analyses, and diagonal lines 95 percent pseudo-confidence intervals around these estimates. Asymmetrical appearances and gaps in the left bottom sides of graphs are indications of small sample effects.

Source: World Bank staff calculations based on studies listed in [Table 1](#).

Nevertheless, and irrespective of the actual reasons behind any asymmetric funnel plots, figures A1 to A3 contain little evidence that publication bias is a grave worry in our context. For interventions targeting the demand for education, the evidence for impacts on native language, mathematics, or composite test scores is

Figure A3. Funnel Plots for Impacts on Composite Test Scores

Notes: Vertical lines denote point estimates from meta-analyses, and diagonal lines 95 percent pseudo confidence intervals around these estimates. Asymmetrical appearances and gaps in the left bottom sides of graphs are indications of small sample effects.

Source: World Bank staff calculations based on studies listed in [Table 1](#).

generally rather limited, but funnel plots appear very symmetric. For supply-centric programs, somewhat more evidence is available, and again the respective funnel plots do not indicate strong evidence of any publication biases.