What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews

David K. Evans and Anna Popova

Over the course of just two years, at least six reviews have examined interventions that seek to improve learning outcomes in developing countries. Although the reviews ostensibly have the same objective, they reach sometimes starkly different conclusions. The first objective of this paper is to identify why reviews diverge in their conclusions and how future reviews can be more effective. The second objective is to identify areas of overlap in the recommendations of existing reviews of what works to improve learning. This paper demonstrates that divergence in the recommendations of learning reviews is largely driven by differences in the samples of research incorporated in each review. Of 229 studies with student learning results, the most inclusive review incorporates less than half of the total studies. Across the reviews, two classes of programs are recommended with some consistency. Pedagogical interventions that tailor teaching to student learning levels—either teacher-led or facilitated by adaptive learning software—are effective at improving student test scores, as are individualized, repeated teacher training interventions often associated with a specific task or tool. Future reviews will be most useful if they combine narrative review with meta-analysis, conduct more exhaustive searches, and maintain low aggregation of intervention categories. Education, Impact Evaluation, Human Capital. JEL codes: O15, I21, I28, J13

Education quality remains an elusive goal in many developing countries. Although countries around the world have made great strides in increasing access to education, much of this education is still of low quality, with low learning outcomes reported in Africa, Latin America, and elsewhere (Bruns and Luque 2015; Filmer and Fox 2014; UNESCO 2014). Furthermore, evidence suggests—

The World Bank Research Observer

doi:10.1093/wbro/lkw004 Advance Access publication October 5, 2016 31:242–270

unsurprisingly—that additional years of schooling have little impact on economic growth in the absence of learning, which is a function of education quality (Hanushek and Wößmann 2007). At the same time that governments seek to increase the quality of education, the use of experimental and quasi-experimental methods to measure the effectiveness of education interventions in developing countries has become increasingly common. This has resulted in hundreds of studies from around the world demonstrating the effectiveness (or ineffectiveness) of various interventions at improving student learning. These interventions range from providing parents with information about the quality of schools to training teachers in scripted literacy instruction to dropping laptops off for students.

To make sense of all this evidence, various researchers have undertaken reviews of these impact evaluation studies.[1] In 2013 and 2014 alone, at least six reviews of studies seeking to improve student learning in primary schools in developing countries were published in journals or edited volumes or released as working papers. These include Conn (2014), Glewwe et al. (2014), Kremer, Brannen, and Glennerster (2013), Krishnaratne, White, and Carpenter (2013), McEwan (2015), and Murnane and Ganimian (2014a).[2] Between them, they review 300 studies from across the developing world: 229 of those studies report learning outcomes and 152 report enrollment or attendance outcomes. There are differences in the scope of the reviews: some focus only on primary education whereas others explore both primary and secondary, some only look at learning impacts whereas others also consider enrollment or attendance, one has a regional focus (Sub-Saharan Africa), two include only randomized controlled trials (RCTs), and three have a well-defined time frame. Yet the expected overlap is substantial: a feature common to all of these reviews is that they include RCTs implemented in Sub-Saharan Africa with learning outcomes at the primary school level published roughly between 1990 and 2010.

Despite that, the main results they highlighted for improving learning appear inconsistent. For example, using a subset of the conclusions for each review, Conn (2014) highlighted pedagogical interventions as the most effective, whereas McEwan (2015) found the largest effects for interventions involving computers and technology. Kremer, Brannen, and Glennerster (2013) highlighted pedagogical reforms that match teaching to student learning levels, as well as the incentives associated with hiring teachers on short-term contracts. Glewwe et al. (2014) emphasized the impact of teacher knowledge, teacher absenteeism, and the availability of student desks on student learning. Krishnaratne, White, and Carpenter (2013) underlined the importance of learning materials. And Murnane and Ganimian (2014a) emphasized providing information about school quality and returns to schooling, among other findings.

Given the massive array of evidence and the apparent divergence in conclusions from the reviews of the evidence, how is one to understand what actually works best to improve learning in developing countries? In this paper, we critically

examine recent reviews of how to improve primary learning outcomes in developing countries in order to understand the underlying reasons for the observed variation in conclusions and to provide recommendations for yielding more reliable inferences from such reviews. We also characterize the heterogeneity of effectiveness within categories of interventions. Finally, we highlight the common themes across the reviews—sometimes obscured by differences in categorization—in terms of what kinds of interventions are more and less effective.

We find that much of the variation in conclusions is driven by strikingly different compositions of studies across the reviews: of the 229 studies that look at learning outcomes, only three are included in all six reviews, whereas almost three-quarters (160) are included in only one or another of the reviews. Although some of these compositional differences are driven by explicit exclusion rules, many are not. While the main conclusions of every review are supported by evidence from papers that attempt to explicitly establish a counterfactual, each review incorporates different evidence, leading to different ultimate conclusions on what kinds of interventions are most effective relative to others. We also observe that much of the variation in outcomes across educational interventions is captured within categories of interventions rather than across them. For example, saying that computer interventions are most effective may be less useful and less accurate than saying that computer-assisted learning programs that are tailored to each student's level of knowledge, that are tied to the curriculum, and that provide teachers with training on how to integrate the technology into their instruction are most effective.

Finally, we find that there is indeed some intersection in recommendations across the reviews, although that intersection is masked by different labels. Even given the small degree of overlap in the composition of review samples, we find support across multiple reviews for pedagogical interventions that match teaching to students' learning—including through the use of computers or technology—and for individualized, sustained in-service teacher training. On the other hand, multiple reviews conclude that interventions that have consistently not led to a significant increase in student learning are cost-reducing interventions and health interventions.

Method

Inclusion Criteria

This paper takes as its population the set of reviews of impact evaluation evidence on improving student learning at primary levels in developing countries identified in 2013 and 2014. We restrict our analysis to reviews of evidence on how to improve learning, as opposed to increasing access (although many of the reviews also

include evidence on the latter). For the purposes of this paper, student learning is measured by test scores in math, language, science, or cognitive assessments, as determined by the inclusion criteria of the six reviews.

We also include only reviews that examine the effectiveness of improving learning at the primary level, although they need not exclusively examine the primary level. We include both published and unpublished reviews, but include only reviews of interventions in developing countries that were either published or posted online (in the case of unpublished work) in 2013 or 2014 in order to maximize the probability that the reviews draw on a similar underlying population of education studies. Reviews of learning interventions in developing countries continue to be written since this time window, including Asim et al. (2015), Masino and Niño-Zarazúa (2015), Glewwe and Muralidharan (2015), and Snilstveit et al. (2015).

Search Strategy

To identify reviews fulfilling the above criteria, we searched four meta-databases—Google Scholar, ERIC, The Campbell Library, and Cochrane Library—for articles posted in 2013 and 2014 containing the terms ["quality" OR "learning"] AND "education" AND ["review" OR "meta-analysis"] AND [developing countries OR low income countries OR "poor countries"]. This search yielded 16,865 results. In addition, we contacted experts in international education for recommendations, which yielded four results. We examined the first 1,069 of these combined results, which reflect the number of records reached after finding 500 consecutive irrelevant results. This review process led to the exclusion of 1,057 records that obviously did not meet the inclusion criteria. We went through the full texts of the remaining 12 papers to assess their eligibility and excluded six that were not eligible according to the conditions described above. This yielded the final six reviews under consideration: Conn (2014), Glewwe et al. (2014), Kremer, Brannen, and Glennerster (2013), Krishnaratne, White, and Carpenter (2013), McEwan (2015), and Murnane and Ganimian (2014a).

Analytical Strategy

In examining the eligible reviews, we address: (a) the characteristics of the reviews in terms of the methodologies used, their coverage, their primary conclusions, and how systematic they are; (b) what drives the differing conclusions of each review, considering the exclusion rules employed in selecting the studies included in each review, the variation in the composition and categorization of included studies for one key conclusion area from each review, and the heterogeneity across results within intervention categories; and (c) what the overlap in conclusions can tell us

about what does and does not work to improve student learning in developing countries.

In order to conduct this analysis, we extract and code data on two levels: first, on the level of the studies underlying the reviews and, second, on the level of the reviews themselves. In terms of the former, we compile a list of the underlying studies from the references of each review and then review their titles, abstracts, and, if necessary, full texts in order to extract and code the following indicators for each study: outcomes reported (learning, access, or both), year of publication, publication status (journal article, working paper, or report), education level of intervention (preschool, primary, secondary, tertiary, or vocational), country of intervention, region of intervention, and in which of the six reviews they are included.

At the level of the reviews, we recorded the following characteristics of each review: methodology, inclusion criteria, number of studies included, intervention categories reviewed, and most recommended intervention categories. We then tallied up the number of recommendations each intervention category has received across the six reviews and discussed those intervention categories recommended by the majority (i.e., at least four out of six) of the reviews to provide more detailed recommendations for education policy.

Results

The Reviews and the Studies Underlying the Reviews

Review methodologies. The six reviews discussed in this study include, fundamentally, three types of review. The first of these, meta-analysis, converts the results of all the included studies to standardized point estimates and then pools the estimates within a category of interventions (e.g., all the studies on providing school meals). Second, the narrative review examines the evidence qualitatively, usually discussing study by study, and then infers conclusions. Third, the vote-counting review shows the pattern of significant and insignificant positive and negative impacts across studies and draws inferences from that.

Each method has its advantages and disadvantages (Koricheva and Gurevitch 2013). Narrative reviews are often written by recognized experts in the field, who may have broad familiarity with the topic. These reviews provide the ability to reflect on nuances across studies and their underlying interventions and to draw conclusions from these. This is particularly valuable where there is variation in the effectiveness at improving student learning within a given intervention category, which there often is. Narrative reviews may also be more effective than other reviews at exploring the mechanisms behind the effectiveness of interventions using economic and education theory. However, these reviews rely on a subjective

weighting of the evidence by the reviewer, which may become less reliable as the number of studies reviewed increases. Also, because the weighting is qualitative, it may not be completely transparent to the reader, especially if not all reviewed studies are reported.

Vote counting has the appeal of simplicity, but it ignores sample size, statistical precision (except for significance cut-offs), and effect size, and so may overemphasize small, significant effects at the expense of large effects that narrowly miss a significance cut-off. Meta-analysis is more labor-intensive to implement, but because it aggregates results across studies into a single meta-result, it incorporates the data that vote counting excludes (e.g., effect size) while potentially increasing statistical power by pooling across smaller studies. A specific application of meta-analysis, called meta-regression, also permits controlling for the quality of studies or other moderating factors, as Conn (2014) and McEwan (2015) did in their reviews. At the same time, meta-analysis can mask heterogeneity; if a class of intervention has strong positive impacts in some cases and strong negative impacts in other cases, a meta-analysis may suggest a near-zero impact on average, which would be a mischaracterization of the true pattern of results. Furthermore, because meta-analysis requires pooling estimates across studies, studies that fail to report certain elements of the underlying data may be excluded, despite the studies being of high quality in other respects (e.g., internal validity). Meta-analyses also tend to use higher levels of aggregation (e.g., "pedagogical interventions") than narrative reviews, which can be less helpful if there is significant variation within the broad class of intervention.

Of the six reviews considered here, three are meta-analyses—Conn (2014), McEwan (2015), and Krishnaratne, White, and Carpenter (2013); two are narrative reviews—Kremer, Brannen, and Glennerster (2013) and Murnane and Ganimian (2014a); and one is a vote count—Glewwe et al. (2014). However, several of the reviews have elements that cross categories. The review by Kremer, Brannen, and Glennerster (2013), although a narrative review, presents standardized coefficients across many of the studies considered, albeit no average effect across studies. Krishnaratne, White, and Carpenter (2013) reported meta-analysis results in the appendix, but the article is written in the format of a narrative review. Conn (2014) presented detailed meta-analysis but also a detailed narrative discussion of individual studies.

Review coverage. The reviews vary extensively in the number of studies incorporated and the official inclusion criteria (table 1). The median number of learning studies reviewed is 61, with a minimum of 29 (Kremer, Brannen, and Glennerster 2013)[3] and a maximum of 96 (Murnane and Ganimian 2014a). The total number of learning studies across the six reviews is 229. These are drawn from across the world, with more than 20 studies in each of China, India, and Kenya. The total number of learning studies available has grown significantly over time, from 30

Table 1. Reviews and Their Composition

Review	Learning studies reviewed (total studies reviewed)	Inclusion criteria (in brief)
Conn (2014)	56 (56)	Any formal education level Learning outcomes RCT and quasi-experimental Sub-Saharan Africa 1980–2013
Glewwe et al. (2014)	67 (79)	Primary and secondary school Learning or access outcomes RCT and quasi-experimental Low- and middle-income countries 1990–2010
Kremer, Brannen, and Glennerster (2013)	29 (32)	Primary school Learning or access outcomes RCT only Low- and middle-income countries
Krishnaratne, White, and Carpenter (2013)	44 (76)	Primary and secondary school Learning or access outcomes RCT and quasi-experimental Low- and middle-income countries 1990–2009
McEwan (2015)	66 (66)	Primary school Learning outcomes RCT only Low- and middle-income countries
Murnane and Ganimian (2014a)	96 (132)	Primary and secondary school Learning or access outcomes RCT and natural experiments (no matching or fixed effects) Low- and middle-income countries
Total learning studies reviewed Total studies reviewed	229 300	

Notes: RCT stands for randomized controlled trial. Learning outcomes are scores in language or reading (in local language or English), mathematics, science, cognitive outcomes, or a composite assessment including any of these. Notably, learning outcomes do not include assessments of computer skills. Access outcomes include enrollment, attendance, and years of schooling. Note that we describe inclusion and not exclusion criteria; for example, where the inclusion criterion is access (learning) outcomes only, this means that only studies that have at least one access (learning) outcome are included in the review, although studies may include other outcomes in addition.

cumulative studies in 2000 to 33 studies coming out in 2013 alone. Taken together, this collection of studies likely reflects a close approximation of the total impact of evaluation evidence on learning in developing countries over the last 25 years.

Two reviews include only randomized controlled trials (Kremer, Brannen, and Glennerster 2013 and McEwan 2015). The others include RCTs as well as quasi-experimental methods, with slightly differing criteria for which methods qualify. Conn's (2014) is the only study with an explicit geographic focus. Two examine primary school only (Kremer, Brannen, and Glennerster 2013 and McEwan 2015), whereas the others include secondary school or other levels in addition to primary school. Only three impose an explicit criterion for study publication date, Glewwe et al. (2014) and Krishnaratne, White, and Carpenter (2013), both roughly 1990–2010, and Conn (2014), 1980–2013. All the reviews include RCTs, primary school learning outcomes, studies in Sub-Saharan Africa, and studies released between 1990 and 2010.

The learning studies included in the reviews fall broadly into three publication categories: published journal articles, unpublished working papers, and reports. Across the reviews, a slight majority of the learning studies included are journal articles (62 percent). Similarly, of the 13 studies cited in the majority of the reviews (i.e., at least four out of six), only four are working papers, while nine are journal articles. This suggests that there may be some degree of publication bias driving the studies included, but the proportion of published articles is not overwhelming and could merely reflect either reviewers' preferences for the inclusion of high-quality studies (with publication being one indicator) or the fact that published studies may be easier to locate.

Review recommendations. As they are reported in the reviews, the main conclusions recommend somewhat different categories of interventions. Conn (2014) found the best results for pedagogical interventions as well as for student incentives. (Conn's 2014 estimate for student incentives is based on only two studies, however, containing four treatment arms in total.) She also found positive results for extending the length of the school day, but only based on one study. Glewwe et al. (2014) found evidence that desks, chairs, and tables improve student learning, as well as teacher subject knowledge and teacher presence. Kremer, Brannen, and Glennerster (2013) identified pedagogical interventions to match teaching to students' learning, school accountability, and incentives as being highly effective. Krishnaratne, White, and Carpenter (2013) identified the provision of school materials as effective. McEwan (2015) identified several effective classes of interventions, including—in descending order of mean effect size—computers or instructional technology, teacher training, smaller classes, smaller learning groups within classes (or ability grouping), contract or volunteer teachers, student and teacher performance incentives, and instructional materials. Finally, Murnane and Ganimian (2014a) recommended providing information about school quality and returns to schooling, providing teacher incentives (in very low performance settings), and providing specific guidance for low-skilled teachers to help them reach minimally acceptable levels of instruction.

Are these systematic reviews? There are many reviews that strive to synthesize evidence on the effectiveness of policy interventions (or classes of interventions) across various fields. Reviews vary in how systematic they are in synthesizing the results of studies. Reviews are often systematic in some aspects of their methodology but not in others, making systematism more of a continuum than a binary notion. There is no single definition of a systematic review, but in considering how systematic each of the reviews we examine is, we turn to guidance from two main registries of systematic reviews, the Campbell Collaboration and Cochrane.

The Campbell Collaboration (2015) defines a systematic review as one that "uses transparent procedures to find, evaluate and synthesize the results of relevant research. Procedures are explicitly defined in advance, in order to ensure that the exercise is transparent and can be replicated." Campbell also describes screening studies for quality and peer review as important elements of systematic reviews. They provide four specific criteria that a review must have in order to be considered systematic: (a) clear inclusion/exclusion criteria; (b) an explicit search strategy; (c) systematic coding and analysis of included studies; and (d) meta-analysis (where possible). Cochrane provides less-specific guidance on what makes a review systematic, but its description of its own reviews is highly correlated with that of the Campbell Collaboration (Cochrane 2015).

We examine each of the six identified reviews of learning interventions in terms of the criteria defined by the Campbell Collaboration. We find that, although all six of the reviews considered here review the literature on what works to improve student learning in developing countries, they vary in how systematically they carry out different aspects of this task. The review by Kremer, Brannen, and Glennerster (2013) does not satisfy any of the criteria highlighted by the Campbell definition. All other reviews have clear inclusion/exclusion criteria, have a more or less explicit search strategy, and have coded and analyzed the studies they included with some systematism. As examples of the varying levels of systematism, Krishnaratne, White, and Carpenter (2013) briefly summarized an extensive search but provided no explicit details, and Murnane and Ganimian (2014a) provided an explicit list of sources and topics searched but not key words used. Meta-analysis was conducted by half of the reviews, although this requirement is less stringent as it is only required "where possible," and many of the studies included in these reviews do not report the necessary data for meta-analysis.

In addition, three out of the six reviews undertake a critical appraisal of the quality of the studies they include to give greater weight to more reliable studies, another important characteristic of systematic reviews highlighted by The Campbell Collaboration (2015). Kremer, Brannen, and Glennerster (2013), Krishnaratne, White, and Carpenter (2013), and Murnane and Ganimian (2014a) implicitly appraised the quality of studies by including only studies using defined econometric methods, but they did not explicitly discuss the quality of studies

Table 2. Inclusion of Learning Studies Across Reviews

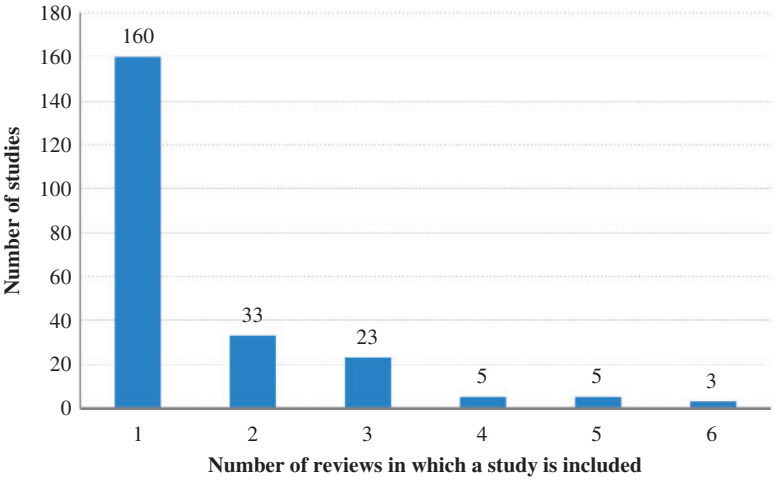
Inclusion of learning studies	Conn (2014)	Glewwe et al. (2014)	Kremer, Brannen, and Glennerster (2013)	Krishnaratne, White, and Carpenter (2013)	McEwan (2015)	Murnane and Ganimian (2014a)	Total
Panel 1: Studies with learning outcomes							
Number of studies in this review	56	67	29	44	66	96	229
As percentage of all studies with learning outcomes	24	29	13	19	29	42	
Panel 2: RCTs with learning outcomes							
Number of studies in this review	44	12	29	33	66	71	134
As percentage of all RCTs with learning outcomes	33	9	22	25	49	53	
Panel 3: RCTs with learning outcomes, primary level, 1990–2010							
Number of studies in this review	33	12	27	26	64	55	106
As percentage of all RCTs with LO, primary, 1990–2010	31	11	25	25	60	52	
Panel 4: RCTs with learning outcomes, primary level, 1990–2010, SSA							
Number of studies in this review	33	4	11	11	22	19	42
As percentage of all RCTs with LO, primary, 1990–2010, SSA	79	10	26	26	52	45	

Notes: LO stands for learning outcomes; RCT stands for randomized controlled trial; SSA stands for Sub-Saharan Africa. Studies are coded as SSA if they include learning outcomes for at least one country in Sub-Saharan Africa.

included in their reviews. In contrast, Conn (2014) and Glewwe et al. (2014) addressed this by reporting results separately for just RCTs, and Conn (2014) and McEwan (2015) controlled for various moderators of study quality, such as attrition and missing data, in meta-regressions.

We observe a spectrum in how systematic the reviews are, with most of the reviews we consider being completely or somewhat systematic fitting the majority of

Figure 1. Distribution of Learning Studies Across Reviews.
Note: The total number of learning studies is 229.



Campbell's criteria. Hence, even when these reviews do not self-identify as "systematic reviews," many of them are as systematic as any other.

What Drives Different Conclusions?

This section investigates how much of the difference in recommendations across the six reviews can be explained by differences in composition—that is, differences in the studies included in each review—and how much can be explained by reviews categorizing the same studies in different ways. It then examines how much of the variation in recommendations comes from differences across intervention categories as compared to within intervention categories.

Variation in composition and categorization. How much of the variation in conclusions is driven by the composition of the studies included, and how much is driven by differing categorization of similar studies? In terms of composition, the reviews include 229 learning studies between them, yet the most inclusive single review (Murnane and Ganimian 2014a) includes just over 40 percent of the total sample of papers. The least inclusive review (Kremer, Brannen, and Glennerster 2013) includes 13 percent of the total sample (table 2, panel 1).

The overlap across these reviews is surprisingly limited. Almost three-quarters of all the learning studies across the six reviews (160 studies) are included in only one of the six reviews. Only three studies (1 percent of the total) are included in all of the reviews (figure 1): a study of textbook provision (Glewwe et al. 2009), a

study of flipchart provision (Glewwe et al. 2004), and a study of student incentives (Kremer, Miguel, and Thornton 2009), all in Kenya.

The natural explanation for the difference in composition is the inclusion rules of the reviews. The most obvious candidate may be that one of the reviews, Conn (2014), only includes studies in Sub-Saharan Africa. However, if one looks at the studies that are included in all but one of the reviews, allowing for the possibility that many studies may be included in all reviews except Conn (2014), one finds only five studies (again, out of a total of 229).

In order to examine the role of other inclusion rules, we consider the intersection of studies that might be included by all the reviews according to their own inclusion rules. We begin with our full sample, which is all underlying studies with learning outcomes (table 2, panel 1), then, within those, we examine RCTs only (table 2, panel 2), then primary school and the 1990–2010 time frame only (table 2, panel 3), and finally implementation in Sub-Saharan Africa (table 2, panel 4).

One way that reviews control for the quality of studies is to include only RCTs, as two reviews do. The other four include RCTs and studies using quasi-experimental methods. However, even with randomized trials, the overlap in studies is limited (table 2, panel 2). Of 134 learning RCTs, over half (73 studies) are included in only one or another of the reviews. As with the wider collection of learning studies, only 13 studies are included in most (four, five, or six) of the reviews. The largely nonoverlapping collection of studies is apparently not driven by quality of studies either.

We next consider two additional areas of overlap in the inclusion criteria, studies that include primary-level outcomes and were published between 1990 and 2010 (table 2, panel 3). Of the 106 studies fulfilling all of the above requirements, only between 11 percent and 60 percent of studies are included in any single review.

Finally, we consider the overlapping inclusion criteria across all reviews: RCTs with learning outcomes at the primary school level published between 1990 and 2010 in Sub-Saharan Africa (table 2, panel 4). Of the 42 studies fulfilling all five of these requirements, still only between 10 percent and 79 percent of studies are included in any single review. This suggests that variation in composition is not remotely explained by the inclusion criteria of the reviews; if it were, then we would expect the coverage of studies—when considering only those fulfilling overlapping inclusion criteria—to be much closer to 100 percent for each review. Although there are differences across reviews in the proportion of studies that are published papers, there is no clear pattern between publication bias and coverage. This suggests that there is more behind variation in composition than systematic inclusion decisions.

At the same time, the reviews sometimes categorize studies in different ways. Many interventions fall into multiple categories, and studies tend not to provide

sufficient information for reviewers to apply a systematic rule for allocating interventions to categories. Thus these discrepancies are not necessarily due to error on the part of the reviewers; rather the allocation of interventions to categories is inherently subjective. In general, the review by Krishnaratne, White, and Carpenter (2013) tends to categorize studies that most other reviews put into some sort of "computer" category simply as "materials," those that others consider "teacher training" also as "materials," and "teacher incentives" simply as "additional teaching resources." Another notable difference in categorization is that of Conn's (2014) "Pedagogical interventions" and McEwan's (2015) "Computers or instructional technology," which are responsible for each review's strongest conclusion. Although the labels of these two groups are quite different, the samples overlap greatly because a significant subset of Conn's (2014) pedagogical interventions is computer-assisted learning programs.

These two examples illustrate that much of the difference in categorization across the reviews is explained by the various reviews either (a) opting for different levels of disaggregation in their analyses (e.g., pedagogy versus computer-based pedagogy) or (b) focusing on a different element of the intervention. McEwan (2015) is the only paper with explicitly overlapping categories. Beyond these examples, however, many of the reviews have categories that are easily recognizable as synonymous or at least widely overlapping. Thus, categorization—especially for Krishnaratne, White, and Carpenter (2013)—can be an additional driver of at least apparently divergent conclusions.

What is the role of composition and categorization in driving the different conclusions? We selected a primary conclusion from each review and then analyzed which studies drive that conclusion and whether those studies are included in the other reviews. For the five reviews for which we conducted this analysis, we selected the primary conclusions of each review by choosing: (a) for the meta-analyses, the category with the largest significant pooled effect size or most prominent result as defined by the review (for Krishnaratne, White, and Carpenter 2013, this is the category with the biggest significant effect when six or more studies are pooled together); and (b) for the other reviews, the first positive conclusion mentioned. (This analysis was not possible for Glewwe et al. 2014 because it does not identify which studies fall into which category.) The results of this analysis are summarized in table 3.

Considering Conn's (2014) finding that pedagogical interventions are the most effective, a tiny fraction of all of Conn's (2014) 17 pedagogical studies are incorporated in any other study (6 percent in three other reviews, none in Kremer, Brannen, and Glennerster 2013, and 18 percent in Murnane and Ganimian 2014a). The three studies with the largest effect sizes are not included in any other review. When considering Kremer, Brannen, and Glennerster's (2013) recommendation of pedagogical interventions that match teaching to students' learning,

Table 3. How Many of the Studies in One Review's Recommended Category of Intervention Are Included in Other Reviews?

Review – recommended intervention category	Percentage of studies included in review						
	n	Conn (2014)	Glewwe et al. (2014)	Kremer, Brannen, and Glennerster (2013)	Krishnaratne, White, and Carpenter (2013)	McEwan (2015)	Murnane and Ganimian (2014a)
Conn (2014) – pedagogical interventions	17	–-	6	0	6	6	18
Kremer, Brannen, and Glennerster (2013) – matching teaching to students' learning	2	50	50	–-	50	100	50
Krishnaratne, White, and Carpenter (2013) – materials provision	6	17	67	50	–-	100	83
McEwan (2015) – computers or instructional technology	10	0	30	30	40	–-	70
Murnane and Ganimian (2014a) – information provision	9	11	0	11	33	33	–-

Note: n is the total number of studies included in a given review's recommended category of intervention, including both those that report positive effects and those that report negative effects.

there is more but still limited coverage: one of the two studies driving this conclusion is in four of the other five reviews, whereas the other study is in three of the other five. Coverage in other reviews is also low for the studies driving the findings in McEwan (2015) and Murnane and Ganimian (2014a). For Krishnaratne, White, and Carpenter's (2013) finding supporting "materials provision," the three studies that seem to be driving this result—Banerjee et al. (2007), He, Linden, and MacLeod (2008), and Lai et al. (2012)—are included in some other reviews (one of the studies is in four other reviews, whereas the other two are in just one or two), but most other reviews categorized those three studies as computer-assisted learning. In that case, categorization may be driving some of the result. We repeated this analysis for RCTs only (results not shown) and found that the composition analysis is almost identical to that which includes all studies, suggesting that the main conclusions of each review are driven by evidence from RCTs.

Thus, differences in composition seem much more likely to drive variation in conclusions than differences in categorization, although categorization also plays a role. No review included even half of the total sample of studies. As a result, it may be unwise to rely on a single review to derive a conclusion about the most effective interventions to improve student learning, but each review relied on clear empirical evidence to determine what works well in some settings. So these reviews may

be more effective at providing ideas for what works well to improve learning rather than definitively characterizing what works best.

Variation within intervention categories. As some of the reviews highlighted, much of the variation in learning results across studies is driven by variation within categories. Just because a given intervention falls into a category that is effective at improving student learning on average, this does not mean that it will perform per the mean of that category; specific details of the intervention determine its effectiveness. When Conn (2014) concluded that pedagogical interventions are most effective or when McEwan (2015) concluded that computer interventions are most effective, these conclusions can mask the massive heterogeneity within the category. Both reviews discuss this. It is important to note that many pedagogical interventions have been ineffective, as have many computer interventions.

For example, although McEwan (2015) found computer-based interventions to be by far the most effective category, the One Laptop Per Child (OLPC) program in Peru had little or even negative effects on student learning, apparently because it distributed computers without any additional training (Cristia et al. 2012). Even within the subcategory of OLPC programs there is great heterogeneity; a recent program that distributed laptops installed with remedial tutoring software to migrant children in Beijing and trained them in their use produced large increases in standardized math scores (Mo et al. 2012). Significant heterogeneity may even exist across estimates within a single study. For example, providing English textbooks in rural Kenya was found to improve test scores among the students who performed best on a pre-intervention exam, while having no significant impact on other students (Glewwe, Kremer, and Moulin 2009). Similar heterogeneity also exists within low-performing intervention categories. Conn (2014) found interventions providing school supplies to have a low average effect (0.02 standard deviations), for example, yet unanticipated school grants for textbooks in Zambia (Das et al. 2013) are roughly five times more effective than the mean of this category. It is crucial to examine not just which categories of interventions are most effective, but rather which specific interventions have been effective within that category and the characteristics of those interventions.

What Works to Improve Student Learning?

Despite differing conclusions from each review, is there any intersection in what works? At first glance, there is no convenient overlap in the categories of interventions deemed most effective. But upon closer analysis, despite the differing samples and some degree of different characterization, there is some agreement. In this analysis, we group interventions using the lowest possible level of aggregation so

as to highlight the specific elements driving the relative effectiveness or ineffectiveness of certain types of programs. For example, we consider "teacher incentives" or "student incentives," rather than the aggregate category of "incentives." We then tally up the number of recommendations each disaggregate intervention category has received across the six reviews (see table 4). We interpret those categories recommended by the majority of the reviews (i.e., at least four out of six) as representing the intersection. Across the six reviews, we find that only two intervention categories fulfill this condition: pedagogical interventions that match teaching to students' learning and individualized, repeated teacher training associated with a specific method or task. In the subsequent discussion, we use the studies with positive effects that fall into these two recommended categories to derive recommendations for education policy.

Pedagogical interventions that match teaching to students' learning. As a category, pedagogical interventions that match teaching to students' learning, including through the use of computers or technology, is recommended by four of the six reviews (table 4). Among the meta-analyses—which calculate average pooled effects by category—it is the category that most commonly produces the largest quantitative impacts on student learning. This comes out particularly strongly in Conn (2014), McEwan (2015), and Kremer, Brannen, and Glennerster (2013). Each of these reviews gives this category a slightly different name ("Pedagogical interventions," "Computers or instructional technology," and "Pedagogical interventions to match teaching to students' learning," respectively) but essentially refers to the same group of driving interventions.

Conn (2014) found that, across her sample of African studies, pedagogical interventions (which she defined as those that change instructional techniques) are more effective at improving student learning than all other types of interventions combined. Within high-quality studies of pedagogical interventions, she found that those interventions that employ adaptive instruction and teacher coaching techniques are particularly effective. Among these interventions, the pooled effect size associated with adaptive instruction is 0.42 standard deviations, whereas that of programs with nonadaptive instruction is about one-quarter of that, at only 0.12 standard deviations.[4] All three studies in Conn's (2014) sample, which evaluate adaptive instruction interventions, reported positive, statistically significant effects on student literacy scores (Korsah et al. 2010; Piper and Korda 2011; Spratt et al. 2013).

Programs with adaptive instruction fall into two categories: (a) computer-assisted learning (CAL) programs that adapt to the student's learning level or (b) teacher-led methods that emphasize formative assessment and individualized and targeted instruction. Although Conn (2014) found both computer-assisted and teacher-led methods to produce a significant improvement in student performance (at the 10 percent level), the effect of the former is twice as large as the latter. One

Table 4. Cumulative Positive Support for Intervention Categories across the Reviews

	Conn (2014)	Glewwe et al. (2014)	Kremer, Brannen, and Glennerster (2013)	Krishnaratne, White, and Carpenter (2013)	McEwan (2015)	Murnane and Ganimian (2014a)	Tally
Pedagogical interventions that match teaching to students' learning	Pedagogical interventions/ teacher training	Teacher subject knowledge	Pedagogical interventions to match teaching to students' learning		Computers or instructional technology		IV
Individualized teacher training	Pedagogical interventions/ teacher training	Teacher subject knowledge			Teacher training	Specific guidance for low-skilled teachers to reach minimally acceptable levels of instruction	IV
Teacher incentives			Incentives		Student and teacher performance incentives	Teacher incentives in very low-performance settings	III
Materials		Desks, tables, and chairs		Materials	Instructional materials		III
Student incentives	Student incentives				Student and teacher performance incentives		II
Accountability			Accountability				I
Contract or volunteer teachers					Contract or volunteer teachers		I
Providing information about school quality and returns to schooling						Providing information about school quality and returns to schooling	I
Smaller classes, smaller learning groups within classes, or ability grouping					Smaller classes, smaller learning groups within classes, or ability grouping		I
Teacher presence		Teacher presence					I

example of teacher-led adaptive instruction is the Early Grade Reading Assessment program in Liberia, evaluated by Piper and Korda (2011), in which students' reading levels were evaluated using a diagnostic exam and teachers were then trained in how to continually assess student progress.

Another example, categorized differently by Conn (2014) but argued to help teachers adapt instruction in Kremer, Brannen, and Glennerster (2013) and included in four of the six reviews, assigned students in Kenya to classes on the basis of initial preparedness so that teachers could focus instruction at the level of learning of the students (Kremer, Duflo, and Dupas 2011). This increased test scores at all levels of initial preparedness (by 0.17 standard deviations in language and 0.16 standard deviations in math). Even for low-performing students, who might stand the most to gain from being integrated into classes with high-performing students, ability grouping improved student performance by 0.16 standard deviations, with results carrying over into the next school year after the program had stopped. An RCT that is too recent to be included in any of the reviews underlines the effectiveness of formative assessment linked with targeted instruction. Giving students in India a brief assessment of basic language skills at the start of the academic year and then setting aside a portion of the school day to teach students in groups according to ability level, independent of grade or age, improved both oral and written language test scores, by 0.15 standard deviations and 0.14 standard deviations, respectively (Duflo et al. 2015).

Along the same lines, McEwan (2015) found computer-assisted learning programs to have a greater impact than other kinds of interventions, with a mean effect size of 0.15 (significant with 99 percent confidence). A successful example included in McEwan (2015) but also highlighted by Kremer, Brannen, and Glennerster (2013) is a CAL program in India, which—using math software that allowed children to learn at their own pace—increased math scores by 0.48 standard deviations, significant with 99 percent confidence (Banerjee et al. 2007). Moreover, the latter program was extremely cost-effective, producing an increase of 3.01 standard deviations in test scores per $100 spent (Kremer, Brannen, and Glennerster 2013).

However, as Murnane and Ganimian (2014a) highlighted, such programs do not improve student achievement unless they change children's daily experiences at school. Computer-assisted learning programs are ineffective when instruction is not tailored to each student's level of knowledge, when technology distribution is unaccompanied by parent or student training as was the case in Peru's One Laptop Per Child program (Cristia et al. 2012), when computers substitute away from useful instructional time during school hours (He, Linden, and MacLeod 2008) or home study (Malamud and Pop-Eleches 2011), or when the treatment is not tied to the curriculum or integrated by teachers into their classroom instruction (Barrera-Osorio and Linden 2009). Here effectiveness is defined in terms of

improving student test scores in math and language. Several of these programs were found to improve children's computing skills, but without improvements in school achievement. Moreover, although these programs may improve computing skills for the specific computers or laptops provided, evidence from Peru suggests that this may not transfer to an improvement in more general computing skills (Beuermann et al. 2013; Murnane and Ganimian 2014a).

Taken together, there is significant overlap in these recommendations: Computer-assisted learning or teacher-led interventions that individualize instruction can be highly effective. But pedagogical interventions or computing interventions generally are not inherently more effective than others; they have to be well implemented and affect students' learning experience.

Individualized, repeated teacher training associated with a specific method or task. The other category of interventions recommended by a majority of the reviews (also four of the six, as in table 4) is teacher training. This intervention type is found to produce the second largest effects in two of the meta-analyses. McEwan (2015) found teacher training to produce a 0.12 standard deviations improvement in learning (significant with 99 percent confidence), for example. (McEwan 2015 and Conn 2014 may not have precisely comparable standardized estimates because they control for different moderators in their regressions.) Again, examining the specific programs is crucial: providing teachers with general guidance tends not to improve student learning, but Murnane and Ganimian (2014a) found that detailed support tailored to the skill levels of teachers can be effective. For example, an Indian program giving teachers diagnostic information about student performance with general tips on how to help them improve had little impact on student learning (Muralidharan and Sundararaman 2010). In contrast, training that provides detailed guidance on what and how teachers should teach has proven to be effective in enhancing the skills of low-performing students (Murnane and Ganimian 2014a). For example, a scripted literacy program in Mumbai that provided storybooks, flashcards, and a child library, as well as instructions for teachers specifying the activities in which these should be used and when, had positive effects on child literacy (He, Linden, and MacLeod 2009).

This highlights the fact that the large improvements in student learning produced by appropriate teacher training may be in part driven by a large degree of overlap with other interventions because many of the successful instructional interventions were coupled with teacher training in how to employ the new method in the classroom (McEwan 2015). For example, a related intervention providing flashcards to teach children English in India improved test scores by much more when it was implemented through a teacher training program than when it was introduced externally without preparing teachers (He, Linden, and MacLeod 2008). Moreover, with regards to variation within the category of teacher training, one-time in-service trainings at a central location, typical of many teacher

Table 5. Cumulative Negative Support for Intervention Categories across the Reviews

	Conn (2014)	Glewwe et al. (2014)	Kremer, Brannen, and Glennerster (2013)	Krishnaratne, White, and Carpenter (2013)	McEwan (2015)	Murnane and Ganimian (2014a)	Tally
Cost-reducing interventions			Cost-reducing interventions		Monetary grants	Reducing the costs of going to school	III
Health interventions	Health interventions		Health interventions		Deworming treatments		III
Alternatives to traditional public schools						Alternatives to traditional public schools	I
Information interventions			Information interventions				I
Resources (unless they change children's daily experiences at school)						Resources (unless they change children's daily experiences at school)	I

training interventions, are not among those found to be highly effective. However, Conn (2014) found pedagogical interventions involving long-term teacher mentoring or in-school teacher coaching to produce a sizeable (albeit not always significant) effect on student learning, at 0.25 standard deviations.[5] An example is the "Read, Educate, and Develop" (or READ) program in rural South Africa evaluated by Sailors et al. (2010), which provides students with high-quality books relevant to their lives and teachers with training on strategies to integrate these books into their lesson plans. This training includes demonstration lessons by READ mentors, monthly coaching and monitoring visits followed by one-on-one reflection sessions, and after-school workshops for both teachers and school administrators. The program had highly significant impacts on a range of reading measures, albeit with a quasi-experimental design. Overall, of the evaluations of programs with ongoing teacher training elements that Conn (2014) reviewed, all four showed statistically significant improvements in student literacy (Brooker et al. 2013; Lucas et al. 2014; Sailors et al. 2010; Spratt et al. 2013), as well as numeracy when it was tested (Lucas et al. 2014).

Other examples of interventions combining instructional methods with teacher training include a combination of student reading groups and in-school supervisors to provide guidance to group leaders in Chile (Cabezas, Cuesta, and Gallego

2011); a remedial education program in India that gives local contract teachers two weeks of initial training followed by reinforcement throughout the school year (Banerjee et al. 2007); a program targeting early reading skills in Mali that offers lesson plans and accompanying instruction materials, together with training, support visits, and grading of teacher guides and student workbooks (Friedman, Gerard, and Ralaingita 2010); and an early grade reading instruction program in Kenya and Uganda that provides schools with materials and trains teachers in the use of the instructional method (local language materials) and in learning assessment, as well as providing them with regular mentoring (Lucas et al. 2014). The success of such programs is consistent with the findings of a more recent systematic review, the results of which show that structured pedagogy programs—which typically combine the development of new content, materials, and training for teachers in delivering the content—have the largest positive average effects on student learning (Snilstveit et al. 2015).

Glewwe et al.'s (2014) finding that teachers' knowledge of the subjects they teach increases student learning also implicitly supports teacher training interventions that effectively boost such knowledge. Kremer, Brannen, and Glennerster (2013) and Krishnaratne, White, and Carpenter (2013) had less to say about teacher training. This is explained in part by composition and in part by categorization. Some of the studies driving the large (and significant) positive effect for teacher training interventions in McEwan's (2015) sample appear in only one or two of the other reviews, and, in the case of an early reading program in Mali (Friedman, Gerard, and Ralaingita 2010), in none of the others. Furthermore, Krishnaratne, White, and Carpenter (2013) reviewed a number of training interventions, but they have no specific category for teacher training and instead code all interventions that have training along with pedagogical materials (e.g., guides) under the broad umbrella of materials provision.

What Has Not Led to Measured Improvements in Student Learning?

We also observe overlap in intervention areas where the evidence is weaker. Tallying up the lack of support for each disaggregated intervention category across the six reviews in the same way as we did the positive recommendations, we find that cost-reducing interventions (such as fee reductions) and health interventions (such as nutritional supplements) are those least commonly found to be effective at improving student learning (table 5). While there is not definitive, precise evidence that interventions across these categories do not improve learning, they have not produced significant effects on student learning across multiple studies. Importantly, there is substantial evidence that these interventions can effectively increase school enrollment and attendance but not reading and math scores. An

education improvement program may couple these kinds of programs to boost access with those programs proven to improve learning.

The conclusion on health interventions is in part driven by the definition of learning as test scores in language and math in some of the reviews; Conn (2014) found that health interventions do significantly improve students' attention and memory. Ozier (2014)—not included in any of the reviews—found that a large-scale deworming intervention in Kenya significantly increased reasoning test scores among the younger siblings of program participants 10 years after implementation. However, if children are more attentive to or better at remembering material that is poorly taught or poorly targeted to their learning level, the cognitive improvements may not translate into academic learning gains. Thus, if the goal is to improve student test scores, these programs may be less likely to be effective.

Discussion

This paper demonstrates that even reviews that are relatively systematic in fact fall far short of exhaustive coverage and, as a result, reach varying conclusions. Authors also make judgments as to how to characterize the studies they include, which further drives differing conclusions. The least systematic form of analysis, the narrative review, can incorporate the largest number of studies but requires nonscientific tallying and weighting across studies and is the most susceptible to influence by authors' prior beliefs. The most systematic form of analysis, the meta-analysis, may limit the included studies because of stringent requirements on the data reported in order to compute strictly comparable effect sizes, and it may fail to illuminate the mechanisms behind the most effective interventions. Each method has flaws that keep it from being both systematic and exhaustive.

Nonetheless, these reviews—when analyzed together—can effectively identify interventions that work well, even if they cannot convincingly identify what works best. Taking the reviews together, we conclude with some confidence that pedagogical interventions that match teaching to students' learning and individualized, repeated teacher training associated with a specific method or task are effective at improving student learning; both of these recommendations are reported in some form across a majority of the reviews.

Even intervention types that are recommended by only a minority of reviews may be effective. Each recommendation in each review is based on studies demonstrating positive, significant impact. For example, McEwan (2015) estimated a mean effect of performance incentives of 0.09 (significant with 95 percent confidence), driven mostly by teacher incentives, although the effectiveness of improving such incentives varies greatly across studies (Kremer, Brannen, and

Glennerster 2013). Despite a lower effect size, providing information on the returns to schooling in Madagascar (Nguyen 2009) is one of the most cost-effective education interventions that has been evaluated using an RCT (Kremer, Brannen, and Glennerster 2013). These interventions may be a good investment in some school systems and they certainly merit further study, but, given their limited coverage across reviews, it would be difficult to claim conclusively that they are the very best investments.

A further limitation of these reviews extends from a limitation of most underlying studies: the reviews focus on effectiveness but say less about the cost-effectiveness of various intervention types due to the fact that most of the studies they review do not report sufficiently detailed and comparable cost data (Evans and Popova 2016; McEwan 2015). Varying costs can lead certain interventions that have lower benefits to have a much higher benefit-per-dollar than others, and policy makers make investment decisions based on costs as well as impacts. Kremer, Brannen, and Glennerster (2013) did provide cost-effectiveness results for a subsample of 18 studies. They found pedagogical interventions that match teaching to students' learning levels, contract teachers, and the provision of earnings information to be the most cost-effective. Informing the expensive end of the spectrum, McEwan (2015) combined his effect sizes with cost estimates from Kremer, Brannen, and Glennerster (2013) to find that interventions focusing on computer-assisted learning and class size reduction may be less cost-effective than others. However, these are based on a small sample (less than 10 percent) of the 229 learning studies included in this review; much additional work is needed.

A third limitation—again, extending from the underlying studies—is that these reviews focus largely on short-term learning impacts. For example, McEwan (2015) highlighted that for his sample of studies, the average follow-up is conducted after nine to 13 months of program exposure, with only about 10 percent of follow-ups occurring even one month after the conclusion of the intervention. Across low- and high-income countries, it has been observed that educational gains are sometimes not sustained over time (Andrabi et al. 2011; Evans, Kremer, and Ngatia 2014; Jacob, Lefgren, and Sims 2010). On the other hand, impacts may take longer to manifest in some cases, as with interventions in which teachers or schools receive continued support (such as annual grants or regular training) but where it takes time to see the return on those investments in terms of learning (King and Behrman 2009). Thus, a clear shortcoming of this literature is its inability to inform the trajectory of longer-term learning impacts.

Finally, many of the studies underlying these reviews are evaluations of smaller-scale interventions implemented by nongovernment organizations (NGOs) or researchers. Illustratively, analyzing the sample of all evaluations of teacher training interventions included across the six reviews, we find that only three of the 20 programs were implemented by governments, while nine were implemented by NGOs

and eight by researchers directly. Of the three government-implemented programs, one was effective at improving learning. Although this indicates that a government-implemented program can be successful, it is not a strong record of success. As such, their findings inform us imperfectly about the kinds of interventions that governments have the institutional capacity necessary to implement at scale to improve student learning. Evidence from an evaluation of a teacher training program in Uganda—too recent to be included in any of the reviews—highlights this challenge. The program, which was highly effective at increasing student learning when the training was implemented by a social enterprise, saw its positive impacts dissipate under a lower-cost version, which used government workers to implement the training (Kerwin and Thornton 2015).

Our review of the individual reviews faces its own limitations. It demonstrates the issues faced when conducting a systematic review but only in a specific sample. While these same issues are likely to apply in other areas, they are not an exhaustive list. An examination of systematic reviews in other areas (outside of education in developing countries), or a similar analysis of reviews that examine a single intervention or class of intervention, would be instructive. However, education provides a unique opportunity to use multiple reviews to highlight where they might fall short, as a high number of reviews have been produced, all with ostensibly the same goal, over a short time period.

Future reviews will benefit from combining methodologies, for example, performing meta-analysis (which allows a highly systematic analysis) accompanied by narrative review (which can explore heterogeneity within categories and the apparent mechanisms behind effective programs). Furthermore, using narrative review will allow the inclusion of studies that are excluded from meta-analyses. Given the high observed level of heterogeneity within classes of interventions, the most useful reviews are likely to use low levels of aggregation, identifying specific characteristics of interventions that are effective rather than broad classes of interventions. Future reviews will also be most useful if they are careful to search out unpublished studies: Less than two-thirds of studies included in the six reviews were published journal articles. In the context of learning reviews specifically, future research could apply these recommendations to the full pool of learning studies identified across the reviews, as well as any new learning studies.

Taken together, these reviews do identify certain key messages about improving learning in developing countries: Both student learning interventions and teacher training interventions will be most effective when tailored to the student or teacher involved. Pedagogical interventions must change students' learning experiences and be adapted to individual student learning levels. Teacher training may be most effective when it is repeated and linked to a specific pedagogical method or tool. Beyond these findings, with the quantity of education research being produced, synthesizing it in a way that is consistent across reviews will be crucial to

identifying those future interventions most likely to benefit students around the world.

Notes

David K. Evans is a Senior Economist in the Office of the Chief Economist of the World Bank's Africa Region; email: devans2@worldbank.org. Anna Popova is a Consultant at the World Bank. This work was supported by the World Bank. The authors are grateful for comments from Jacobus Cilliers, Katharine Conn, Deon Filmer, Alejandro Ganimian, Peter Holland, Howard White, Jeffery Tanner, and Viticia Thames; conference participants at the Comparative and International Education Society and the Center for the Study of African Economies; for background materials provided by Katharine Conn, Alejandro Ganimian, and Paul Glewwe; to three anonymous reviewers; and to the editor of the World Bank Research Observer.

1. These build on a previous generation of reviews also seeking to analyze the determinants of student learning in primary schools in developing countries. See, for examples, Hanushek (1995) and Kremer (1995).

2. The review by Murnane and Ganimian was published in July 2014 as a National Bureau of Economic Research working paper (Murnane and Ganimian 2014b). For this study, we draw on an updated, unpublished version provided by the authors, dated November 18, 2014. Although the sample of studies varies across the two versions, the conclusions are exactly the same. The paper has subsequently been revised and is forthcoming (Murnane and Ganimian 2016).

3. We arrive at Kremer, Brannen, and Glennerster's (2013) sample of 29 studies by including all those studies for which they provide a point estimate of the evaluated program's impact on test scores (18 studies), as well as those whose impacts (positive or negative) are explicitly discussed in the text.

4. The samples are small (three studies in adaptive instruction and five studies in nonadaptive instruction), so Conn (2014) did not report *p* values.

5. As Conn (2014) reports, with four studies, the sample size does not allow estimation of a reliable *p* value. But as suggestive evidence, the coefficient divided by the standard error yields a t-statistic of 1.87, which is normally considered significant with between 90% and 95% confidence.

References

Abdu-Raheem, B. O. 2011. "Effects of Problem-Solving Method on Secondary School Students' Achievement and Retention in Social Studies in Ekiti State, Nigeria." *Journal of International Education Research* 8 (1) : 19–26.

Ahn, S., A. J. Ames, and N. D. Myers. 2012. "A Review of Meta-Analyses in Education: Methodological Strengths and Weaknesses." *Review of Educational Research* 82 (4) : 436–76.

Ajaja, O. P., and O. U. Eravwoke. 2011. "Effects of Cooperative Learning Strategy on Junior Secondary School Students' Achievement in Integrated Science." *Electronic Journal of Science Education* 14 (1) : 1–18.

Andrabi, T., J. Das, A. I. Khwaja, and T. Zajonc. 2011. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics* 3 (3) : 29–54.

Angrist, J., and V. Lavy. 2001. "Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools." *Journal of Labor Economics* 19 : 343–69.

Asim, S., R. S. Chase, A. Dar, and A. D. Schmillen. 2015. "Improving Education Outcomes in South Asia: Findings from a Decade of Impact Evaluations." Policy Research Working Paper 7362. World Bank, Policy Research Department, Washington, DC.

Banerjee, A., S. Cole, E. Duflo, and L. L. Linden. 2007. "Remedying education: Evidence from two randomized experiments in India." *The Quarterly Journal of Economics* 122 : 1235–64.

Barrera-Osorio, F., and L. L. Linden. 2009. "The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia." Policy Research Working Paper 4836. World Bank, Policy Research Department, Washington, DC.

Beuermann, D. W., J. P. Cristia, Y. Cruz-Agayo, S. Cueto, and O. Malamud. 2013. "Home Computers and Child Outcomes: Short-Term Impacts from a Randomized Experiment in Peru." Working Paper 18818. National Bureau of Economic Research, Cambridge, MA.

Brooker, S., H. Inyega, B. Estambale, K. Njagi, E. Juma, C. Jones, C. Goodman, and M. Jukes. 2013. "Impact of Malaria Control and Enhanced Literacy Instruction on Educational Outcomes among Kenyan School Children: A Multi-Sectoral, Prospective, Randomized Evaluation." 3ie Draft Grantee Final Report. International Initiative for Impact Evaluation, Washington, DC.

Bruns, B., and J. Luque. 2015. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, DC: World Bank.

Cabezas V., J. I. Cuesta, and F. A. Gallego. 2011. "Effects of Short-Term Tutoring on Cognitive and Non-Cognitive Skills: Evidence from a Randomized Evaluation in Chile." Unpublished manuscript. Pontificia Universidad Católica de Chile, Santiago.

Cochrane. 2015. "What Is Cochrane Evidence and How Can It Help You?" Accessed March 30, 2015. http://www.cochrane.org/what-is-cochrane-evidence.

Conn, K. 2014. *"Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Rigorous Impact Evaluations."* PhD diss. Columbia University, New York.

Cristia, J., P. Ibarrarán, S. Cueto, A. Santiago, and E. Severín. 2012. "Technology and Child Development: Evidence from the One Laptop Per Child Program." Discussion Paper 6401. Institute for the Study of Labor, Bonn, Germany.

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(2): 29–57.

Duflo, E., J. Berry, S. Mukerji, and M. Shotland. 2015. "A Wide Angle View of Learning Evaluation of the CCE and LEP Programmes in Haryana, India." 3ie Impact Evaluation Report 22. International Initiative for Impact Evaluation, Washington, DC.

Evans, D. K., M. Kremer, and M. Ngatia. 2014. "Schooling Costs, School Participation, and Long-Run Outcomes: Evidence from Kenya." Unpublished working paper. World Bank, Washington, DC.

Evans, D. K., and A. Popova. 2014 "Cost-Effectiveness Analysis in Development: Accounting for Local Costs and Noisy Impacts." *World Development*, 77: 262-276.

Filmer, D., and L. Fox. 2014. *Youth employment in Sub-Saharan Africa (Africa Development Series)*. Washington, DC: World Bank.

Friedman, W., F. Gerard, and W. Ralaingita. 2010. "International Independent Evaluation of the Effectiveness of Institut Pour l'Education Populaire's 'Read - Learn – Lead' (RLL) Program in Mali." Mid-Term Report. RTI International, Durham, NC.

Gee, K. 2010. "The Impact of School-Based Anti-Malarial Treatment on Adolescents' Cognition: Evidence from a Cluster-Randomized Intervention in Kenya." PhD diss. Harvard University, Boston, MA.

Glewwe, P., E. A. Hanushek, S. D. Humpage, and R. Ravina. 2014. "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*. Chicago and London: University of Chicago Press.

Glewwe, P., M. Kremer, and S. Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1 (1): 112–35.

Glewwe, P., M. Kremer, and S. Moulin, and E. Zitzewitz. 2004. Retrospective Vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya. *Journal of Development Economics* 74: 251–68.

Glewwe, P., and K. Muralidharan. 2015. "Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." Working Paper RISE-WP-15/001. RISE. University of Oxford, Research on Improving Systems of Education (RISE), Oxford.

Hanushek, E. A. 1995. Interpreting Recent Research On Schooling in Developing Countries. *The World Bank Research Observer* 10: 227–46.

Hanushek, E. A., and L. Wößmann. 2007. "The Role of School Improvement in Economic Development." Working Paper 12832. National Bureau of Economic Research, Cambridge, MA.

He, F., L. L. Linden, and M. Macleod. 2008. "How to Teach English in India: Testing the Relative Productivity of Instruction Methods with Pratham English Language Education Program." Unpublished manuscript. Columbia University, New York.

Jacob, B. A., L. Lefgren, and D. P. Sims. 2010. "The Persistence of Teacher-Induced Learning." *Journal of Human Resources* 45: 915–43.

Kerwin, J. T., and R. Thornton. 2015. "Making the Grade: Understanding What Works for Teaching Literacy in Rural Uganda." Unpublished manuscript. University of Illinois, Urbana, IL.

King, E. M., and J. R. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *The World Bank Research Observer* 24 (1): 55–82.

Koricheva, J., and J. Gurevitch. 2013. "Place of Meta-Analysis among Other Methods of Research Synthesis." In J. Koricheva, J. Gurevitch, and K. Mengersen, eds., *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton, NJ: Princeton University Press.

Korsah, G. A., J. Mostow, M. B. Dias, T. M. Sweet, S. M. Belousov, M. F. Dias, and H. Gong. 2010. "Improving Child Literacy in Africa: Experiments with an Automated Reading Tutor." *Information Technologies and International Development* 6 (2): 1–19.

Kremer, M. R. 1995. Research On Schooling: What We Know and What We Don't: A Comment on Hanushek. *The 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: 297–300.

Kremer, M., E. Duflo, and P. Dupas. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking." *American Economic Review* 101: 1739–74.

Kremer, M., E. Miguel, and R. Thornton. 2009. "Incentives To Learn." *The Review of Economics and Statistics* 91: 437–56.

Krishnaratne, S., H. White, and E. Carpenter. (2013). "Quality Education For All Children? What Works in Education in Developing Countries." 3ie Working Paper 20. International Initiative For Impact Evaluation, Washington, DC.

Lai, F., R. Luo, L. Zhang, X. Huang, and S. Rozelle. 2012. "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Migrant Schools in Beijing." Working Paper 228. Stanford University, Rural Education Action Project, Stanford, CA.

Linden, L. L. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India." InfoDev Working Paper 17. World Bank, Washington, DC.

Lucas, A. M., P. J. McEwan, M. Ngware, and M. Oketch. 2014. "Improving Early Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda." *Journal of Policy Analysis and Management* 33: 950–76.

Malamud, O., and C. Pop-Eleches. 2011. "Home Computer Use and the Development of Human Capital." *The Quarterly Journal of Economics* 126: 987–1027.

Masino, S., and M. Niño-Zarazúa. 2015. "What Works to Improve the Quality of Student Learning in Developing Countries?" Working Paper 2015/033. World Institute For Development Economics Research, United Nations University, Tokyo, Japan.

McEwan, P. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85: 353–94.

Mo, D., J. Swinnen, L. Zhang, H. Yi, Q. Qu, M. Boswell, and S. Rozelle. 2012. "Can One Laptop per Child Reduce the Digital Divide and Educational Gap? Evidence from a Randomized Experiment in Migrant Schools in Bejing." Working Paper 233. Stanford University, Rural Education Action Project, Stanford, CA.

Muralidharan, K., and V. Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *The Economic Journal* 120: F187-F203.

———. 2011. "Teacher Performance Pay: Experimental Evidence from India." *The Journal of Political Economy* 119: 39–77.

Murnane, R. J., and A. J. Ganimian. 2014a. *"Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations."* Unpublished manuscript. Harvard University, Cambridge, MA.

———. 2014b. "Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations." Working Paper 20284. National Bureau of Economic Research, Cambridge, MA.

———. 2016. "Improving Education in Developing Countries: Lessons from Rigorous Impact Evaluations." *Review of Educational Research* 86: 719–755.

Nguyen, T. 2009. "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar." Unpublished manuscript. Massachusetts Institute of Technology, Cambridge, MA.

Ozier, O. 2014. "Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming." Policy Research Working Paper 7052. World Bank, Policy Research Department, Washington, DC.

Piper, B., and M. Korda. 2011. "EGRA Plus: Liberia." Program Evaluation Report. RTI International, Durham, NC.

Sailors, M., J. V. Hoffman, P. D. Pearson, S. N. Beretvas, and B. Matthee. 2010. "The Effects of First- and Second-Language Instruction in Rural South African Schools." *Bilingual Research Journal* 33: 21–41.

Snilstveit, B., J. Stevenson, D. Phillips, M. Vojtkova, E. Gallagher, T. Schmidt, H. Jobse, M. Geelen, M. G. Pastorello, and J. Eyers. 2015. "Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle-Income Countries: A Systematic Review." 3ie Final Review. International Initiative For Impact Evaluation, London.

Spratt, J., S. King, and J. Bulat. 2013. "Independent Evaluation of the Effectiveness of Institut Pour l'Education Populaire's 'Read-Learn-Lead' (RLL) Program in Mali." Endline Report. RTI International, Durham, NC.

Tamim, R. M., R. M. Bernard, E. Borokhovski, P. C. Abrami, and R. F. Schmid. 2011. "What Forty Years of Research Says about the Impact of Technology on Learning a Second-Order Meta-Analysis and Validation Study." *Review of Educational Research* 81 (1): 4–28.

The Campbell Collaboration. 2015. "What Is a Systematic Review?" Accessed March 30, 2015. http://www.campbellcollaboration.org/what_is_a_systematic_review/index.php.

UNESCO. 2014. "Teaching and Learning: Achieving Quality For All." EFA Global Monitoring Report. United Nations Educational, Scientific, and Cultural Organization, Paris.