How Can Developing Countries Address Heterogeneity in Students' Preparation for School? A Review of the Challenge and Potential Solutions*

Alejandro J. Ganimian†

Sharnic Djaker‡

1. Introduction

Over the past two decades, in low- and middle-income countries (LMICs) have rapidly expanded access to schooling by building new schools and making education more affordable. In 1999, gross enrollment in low- and middle-income countries was 11 and 32% in pre-primary education, 78 and 102% in primary education, and 30 and 38% in secondary education, respectively (UNESCO, 2014).[1] By 2015, the corresponding figures were 18 and 42% in pre-primary, 103% and 105% in primary, and 39 and 78% in secondary (UNESCO, 2017).

These expansions have included segments of the population in these countries that had remained at the margins of the school system. Yet, they also presented schools with a new

---

* We gratefully acknowledge the funding provided by the Jacobs Foundation for this review. We also thank Gelgia Fetz and Jessica Tsang for useful comments on earlier drafts of this manuscript.

† Assistant Professor of Applied Psychology and Economics, Steinhardt School of Culture, Education, and Human Development, New York University. Email: alejandro.ganimian@nyu.edu.

‡ Doctoral candidate, Psychology and Social Intervention, Steinhardt School of Culture, Education, and Human Development, New York University. Email: sharnic.djaker@nyu.edu.

[1] Gross enrollment rates surpass 100% when a given education level enrolls students outside the theoretical age for that level (e.g., those who started school rates, repeated a grade, or dropped out and returned to school).

challenge: how to educate larger and more heterogeneous student groups. For example, according to a recent survey representative of rural India, by the end of primary school (i.e., grade 5), only 50% of children could read a grade 2 text, 19% could read a grade 1 text, 13% could only recognize words, 12% could recognize letters, and 6% could not do any of the above. In math, only 18% of children could divide, 25% could subtract, 34% could only recognize two-digit numbers, 19% could recognize one-digit numbers, and 4% could not do any of the above (ASER, 2018). Similar patterns have been observed elsewhere in South Asia and Sub-Saharan Africa (see, e.g., ASER Pakistan, 2020; Uwezo, 2015, 2016; Uwezo, 2021).

In light of this heterogeneity in students' preparation for school, this paper has four goals: (a) reviewing the evidence on the degree to which students of different preparation levels coexist in the same classrooms; (b) proposing a conceptual framework to understand why current incentives and capacity do not lead principals and teachers to address heterogeneity; (c) discussing the extent to which existing experimental and quasi-experimental evidence has addressed these constraints; and (d) identifying unexplained pieces of this puzzle and recommend future directions. Each of the following sections pursues one of these objectives.

2. The Problem: How Much Do Developing-Country Students Differ in their Preparation?

There is mounting evidence that many children in LMICs lag behind curricular expectations. For example, Muralidharan et al. (2019) used data from a diagnostic assessment administered to participants in a blended-instruction program in Delhi, India to show that the typical grade 6 student is 2.5 grades behind in math and 0.5 grades in language. These findings have been replicated elsewhere. For example, in an evaluation of a scaled-up version of the same program in Rajasthan, India Muralidharan and Singh (2020) found that the average grade 6 student was 2.9 grade levels behind in math and 1.5 levels in language. Similarly, in an evaluation of a comparable program in Morazán, El Salvador Büchel et al. (2022) found that the average sixth grader performed 3.5 grade levels behind in math and that achievement spans five grade levels.

These data also suggest that students' achievement increases more slowly with every grade than the curriculum expects. In Delhi, while the average sixth grader performed at a third-grade level (i.e., a three-grade difference), the average ninth grader performed at a fourth-grade level (i.e., a five-grade difference), indicating that the gap between expectations and performance widens as students go through middle school (Muralidharan et al., 2019). In Rajasthan and Morazán, the pattern are remarkably similar (Büchel et al., 2022; Muralidharan & Singh, 2022).

As a result of the two patterns above, student achievement varies widely within each grade. In Delhi, grade 6 students span six grade levels in math and seven grade levels in language (Muralidharan et al., 2019). In Rajasthan, the corresponding figures are six and five, respectively (Muralidharan & Singh, 2022) and in Morazán, it is five levels in math (Büchel et al., 2022).[2]

[2] The magnitude of this heterogeneity in achievement is more stark if expressed using international benchmarks. Das and Zajonc (2010) showed that while the median child in Rajasthan and Orissa, India is failing—42% of enrolled children in the former and 50% of the latter fail to meet a basic international low benchmark of math knowledge—the highest-achieving children in those states perform exceptionally well—the top 5% performers perform far better than their low-income counterparts and fare even better than some high-income countries like Norway. Similarly, Sandefur (2018) found that the average student across 14 African countries performed half a standard deviation below their counterparts in the Organization for Economic Cooperation and Development (OECD) and that students in the region converged with their more advantaged peers very slowly, if at all, during the 2000s. Using data on 164 countries, Angrist et al. (2021) also found limited progress in learning from 2000 to 2017.

Heterogeneity in students' preparation matters for three main reasons. First, it makes teaching more challenging—especially, in contexts of low teacher capacity (see, e.g., Araujo et al., 2016; Bietenbeck et al., 2018; Bold et al., 2017; Ganimian et al., 2022a; Metzler & Woessmann, 2012; Santibañez, 2006; World Bank, 2016). Second, learning outcomes are "sticky": low learning levels become harder to remedy later in a child's schooling trajectory (see, e.g., Bau et al., 2021; Singh, 2019). Lastly, learning outcomes correlate with important life outcomes—including employment and wages and propensity of teenage pregnancy, addition, and/or incarceration (see, e.g., Danon et al., 2022; Das et al., 2022).

Addressing such heterogeneity is possible. Notably, some of the school systems that have improved the most in international assessments over the past decades have done so by reducing the scope and degree of abysmally low performance in their schools (Mourshed et al., 2011). These school systems are not exceptional in their capacity. The frontier challenge in improving learning in developing countries, therefore, is to identify the combination of strategies that best fits a context, to address heterogeneity at scale.

3. The Framework: Why Does Heterogeneity in Student Preparation Emerge and Persist?

Heterogeneity in students' preparation for school is not a novel phenomenon. Yet, its current manifestation in LMICs is distinctive. It affects a large share of school-goers: almost 9 out of 10 school-aged children and youths live in a low- or middle-income country (UNESCO, 2021). Its magnitude is unprecedented: for example, in Rajasthan, India the average student is nearly three grade levels below the curriculum in math by the time he/she reaches grade 6 and almost four grade levels behind by grade 8 (Muralidharan & Singh, 2020). Lastly, its effects are compounded by the large class sizes that are characteristic of LMICs: in a typical school, there are 39 students per teacher in low-income countries and 22 in middle-income countries, compared to only 12 in high-income countries (UNESCO, 2021).

The fundamental question, on which we have made relatively little progress, is why this heterogeneity persists in LMICs (i.e., why teachers seem to be unwilling or unable to address it). Conventional wisdom focuses on the incentives that teachers in these settings face. Specifically, economists have pointed to three patterns that are common to many developing countries. Low-performing children are often pulled out of school by their parents, so teachers have little reason to invest in them given that they will eventually drop out (Banerjee & Duflo, 2011). The ambitious curricula that these countries once adopted to rigorously prepare their political and economic elites for university have left teachers little margin to slow down and make sure all students understand the material (Bhattacharjea et al., 2011; Pritchett & Beatty, 2015; Sinha et al., 2016). Third, the high-stakes exams that many school systems in these contexts have at the end of middle- and high-school encourage teachers to focus on the students mostly likely to take them, who tend to be from relatively more advantaged backgrounds (Duflo et al., 2011).

It is plausible, if not likely, that outdated incentive structures in LMICs play an important role in explaining the persistence of heterogeneity in student preparation. Yet, teacher capacity (i.e., what teachers know and can do)—which has traditionally been studied by psychologists—warrants more attention than it has received thus far. First, most teachers are unaware of the extent of students' underperformance: even though they see students every day, assign and grade their homework, and assess them, too many of them do not know how many of their students perform below curricular expectations or the extent of there are missing such expectations (Djaker et al., 2021). Second, many teachers do not believe that low-performing students can catch up; they overestimate the influence of a child's socio-economic background on their developmental potential and/or underestimate the malleability of their intelligence—especially for disadvantaged sub-groups, such as female students or those from racial/ethnic minorities and low-income families (Farfan Bertran et al., 2021; Hanna & Linden, 2012; Sabarwal et al., 2022).[3] Finally, many teachers do not know how to address foundational gaps in students' skills: they lack subject-specific pedagogical knowledge to explain concepts in multiple ways and address student errors (Araujo et al., 2016; Bietenbeck et al., 2018; Bold et al., 2017; Ganimian et al., 2022a; Metzler & Woessmann, 2012; Santibañez, 2006; World Bank, 2016).

More specifically, while we agree that incentives-focused explanations can help us understand why some teachers do not try to address heterogeneity in student preparation, we see teacher effort as only one link of a causal chain that explains why such variability persists (i.e., why variability in inputs results in variability in outputs). First, teachers need to know the problem exists; if they are not aware that their students vary widely in their performance, they may not think changes in their regular instructional practices are needed. Second, teachers need to believe that the problem is solvable; if they assume that students' backgrounds are too influential and/or that intelligence is innate, or if they doubt their capacity as individuals to make a meaningful change in their students' learning, they may not see the point in changing their instruction. Third, teachers need to be able to act on their beliefs; if they perceive rules or

[3] Existing studies may offer a lower bound estimate of the prevalence of these beliefs, given that they are typically measured through self-reports, which are prone to social-desirability bias (Duckworth & Yeager, 2015).

norms as preventing/discouraging them from focusing on the students who need most help, they may refrain from trying. Lastly, teachers need to know how to solve the problem; if they are not trained on how to explain foundational concepts and procedures to students and they do not have colleagues and/or resources to help them do so, they may try some new strategies, become frustrated, and ultimately revert to what they know. Lastly, the extent to which students vary in their preparation for school cuts across all these steps; the more they vary, the more difficult it will be for teachers to take any of these steps.

We find this framework useful to both to diagnose and solve the problem of heterogeneity. The reasons why teachers do not address such heterogeneity are likely to differ across LMICs and the framework offers a roadmap for descriptive research that can help identify the binding constraints in each setting (e.g., in some contexts, the main obstacle may be that teachers are unaware of their students' skills; in others, it may be that teachers feel unable to address it). The framework can also highlight components that are common to interventions that are successful and contribute to our understanding of what makes these interventions work (e.g., nearly all interventions seeking to address heterogeneity involve assessing students, which suggests that getting teachers to acknowledge the problem may be a key step to address it). We see this as an important contribution, given that most interventions in this space are "bundled" (i.e., they combine multiple components, each trying to address a different potential constraint). In the next section, we use this framework to generate working hypotheses about what makes some interventions targeting heterogeneity more successful than others.

4. The Evidence: What Works to Address Heterogeneity in Student Preparation for School?

To date, LMICs have adopted a wide variety of strategies that can reduce heterogeneity in students' preparation for school, including: (a) expanding access to high-quality early-childhood education; (b) informing principals and teachers of students' achievement; (c) offering remedial education to the lowest-achieving students; (d) assigning students to smaller, more homogeneous classes; (e) using technology for customized or personalized instruction; (f) standardizing lessons to compensate for gaps in teacher capacity; and (g) assigning students to activities based on their achievement. Not all of them were devised with the purpose of reducing heterogeneity, but we include them here because they have potential to address one or more of the steps in the hypothesized causal chain in the prior section.

Importantly, this review is neither comprehensive nor systematic. It simply seeks to illustrate what we have learned about the effectiveness of interventions that may be helpful in addressing heterogeneity. We focus on countries classified by the World Bank as low- or middle-income, with a few exceptions to account for recent transitions across groups.[4]

Our discussion of the evidence proceeds as follows. In each sub-section below, we discuss the main take-aways from studies designed to identify the causal effects of policies and programs—either randomized experiments or quasi-experiments (i.e., differences in differences analyses, regression discontinuity designs, and instrumental variables estimation).[5] We illustrate each take-away by describing one or two studies in some detail—favoring those

[4] The most updated classification is at: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519. We discuss evidence from Chile and Uruguay, even if they are both currently classified as high-income countries.

[5] We exclude studies that do not capitalize on plausibly exogenous sources of variation in treatment assignment because they are not well suited to identify cause-and-effect relationships (Murnane & Willett, 2011).

that are recent and well designed—and we introduce caveats by referring to additional studies. When discussing each study's results, we focus on those that are most relevant to this review (e.g., if an initiative improved both education and nutrition outcomes, we highlight the former). And if a study compares two or more intervention types (e.g., hiring new teachers v. training existing teachers), we discuss each intervention type in different parts of the review (e.g., we discuss the first intervention when considering the evidence on hiring new teachers and discuss the second intervention when considering the evidence on training existing teachers).

Expanding Access to High-Quality Early-Childhood Education

Providing all students—especially, those from disadvantaged backgrounds—with access to high-quality early-childhood education could level the playing field before school starts and prepare children to better understand the material during their schooling trajectory.

Expanding public provision of early-childhood education, however, has only improved children's learning in upper-middle-income countries—where governments can ensure minimum standards of quality. For example, Berlinski and Galiani (2007) exploited variation across regions and cohorts in exposure to a program that built pre-primary education classrooms (for children ages 3 to 5) in Argentina from 1993 to 1999 and found that it increased the probability of pre-primary school attendance by 7.5 percentage points (pp.) Berlinski et al. (2009) used a similar strategy and found that a year of pre-school attendance raised achievement in third-grade, on average, by 0.24 standard deviations (SDs) in math and by 0.23 SDs in Spanish. The program also had large positive impacts on students' attention, effort, discipline, and class participation (as reported by teachers), which are potential mechanisms. Evidence from Uruguay suggests that these gains magnify with time (Berlinski et al., 2008).

In lower-middle- and low-income countries, expansions of public pre-primary education have rarely had this effect—presumably, because they struggle to deliver quality instruction. For example, Bouguen et al. (2018) conducted a randomized evaluation of the construction of pre-school classrooms in primary schools of poor rural villages in Cambodia, accompanied by training, development, and supervision of teachers and materials. After two years, the initiative had small and statistically insignificant impacts overall and a *negative* impact on cognition for five-year-olds—especially, if they had poorer and less educated parents. These effects were partly explained by these parents switching their children from primary schools to pre-schools. Similarly, Blimpo et al. (2019) experimentally evaluated the construction of community-based early-childhood development centers in The Gambia and also found they had negative effects—in this case, among children with wealthier and more educated parents, partly because the reform encouraged them to switch from better-quality early-childhood settings in their homes.[6]

In these settings, only pre-school centers run by non-profits or private providers have improved learning outcomes, but it is not yet clear whether these initiatives can be scaled-up. For example, Martinez et al. (2013) used a randomized experiment to evaluate the construction, equipment, and training for 67 pre-schools run by a non-government organization

[6] Both studies suggest that, when alternatives to traditional public provision are good enough, encouraging parents to switch away from them is not in their children's best interests. Similarly, Bernal et al. (2019) used a randomized experiment to evaluate the effect of offering low-income families in Colombia the possibility of transferring their children (ages 6 to 60 months) from non-parental family daycare units to large childcare centers in urban areas. The reform had a negative impact on cognitive development and no statistically significant effect on socio-emotional development, possibly because the centers were of lower quality than the units.

in the Gaza Province of Mozambique. After two years, the program increased primary-school enrollment (by 24%) and time spent on schooling each week (by 7.2 hours). It also improved children's cognitive and problem-solving abilities, fine-motor skills, and socio-emotional and behavioral outcomes, but it had a null effect on communication and language development. Similarly, Dean and Jayachandran (2020) randomly assigned 808 children ages 3 to 5 to scholarships for private kindergartens in Karnataka, India and, after two years, children induced to attend these centers outperformed their peers by 0.8 SDs on cognitive development, and 40% of this effect persisted through grade 1, but they fared no better on social development.[7]

Attempts to improve public early-childhood education by hiring additional workers have had mixed results. When the existing worker is overburdened and has little margin to farm out his/her responsibilities, an extra worker can increase time on task and children's learning. For example, Ganimian et al. (2022b) ran a randomized evaluation of the hiring of facilitators exclusively devoted to teaching pre-school education in (public) *anganwadi* centers in Tamil Nadu, India. The intervention doubled instructional time and boosted numeracy and literacy test scores by 0.29 and 0.46 SDs after 18 months. It also reduced stunting and malnutrition by freeing up the time of *anganwadi* workers for health and nutrition tasks. Yet, when there are unclear expectations for and/or weak monitoring of the division of labor among workers, hiring

[7] Subsidizing demand for private education has been far less successful when providers are of low quality. For example, Wong et al. (2013) used a randomized experiment to evaluate the free provision of a year of pre-school tuition and a cash transfer conditional on attendance in Lushan County, in China's Henan Province. After a year, children in the treatment group were more likely to attend pre-school (by 20 pp. or 35%), but they scored no better on school readiness. The results appear to be related to the low quality of the available pre-schools.

additional workers can lead to shirking. For example, Andrew et al. (2022) experimentally evaluated extra funds for public pre-school centers in Colombia, which were mostly earmarked for hiring teaching assistants. The intervention reduced the time teachers spent in the classroom and did not improve learning, mimicking the results of hiring teachers by contract in primary schools in Kenya (Duflo et al., 2015) and India (Muralidharan & Sundararaman, 2013).[8]

Improving public provision by training existing teachers has been challenging. One possible reason is that such training often fails to produce the intended changes in teachers' instruction. For example, Yoshikawa et al. (2015) randomly assigned 64 pre-kindergarten and kindergartens in low-income municipalities of Santiago, Chile to a teacher professional-development program supplemented by coaching and a library of 100 books (i.e., an intervention group) or a self-care workshop and four books (i.e., a comparison group). After two years, the intervention had no effect on children's language, literacy, or social-emotional skills—possibly, because it did not increase the time spent on language and literacy (Bowne et al., 2016; Mendive et al., 2016).[9] Some have sought to overcome this challenge by combining trainings for teachers and

[8] Combining the teaching assistants with in-service training for existing teachers, however, improved children's cognitive development by 0.15 SDs (Andrew et al., 2022). The authors argue that it was teachers' incorrect beliefs about the extent to which they can complement the work of assistants that led them to reduce effort, and that combining the assistants with training can correct teachers' views, preventing them from decreasing their effort.

[9] Blimpo et al. (2019) randomly assigned existing early-childhood education centers in The Gambia to receive intensive teacher training on a new curriculum and also found statistically insignificant effects on child learning. Yet, this study did not track changes in teachers' instructional practices, so its null results are harder to interpret.

parents,[10] but when teachers and parents are not aligned in their goals, such combinations may backfire.[11]

Informing Principals and Teachers of Students' Achievement

Informing schools of their students' learning levels could alert them to gaps in their knowledge and skills and help principals allocate resources and teachers allocate time and effort to the subjects (and topics within those subjects) in greatest need of support.

Providing principals with reports on the performance of their students on standardized tests of core subjects (e.g., math and language) has worked in upper-middle-income countries, where most teachers attend school, and the challenge is to change what they do when they go. For example, de Hoyos et al. (2021) experimentally evaluated brief, user-friendly reports on the performance of students in grades 3 to 5 on standardized tests of math and language administered at the end of the prior school year in public primary schools in La Rioja, Argentina. After two years, the reports increased math and language achievement by 0.33 and 0.36 SDs, respectively. In fact, the students who attended schools that were offered such reports still outperformed those who attended schools that were not offered such reports a year later by

[10] For example, Ozler et al. (2018) conducted a randomized evaluation in community-based childcare centers in Malawi to compare training and mentoring for teachers alone and combined with a parenting education program. After 18 months, the former had no effect on learning, but the latter raised language and social-emotional skills.

[11] For example, Wolf et al. (2019) experimentally evaluated a teacher professional development program for public and private kindergartens in the Greater Accra Region of Ghana and found that it improved children's school readiness when provided by itself, but that it had null effects when combined with parental-awareness meetings because parents tended to favor academic learning as opposed to structured play—the focus of the training.

0.26 SDs in math and 0.22 SDs in language, respectively, in the national student assessment. The initiative seems to have achieved such impacts by leading principals to use achievement data to inform management decisions and by increasing the activities that teachers assigned.

Information alone has been less successful in lower-middle-income settings, where many teachers still do not go to school regularly (Chaudhury et al., 2006; Muralidharan et al., 2017). For example, Muralidharan and Sundararaman (2010) randomly assigned 100 public primary schools in rural areas of Andhra Pradesh, India to receive reports on the achievement of students in grades 1 to 5 by grade-appropriate competence that included sub-district, district, and state averages against which the principal could benchmark his/her schools' performance. After one year, the teachers working in schools that were offered the reports fared better than those in schools without such reports on classroom observations (e.g., they were more likely to assign students more work during/after school), but students in both sets of schools performed similarly math and language tests, leading the authors to conclude that some teachers delivered better lessons only when observed. A more intensive initiative in a similar context—Madhya Pradesh, India—had similarly disappointing results (Muralidharan & Singh, 2022).

It is not yet clear whether combining these reports with other interventions (e.g., training) necessarily enhances their impact. In the experiment in La Rioja, Argentina, combining reports with professional development for principals (e.g., on how to use reports to inform school-improvement plans) yielded similar results as the reports by themselves (de Hoyos et al., 2021). In another experiment in Salta, Argentina the same combination of reports and workshops improved passing and repetition rates for the cohort of students with higher dosage (i.e., those who were assessed twice during the study), but it had no such effects for any of the other cohorts, and it had no discernible impact on student achievement (de Hoyos et al., 2022). de Hoyos et al. (2017) conducted a quasi-experimental evaluation of a program in Colima, Mexico that combined reports on students' test scores with a technical advisor to help schools design improvement plans by both exploiting its sudden introduction and school eligibility cutoffs. Both strategies estimated math and language impacts to be 0.12 SDs, but the authors cannot parse out the contribution of each intervention component towards the observed effects.

It is also not clear whether delivering the reports to more stakeholders (e.g., teachers and parents, instead of teachers alone) invariably produces better results. To our knowledge, there is only one study on this question. Andrabi et al. (2017) experimentally evaluated the effect of reports for principals and parents that compared each child's and school's test scores in English, math, and Urdu to those of other children and schools in their village in 823 public and private primary schools in Punjab, Pakistan. After two years, schools that were offered the reports scored 0.11 SDs better across all subjects than schools that were not offered such reports. Yet, given the high share of students enrolled in private schools in this context and the authors' choice to conduct the study on villages with at least one private school, it seems possible that the reports were perceived more as an accountability mechanism (i.e., a tool to threaten schools with the possibility of parents switching their children to schools with greater progress) than as a formative instrument to guide teachers' instructional practices. Thus, we see this intervention as more comparable to school test score disclosures to inform families' enrollment decisions (e.g., Camargo et al., 2018) than with the other programs in this sub-section.

Offering Remedial Education to the Lowest-Achieving Students

Providing the lowest-achieving students with opportunities to learn foundational knowledge and skills could help them catch up with their higher-achieving peers and thus reduce the need for teachers to cater to a wide array of preparation levels in regular lessons.

Pulling children who struggle with the material out of class to reinforce basic concepts and procedures has consistently improved learning outcomes. For example, Banerjee et al. (2007) used a randomized experiment to evaluate a program in Vadodara and Mumbai, in the Indian state of Maharashtra, in which grade 3 and 4 students in public primary schools were taken out of their regular classes by a local instructor for two hours per day. After one year, students in the program outperformed their control counterparts by 0.14 SDs across math and language, and after two years, they did so by 0.28 SDs across both subjects. As intended, the lowest-achieving students (i.e., those in the lowest tercile of baseline test scores) benefited the most: after two years, they outperformed their control counterparts by 0.40 SDs across both subjects. A similar program to improve the reading fluency of grade 3 students in Manizales, Colombia had comparable results, as well as spillover effects in math (Álvarez-Marinelli et al., 2021).

Offering remediation before or after school has also had positive effects on achievement. For example, Saavedra et al. (2017) experimentally evaluated an inquiry-based, after-school remedial-education program for grade 3 students in 48 public primary schools in Lima, Peru. Even if the average student assigned to the intervention only received five to six lessons, they scored 0.12 SDs higher than their non-intervention peers in science after only five months.[12]

[12] Chiplunkar et al. (2022) conducted a randomized evaluation of a remedial after-school program for grade 9 students in public schools in Chennai, in the Indian state of Tamil Nadu, which met for an hour five days per week. The program combined short videos that explained concepts through real-world examples, small-group work in

Yet, few studies can distinguish the effects from additional instructional time and remediation. Duflo et al. (2021) used a randomized experiment to compare a pull-out and an after-school remedial-education program for grade 3 students in public schools in Ghana. After two years, both programs improved English, math, and local-language test scores by about 0.14 SDs, suggesting that the extra time in the after-school version was not a main driver of impact.

Many programs have hired volunteers to deliver remedial instruction to keep costs low. These programs have often improved learning. For example, Cabezas et al. (2021) experimentally evaluated a program in which college students read texts to grade 4 students from public and private subsidized schools for 15 90-minute sessions in Bío Bío and Gran Santiago and, Chile. After three months, the program increased school attendance by 0.81 pp., it increased school grades by 0.11 SDs, it had a 0.07 SDs effect in language, it decreased students' likelihood of dropping out of school (by 1.8 pp.) and it increased their probability of graduating from primary (by 4.5 pp.) and secondary school (by 3.2 pp.) on time. Yet, volunteer-led programs have struggled to ensure lessons are held regularly. For example, Banerjee et al. (2010) conducted a randomized evaluation of an after-school program in which volunteers taught children basic reading skills for two-to-three months in Uttar Pradesh, India. Children who attended benefited considerably: after a year, those who could not read anything were 60 pp. more likely to recognize letters than those who were not offered this program. Yet, 16% of

which students had to solve worksheets, and trained facilitators who led lectures to summarize key concepts based on discussions. After a year, it improved reading and math test scores by 0.47 and 0.52 SDs, respectively, but had no effect on scores on the grade 10 national exam because students lagged too far behind the curriculum.

villages assigned to this condition never implemented the program, those that did varied widely in the number of groups composed, and only 8% of children attended the camps.

Assigning Students to Smaller, More Homogeneous Classrooms

Assigning students to smaller classrooms based on their initial achievement could make instruction more manageable by narrowing the range of material that teachers need to cover and thus potentially improving the match between students' preparation and instruction.

Existing evidence on whether students in LMICs benefit from smaller classrooms is mixed. Quasi-experimental studies that exploit maximum class size rules, comparing the achievement of students just below a threshold that triggers the hiring of a new teacher, and that of students just above that threshold, have consistently found smaller classes positively impact test scores. For example, Urquiola (2006) used that strategy to estimate the impact of a rule that split grade 3 classes into two when more than 30 students were enrolled in rural schools in Bolivia and found that an eight-student reduction raised math and language test scores by up to 0.3 SDs.[13] A subsequent analysis, however, used data from Chile to demonstrate that schools near maximum class-size thresholds often adjust prices or enrollments to avoid adding a classroom (Urquiola & Verhoogen, 2009), raising questions about the validity of this analytical approach.[14]

---

[13] His results are consistent with those of Angrist and Lavy (1999), who first pioneered this approach in Israel. We do not discuss this study in detail in the main text because Israel has long been considered a high-income country.

[14] In fact, when Angrist et al. (2019) applied the same approach to later data from Israel, they found no effect on achievement. They also found evidence of manipulation around the maximum class-size threshold as in Chile, but because it was weakly correlated with socio-economic status, they concluded it could not explain the null results. Further, the same approach suggested a *negative* effect from class reductions in Bangladesh (Asadullah, 2005).

Chin (2005) exploited variation by birth cohort and state to estimate the effect of a national policy in India that sought to provide all one-teacher primary schools with a second teacher. However, the reform was poorly implemented, ultimately failing to reduce class sizes.

The only randomized evaluation of a class-size reduction conducted in a LMIC found that simply reducing the number of students does not improve achievement. Duflo et al. (2015) evaluated a program in Western Province, Kenya that provided 70 public primary schools with funds to hire additional teachers on renewable contracts over two years. They found that the achievement of students in large classes (on average, about 82 students) was indistinguishable from that of students taught by regular (i.e., civil-service, non-contract) teachers in small classes (about 44 students). These null effects could be due to classes of 44 students still being too large or due to the response of regular teachers in intervention schools: on average, they were 16 pp. less likely to be in class teaching if their school received funding.

Assigning students to smaller *and* more homogeneous classes seems to be more promising. In the same experiment as above, Duflo et al. (2011) also compared 61 schools in which grade 1 students were assigned to one of two sections randomly to 60 schools in which they were assigned based on their test scores (i.e., students in the lower half of the score distribution were assigned to one section and those in the upper half were assigned to another) and found that, after 18 months, the latter outperformed the former by 0.14 SDs in math and language. In fact, “tracking” benefited all students: those with initially low achievement improved by 0.16 SDs and those with initially high achievement improved by 0.19 SDs across both subjects. Further, a year after the intervention, students in tracking schools still performed 0.16 SDs better. These findings as suggesting that tracking does not simply benefit students assigned to higher-achieving peers, but rather by changing teachers' responses to the class composition. To our knowledge, however, these results have not yet been replicated in other contexts—possibly, because tracking is highly contentious in education policy circles.[15]

Using Technology for Customized or Personalized instruction

Leveraging technology's capacity to simultaneously present different material to different students could potentially allow for customized or even individualized learning and reduce the need for teachers to have to take time away from lessons to remedy gaps in knowledge/skills.

Providing students with hardware (e.g., desktops or laptops), by itself or pre-loaded with educational software, in hopes that they will use it for independent and/or self-paced learning has typically failed to improve achievement. For example, Cristia et al. (2017) conducted a randomized evaluation of an initiative that provided students in public, rural, multigrade primary schools with electricity in Peru with laptops loaded with 39 open-source applications and 200 age-appropriate e-books. Despite dramatically increasing access to computers, the

[15] Banerjee et al. (2007) leveraged the pull-out remedial education program in Maharashtra, India described in the prior sub-section to compare students who remained in the class when their lower-achieving peers were taken out to students in schools where the program was not offered, and class size and composition remained unchanged. The authors found no discernible effects on students who remained in smaller, more homogeneous classrooms. In the context of the larger experiment in Ghana also described in the previous sub-section, Duflo et al. (2021) compared students assigned to the assistant-led remedial-education program delivered during the school day to those assigned to the version of the same program delivered after school and, upon observing similar impacts, concluded that smaller classes and an extra hour of instruction are almost perfect substitutes. The use of active pedagogy, remedial instruction, and assistants in both programs, however, raises the question of whether the authors would observe the same pattern if only class-sizes were reduced, or school times were expanded.

program had no effect on the math and language test scores of students in grades 2 and 6—possibly, because it did not increase the time students devoted to school activities, school attendance, homework completion, motivation for reading, or the quality of instruction (netbooks were typically used for activities that were unlikely to impact learning outcomes). These results were consistent with those of an initiative that distributed similar laptops for home use among students in public primary schools in Lima, Peru (Beuermann et al., 2015).[16]

To our knowledge, only one laptop-distribution program in an LMIC improved achievement. Mo et al. (2013) experimentally evaluated an initiative in migrant schools in Beijing, China that distributed laptops loaded with three remedial tutoring software products in math and language and trained students and their parents on how to use the laptop and the software. After six months, grade 3 students assigned to the program outperformed control peers by 0.17 SDs in math, but performed on par with them in language. Importantly, however, this effect only emerged once the authors accounted for students' characteristics and class assignment. Before they did so, the estimated effect on math was 0.07 SDs and statistically insignificant. The

[16] They are also consistent with those of Malamud and Pop-Eleches (2011), who evaluated a program in Romania that gave public-school students from low-income families in grades 1 to 12 a voucher for a desktop computer by comparing students right below and above the eligibility cutoff. After a year, students who won a voucher performed better in a test of computer skills and self-reported measures of computer fluency by 0.25 SDs, but they had *lower* grades in school in math, English, and Romanian by 0.25 to 0.33 SDs—likely, because they reported using the computers to play games and reduced the time that they devoted to homework and reading. These results, in turn, are consistent with those of a randomized evaluation of an initiative that sought to integrate free computers into the teaching of language in public schools in Colombia (Barrera-Osorio & Linden, 2009).

change in the magnitude of the effect after covariate adjustment is a cause for concern, given that randomization alone should provide an unbiased estimate of the program's impact.[17]

Getting students to interact with software that allows them to learn and/or review material at their own pace seems to be more promising to address heterogeneity in preparation. Most programs of this type that have been rigorously evaluated, however, encourage students to review the material covered at school on a given week. While this may be useful for some, it is unlikely to help those lagging several grade levels behind the curriculum. Thus, it is perhaps not surprising that these initiatives have had small-to-moderate effects. For example, Lai et al. (2015) conducted a randomized evaluation of a program in 24 migrant schools in three districts of Beijing, China in which grade 3 student attended two 40-minute sessions per week outside of school hours (i.e., during lunch or after school) to play animation-based math games to review and practice the material taught in their regular school math classes. After three months, the program improved the math test scores of the average student by 0.15 SDs.[18] A similar initiative

[17] One study suggests that laptops may not even benefit learning much more than textbooks. Bando et al. (2016) randomly assigned 271 public primary schools in Honduras to receive either laptops with digital content and Internet access or math and Spanish textbooks and workbooks. After seven months, students in grades 3 and 6 in both groups performed similarly on tests of math and Spanish, processing speed and memory, and verbal fluency. Using a randomized experiment, de Hoop et al. (2020) found that an initiative that included free tablets for grade 1 students in Zambia's Eastern Province improved early-grade reading (0.4 SDs) and math (0.22 SDs) and vocabulary (0.25 SDs) after 14 months. Yet, given that it also included other components (e.g., solar power, projectors, and lesson plans for teachers and interactive lessons for students), it is not possible to know what caused these effects.

[18] Evidence from Chile suggests that "gamifying" these programs (e.g., awarding student tokens for completed exercises and organizing competitions) can increase achievement, but also learning anxiety (Araya et al., 2019).

in rural boarding schools in Shaanxi, China raised math test scores by 0.12 SDs (Lai et al., 2013). Neither of these programs, however, had spillover effects on language skills.[19]

The programs above were implemented outside of regular school hours, so it is not possible to determine whether they worked because of the content of the software or because of the additional instructional time.[20] Ma et al. (2020) used a randomized experiment to compare additional instruction with traditional inputs (i.e., workbooks), additional instruction with technology (i.e., game-based software), and no additional instruction among boarding grade 4 and 6 students in 130 rural schools in Shaanxi, China. The two interventions were held on the same days of the week, for the same amount of time, and number of times per year. They also had the same curricular content each week, which was aligned with the official curriculum. After four months, students in the group that was offered additional time with technology performed on par with the control and additional time without technology in tests of math, but it fared slightly better than the control group in teacher-assigned grades in math. The authors

[19] Lai et al. (2012) experimentally evaluated a program held at lunch or after school in 57 primary schools in poor, rural areas of Qinghai, China with a high concentration of ethnic minorities in which grade 3 students attended two 40-minute sessions per week to play Mandarin games to review the material they learned in school that week. After three months, the program improved Mandarin test scores by 0.14 SDs. Given that a non-trivial share of students were not native Mandarin speakers, it is perhaps not surprising that the program also improved their math test scores by 0.22 SDs—possibly, by improving their capacity to understand math material.

[20] Mo et al. (2014) conducted a randomized evaluation of a version of the program previously evaluated by Lai et al. (2013) held during school hours in 72 rural schools in Shaanxi, China. However, this iteration took place during computer class, so it still constituted an increase in instructional time spent in math. The same is true of a similar program that sought to improve the English skills of grade 4 students in Qinghai, China (Mo et al., 2020).

interpreted these findings as indicative of effects being driven by time. These results seem to be consistent with those of a program in Brazil that replaced a period of regular math instruction per week with interaction with a software and did not impact learning (Ferman et al., 2019).

It is possible that the answer to this question may vary based on the details of the program. For example, Büchel et al. (2022) randomly assigned students in grades 3 to 6 in Morazán, El Salvador to additional instructional time in math led by a teacher or to additional instructional time in math based on an educational software, led either by technical supervisors or a teacher. Each group attended two 90-minute lessons per week. After six months, extra time based on software was more effective that extra time led by a teacher, and extra time based on software was equally effective when it was led by technical supervisors or teachers. It could also be that there are diminishing returns to additional time with technology. For example, Bettinger et al. (2020) randomly assigned students in grade 3 in two regions of Russia to interact with an educational software for 45 or 90 minutes per week or to no additional instructional time and found that going from no time to 45 minutes improved math achievement, but doubling time did not produce further improvements in both math and (especially) language.

Software that *dynamically* adjusts the content and difficulty of the material based on students' performance has had moderate-to-large impacts on achievement. For example, Muralidharan et al. (2019) experimentally evaluated a blended-instruction program in Delhi, India in which students in grades 4 to 9 attended after-school centers for 90 minutes per week six times per week and interacted for 45 minutes with a software and 45 minutes with an instructor in small groups (in the latter case, for homework completion and exam preparation).[21] After four months and a half, the program improved math test scores by 0.37 SDs and language (Hindi) test scores by 0.23 SDs. The program had similar absolute test-score impacts for all students, but it had much higher relative effects for the lowest-performing students, who learned little at school over the same period (based on the control group's progress). These results are consistent with those of Banerjee et al. (2007), who conducted a randomized evaluation of a program in Vadodara, Gujarat in which grade 4 students interacted for two hours per week with a software to play games and solve math problems and whose difficulty responded to students' performance. The program increased math scores by 0.35 SDs after one year and by 0.47 after two years and it was equally effective for all students.

Like the software products previously discussed, however, these have also been delivered outside of regular school hours, so their effects may be due to additional instructional time. Linden (2008) tried to address this question by randomly assigning schools in Gujarat, India to two versions of the computer-assisted learning program evaluated by Banerjee et al. (2007): one administered during school hours and one administered before or after school. A key limitation of this study, however, is that the program was administered in a relatively well-functioning network of schools run by a non-profit. Thus, the negative effects that the author found for the in-school version could be because these schools offer a better quality of instruction than the public schools where the program was originally administered.

Standardizing Lessons to Compensate for Gaps in Teacher Capacity

[21] Among other things, the software adjusted the topics and difficulty of exercises based on students' ability to answer them correctly and it provided feedback on incorrect answers.

Pre-determining the content and pedagogical strategies for each lesson could potentially compensate for challenges in teachers' recruitment, selection, and training by getting them to employ strategies that have been found effective to build students' foundational skills.

In the most extreme version of this approach, full lessons are pre-recorded for teachers to play in their classrooms. This strategy has mainly been adopted in hard-to-staff areas, which struggle to recruit teachers with minimum qualifications. It seems to conceive teachers primarily as facilitators to deliver pre-packaged content to students and affords them little discretion. For example, Naslund-Hadley et al. (2014) experimentally evaluated an initiative in Cordillera, Paraguay that distributed 108 audio CDs that covered the entire pre-school math curriculum. Given that classrooms were mostly bilingual, audios used Spanish and Guaraní. The audios were used four days a week and the fifth day was set aside for teachers to review what had been learned during that week, based on the topics they believed students needed reinforcement. Each 60-minute lesson included preparation, playing of the audios, and practice exercises. After five months, the program improved math test scores by 0.16 SDs, narrowing gaps between high- and low-achieving students, and those with and without trained teachers.[22]

[22] In schools that struggle to recruit teachers, lessons can be broadcasted live from a central location. For example, Johnston and Ksoll (2017) conducted a randomized evaluation of a program in Ghana that broadcasted math and English lessons live via satellite to primary-school students in rural areas. After two years, it improved numeracy scores by 0.23 SDs but had no impact on literacy. These results are consistent with those of Naik et al. (2019), who experimentally evaluated a program in Karnataka, India that broadcasted math, science, and English lessons via satellite to students in grades 5 to 10 in rural public schools, replacing a third of regular instruction in the subjects.

In an only slightly more flexible version, teachers are given scripts that spell out what they should say and do at every moment of a lesson. It has been deployed in settings with novice teachers with minimal training. Teacher discretion is limited to choosing among recommended strategies to respond to unforeseen situations (e.g., behavioral disruptions). For example, Gray-Lobe et al. (2022) exploited lotteries that awarded scholarships to pre-primary and primary students in Kenya to attend a network of standardized private schools to evaluate their impact. At these schools, teachers received tablets with scripts, principals were trained on how to observe lessons and give teachers feedback to encourage script compliance, and students are assessed seven times a year to track their progress, inform teacher training, and refine scripts. (Importantly, only 23% of primary-school teachers at these schools had a teaching certificate.) After two years, lottery winners in pre-primary and primary schools performed 1.35 SDs and 0.81 SDs above lottery losers across five tests based on the national curriculum, respectively. These effects were larger for initially low-achieving students, they were similar even among teachers with little experience and low content knowledge, and they were largely uniform across sites, thus reducing heterogeneity in student preparation along multiple dimensions.[23]

After two years, it improved test scores in these subjects by 0.12 SDs. Two evaluations of a similar initiative in Mexico suggests that these programs can also increase years of schooling (Fabregas, 2019; Navarro-Sola, 2021).

[23] These results are aligned with those of Ganimian et al. (2022a), who experimentally evaluated a program that recruited college students in science, technology, engineering, and math and gave them scripts to teach math and science to grade 5 and 6 students in public schools in Pune, in Maharashtra, India. After a year, the program raised students' test scores by 0.34 SDs in math and 0.22 SDs in science, and it had a 0.15 SDs spillover effect in language. The above results are also consistent with those of Piper et al. (2014), who conducted a randomized evaluation

Most of the initiatives of this type, however, are less prescriptive than the two programs described above. They typically provide teachers with a combination of pre-recorded lesson segments (not full lessons), teaching and learning materials (e.g., textbooks and workbooks), and/or technology (e.g., tablets, whiteboards, TV screens). They seek to influence and/or support teachers' pedagogy rather than to prescribe it. Their effect, however, has been mixed for reasons that are not entirely clear. For example, de Barros (2020) randomly assigned public-school students in grades 9 and 10 of Haryana, India to: a program that combined short videos mapped to the math and science curriculum, infrastructure upgrades (two smart classrooms, TVs, tablets, and a power inverter), workbooks for students, and teacher training; a program that combined only the workbooks and training; or a business-as-usual control group. After six months, the first program *negatively* impacted math test scores by 0.14 SDs and had no impact on science and the second one had null effects on both subjects. Consistent with these results, the first program but not the second one negatively impacted instructional quality.

One possibility is the effect of these programs depends on whether teachers perceive the materials as complements to or substitutes for their effort. For example, Beg et al. (2020), compared two versions of a program in Punjab, Pakistan: one, for grade 8 students, combined

among grade 1 and 2 students in formal (public) and non-formal schools in Kenya of an initiative that combined lesson scripts with teacher training, support from tutors, and teacher observation and feedback. After a year, the program raised English and Kiswahili test scores by 0.34 to 0.58 SDs, depending on the grade and subject. Lastly, they are consistent with those of Eble et al. (2021), who used a randomized experiment to evaluate an initiative that combined teachers hired on a contract basis, a scripted curriculum, and monitoring of teachers. After three years, the program improved literacy and numeracy test scores by 3.2 SDs among students in grades 1 to 3.

short video lectures in which an expert teacher explained key concepts in math and science, tablets for teachers with questions for them to check student understanding and the official textbooks, and LED television screens for teachers to project the material and use the board interactively; the other one, for grade 6 students, shared the same two components but not the third one, and students were given tablets to self-paced learning. (Importantly, the videos had 30 hours of content for the entire year). After four months, the first version of the program improved test scores by 0.30 SDs; the second one *negatively* impacted test scores by 0.43 SDs. It is possible that while the screens "crowded in" teacher effort by making them responsible for improving their instruction, the student tablets "crowded out" their effort by bypassing them.

Providing teachers with materials that can complement their instruction at specific moments of their lessons has had much more encouraging results. For example, He et al. (2008) experimentally evaluated two materials to teach English in grades 1 to 5 in public schools: a machine that allows students to point to objects and hear them pronounced aloud and to identify words and receive feedback, or flashcards with activities for teachers (e.g., chants and poems). This study evaluated the effect of these two components combined and delivered by assistants (in year 1) and combined or separately and delivered by regular teachers (in year 2). All versions of the program improved English test scores (both when the components were combined and separate), but the version taught by the teachers also improved math scores.

This study, combined with the one above, raises the possibility that some of the scaffolding provided to teachers may serve as a model and help them improve their own practices.[24]

Assigning Students to Activities Based on their Achievement

Administering diagnostic tests to students, grouping them according to their performance on those tests within the same classroom, and assigning different activities to each group—also known as "differentiated instruction" or, in some circles, "teaching at the right level"—could allow teachers to make instruction more closely match their students' preparation levels.

This approach has positively impacted learning when it is held outside of regular lessons, but teachers have been reluctant to take it up on their own during the school day. For example, Banerjee et al. (2011) conducted a randomized evaluation comparing four versions of a differentiated-instruction program in two districts of Bihar and Uttarakhand, India: one in which teachers administered the program during the summer, another one in which volunteers implemented it (in Bihar, in the community, independently from teachers; in Uttarakhand, at school alongside teachers), and two in which teachers were expected to administer the

[24] Similarly, He et al. (2009) conducted a randomized evaluation of an early-literacy curriculum in Mumbai, in the Indian state of Maharashtra, which combined materials (e.g., storybooks, flashcards, alphabet charts) and a library. It was evaluated among grade 1 students in public schools (during and after school); pre-school students in stand-alone classes; and pre-school students in non-profit or public centers. The first and last versions improved reading scores: the first one between 0.26 SDs (during school) and 0.55 SDs (after school) and the second one by 0.7 SDs. Yet, evidence from Karnataka, India suggests that the not all materials improve learning (de Barros et al., 2021). Kerwin and Thornton (2021) use a randomized experiment to compare two versions of an early-literacy program in Uganda and find that teachers may be less productive over the short-term if they need to overhaul their instructional practices and practice to impact learning outcomes over the medium-to-long term.

program during regular lessons—with materials and/or training. The first version raised math and reading test scores by 0.08 SDs by the end of the summer, with effects persisting (0.06 and 0.07 SDs in math and reading) two years after the program ended. The second version improved math and reading test scores by 0.11 SDs when implemented outside of schools and it had a null effect when implemented at schools. The last two versions had mostly null effects. The authors concluded that this pattern of results was explained by teachers believing that implementing this pedagogy would make it harder for them to cover the official curriculum.[25]

Recent evidence suggests that, even in contexts where teachers are reluctant to take up differentiated instruction, it can still improve school management and classroom instruction. For example, Duflo et al. (2021) experimentally evaluated a program in Ghana in which public-school teachers in grades 1 to 3 were trained to implement the approach pioneered in India. Teachers reported that grouping students within their classrooms was too challenging, so they after two 13-week terms, they started grouping them across grades with one teacher per grade. Even after these changes, by the end of the third year, only 6% of teachers adopted the model. Yet, despite the low take up, schools randomly assigned to the program were more likely to have a principal present during unannounced visits and to have teachers in their classroom,

[25] There is some reason to believe, however, that this constraint may not be as strong in other programs/settings. For example, Gallego et al. (2021) experimentally evaluated a program in three cities of Peru that sought to shift the instructional practices of pre-school teachers in math, from rote to inquiry- and problem-based methods. After nine months, the program improved math test scores by 0.14 SDs, although effects faded one year after it ended. Even if this initiative required a radical change in teaching practices, it seems plausible that teachers in pre-schools in Peru feel less pressure to complete the curriculum than their peers in other contexts and/or education levels.

engaged with students, and using the intervention materials. Further, test scores improved, albeit slightly (by 0.07 SDs in math after 6 months and by 0.09 in English after two years).

5. The Frontier: What Else Might We Try to Address Heterogeneity in Student Preparation?

Our review allows us to draw some general lessons about the progress that economists, psychologists, and education researchers more generally have made to date in identifying approaches to address heterogeneity in students' preparation for school in LMICs. Below, we outline these lessons, identify gaps in the evidence, and propose directions for future research.

Three Lessons from This Review

First and foremost, we hope our review makes clear that there are many more strategies to tackle this problem than is commonly acknowledged. Differentiated instruction has received the lion's share of the attention (see, e.g., Global Education Evidence Advisory Panel, 2020). This focus is understandable: this intervention has been evaluated more times than most other approaches covered by this review by some of the world's leading experts (e.g., Banerjee et al., 2010; Banerjee et al., 2011; Banerjee et al., 2007), and it has been implemented under various modalities (e.g., summer camps, during and after school) by one of the largest and most competent education non-profits in the developing world, which has in turn inspired a global network of like-minded organizations.[26] Yet, given how challenging it has proven to scale up, school systems might consider complementing this approach with others in this review.

[26] For an up-to-date list of such organizations, visit: https://palnetwork.org.

Second, our review also sheds light on why interventions to address heterogeneity should not be viewed as interchangeable. The temptation to do so is comprehensible: evidence reviews often highlight the most (cost-)effective interventions to improve an outcome (e.g., enrollment or achievement; see Conn, 2017; Evans & Popova, 2017; Glewwe & Muralidharan, 2016; Kremer et al., 2013; Masino & Niño-Zarazúa, 2015; McEwan, 2014; Snilstveit et al., 2015), leading governments and donors to devote their efforts to making these reforms work in LMICs. Yet, as one of us has previously noted (see Ganimian & Murnane, 2016), broad goals such as addressing heterogeneity in student preparation can be broken down into specific objectives (e.g., seeking to ensure low-performing students reach minimally acceptable levels of learning is not the same as attempting to improving the learning of students at all preparation levels), each of which calls for a different intervention (e.g., the former objective would be best pursued through remedial education; the latter through differentiated instruction). And, as he has argued elsewhere (see Ganimian et al., 2020), instead of chasing the most effective intervention for a broad goal, each school system ought to focus on evaluations of interventions that match its objectives and that can be implemented given its capacity.

Lastly, and relatedly, it is sobering to see that—with the exception of expansions in access to pre-school education, which are politically popular (see, e.g., Corrales, 1999; Grindle, 2004; Kaufman & Nelson, 2004)—most interventions featured in our review have not been scaled-up, even those with consistently positive effects on learning outcomes. This pattern should make us reconsider the traditional model of seeking to understand the effect of an intervention by first evaluating it in small scale under the best conditions possible (i.e., what is known as an "efficacy trial") and then evaluating it at scale (i.e., what is known as an "effectiveness trial").

This model was borne out of a legitimate concern: we should not induce system-level changes before we have reason to believe they might work. Yet, education interventions change when they are scaled-up—often, bearing little resemblance to the versions assessed at small scale. What have been typically regarded as implementation details (e.g., how to hire/train implementers, how to monitor dosage/fidelity, how to provide feedback on implementation) are actually *core components* of such interventions that necessarily need to change when an intervention is enacted at scale, and that often explain why an initiative that proved successful at small scale can have null effects at large scale (see, e.g., Bold et al., 2018; Duflo et al., 2015). We are thus increasingly convinced of the need to evaluate interventions directly at scale (see Muralidharan & Niehaus, 2017), conducting small-scale descriptive work to better understand the context and its constraints whenever necessary (see Yeager, 2021).

Five "Blind Spots" and Ideas to Address Them

Despite the remarkable progress that researchers have made in identifying ways to address heterogeneity in student preparation, several approaches remain relatively underexplored. First, the evidence on teachers' misestimations of students' skills suggests that providing teachers with this information (e.g., through diagnostic assessments that highlight students' strengths and weaknesses; see, e.g., de Hoyos et al., 2021; de Hoyos et al., 2022; de Hoyos et al., 2017)[27] and/or training and supporting them to collect it on their own regularly (e.g.,

[27] An unsuccessful experiment with this approach in India was interpreted as suggesting that information without incentives for teachers has little bite. Yet, teacher absence rates in India are much higher than in the middle-upper income countries where this intervention has worked (see Chaudhury et al., 2006; Muralidharan et al., 2017),

conducting one-on-one assessments with their lowest performers to understand the misconceptions underlying their errors) could at least get them to acknowledge the problem.[28]

Second, research on teachers' intelligence mindsets and biases against student sub-groups suggests that providing them with information on children's and youth's capacity to overcome environmental risk factors and/or on how their students have improved could persuade them to exert more effort. We know of no studies in LMICs that have attempted this approach,[29] but moderation analyses on the impact of growth-mindset interventions in high-income countries suggest it might be promising (Walton & Yeager, 2020; Yeager et al., 2021; Yeager et al., 2019).

Third, evidence on the incentives that teachers in LMICs face indicates that changing such incentives (e.g., having government officials communicate to teachers that they want them to focus on improving the skills of lowest-performing students) may make teachers less reluctant to address heterogeneity. Given that these are system-level changes, there is relatively little causal evidence on their effects (students cannot be randomly assigned to receive or not receive these changes to estimate their impact). However, a recent study in Tanzania found

---

which suggests that the frontier challenge in this context is to improve the external margin of teacher effort (i.e., getting teachers to go to school) rather than the internal margin (i.e., increasing their effort when they show up).

[28] In fact, the evidence on the mismatch between teachers' beliefs and students' skills provides yet another plausible mechanism for how differentiated instruction may improve learning outcomes: by making teachers aware of the skills (and gaps therein) of their students.

[29] In some of the studies above, such information was given to school principals, who often shared it with teachers (de Hoyos et al., 2021; de Hoyos et al., 2022). Others have provided it to parents (Andrabi et al., 2017). But, to our knowledge, no study to date has specifically provided it to teachers to persuade them of their students' potential.

that when teachers in grades 1 and 2 were asked to focus on foundational skills, their students' performance on these skills improved (Rodriguez-Segura & Mbiti, 2022). These results are consistent with those of the tracking study discussed above (Duflo et al., 2011).

Fourth, studies on teachers' subject-matter expertise, pedagogical knowledge, and understanding of students' misconceptions in LMICs suggests that getting teachers to address heterogeneity in student preparation will ultimately require that they are trained on both diagnosing and remedying gaps in foundational skills. The former includes selecting test items that identify gaps in foundational skills, choosing response formats to elicit misconceptions, and aggregating responses to uncover topics/skills in need of reinforcement. The latter entails explaining concepts and procedures that students are supposed to learn in lower grades.[30] What works in this front will likely look different across subjects and grades (and possibly, even within subject-by-grade combinations and even topics), but one high-leverage place to start would be lower grades and core subjects (math and reading). Traditional (i.e., lecture-style) training is unlikely to be enough for these purposes (see Popova et al., 2022). A promising alternative entails providing teachers with opportunities for constant practice and feedback (e.g., Bruns et al., 2018; Cilliers et al., 2022; Cilliers et al., 2018; Majerowicz & Montero, 2020).[31]

[30] The field has largely sought to circumvent this challenge by routinizing instruction of basic topics and skills (see, e.g., Abeberese et al., 2014; Lucas et al., 2014; Piper et al., 2018). Yet, data from high-income countries suggests that, to meaningfully improve their instruction, teachers will eventually need training on both general and subject-specific pedagogy (Kane et al., 2014; Kane & Staiger, 2012).

[31] This evidence is consistent with that of low-capacity teachers in high-income countries, who also stand to benefit from this type of intensive support (see Kraft et al., 2018).

Finally, a relatively under-discussed dimension of heterogeneity in students' preparation for school is the social-emotional domain. Students who are the first in their families to attend school may need additional scaffolding to believe they can succeed, regulate their behavior to pay attention in class, resolve conflicts with others productively, and collaborate with peers. There is surprisingly little evidence on how to address these challenges in LMICs (for exceptions, see, e.g., Alan et al., 2019; Alan & Ertac, 2018; McCoy et al., 2021). Yet, a number of psychological interventions (e.g., cognitive behavioral therapy) for high-risk populations in the U.S. (Heller et al., 2017) are starting to be replicated and evaluated in these settings.

6. References

Abeberese, A. B., Kumler, T. J., & Linden, L. L. (2014). Improving reading skills by encouraging children to read in school: A randomized evaluation of the Sa Aklat Sisikat reading program in the Philippines. *The Journal of Human Resources, 49*, 611-633. https://doi.org/10.1353/jhr.2014.0020

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*.

Alan, S., & Ertac, S. (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy, 126*(5), 1865-1911.

Álvarez-Marinelli, H., Berlinski, S., & Busso, M. (2021). Remedial education: Evidence from a sequence of experiments in Colombia. *Journal of Human Resources*, 0320-10801R10802.

Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review, 107*(6), 1535-1563. https://doi.org/10.2139/ssrn.2538624

Andrew, A., Attanasio, O., Bernal, R., Cardona Sosa, L., Krutikova, S., & Rubio-Codina, M. (2022). *Pre-school quality and child development*. National Bureau of Economic Research (NBER).

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics, 114*(2), 533-575.

Angrist, J. D., Lavy, V., Leder-Luis, J., & Shany, A. (2019). Maimonides' rule redux. *American Economic Review: Insights, 1*(3), 309-324.

Angrist, N., Djankov, S., Goldberg, P. K., & Patrinos, H. A. (2021). Measuring human capital using global learning data. *Nature, 592*, 403-408.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics, 131*(3), 1415-1453.

Araya, R., Arias Ortiz, E., Bottan, N. L., & Cristia, J. P. (2019). *Does gamification in education work?: Experimental evidence from Chile*. Inter-American Development Bank.

Asadullah, M. N. (2005). The effect of class size on student achievement: Evidence from Bangladesh. *Applied Economics Letters, 12*(4), 217-221. https://doi.org/10.1080/1350485042000323608

ASER. (2018). *Annual status of education report (ASER) 2018: Provisional*. ASER Centre.

ASER Pakistan. (2020). *Annual Status of Education Report (ASER) Pakistan 2019: Provisional*.

Bando, R., Gallego, F. A., Gertler, P. J., & Romero, D. (2016). Books or laptops? The cost-effectiveness of shifted from printed to digital delivery of educational content *Economics of Education Review, 61*, 162-173.

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

Banerjee, A. V., Banerji, R., Duflo, E., & Walton, M. (2011). *Effective pedagogies and a resistant education system: Experimental evidence on interventions to improve basic skills in rural India*.

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

Banerjee, A. V., & Duflo, E. (2011). Top of the class. In *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs.

Barrera-Osorio, F., & Linden, L. L. (2009). *The use and misuse of computers in education: Evidence from a randomized experiment in Colombia*. The World Bank.

Bau, N., Das, J., & Yi Chang, A. (2021, 2021/07/01/). New evidence on learning trajectories in a low-income setting. *International Journal of Educational Development, 84*, 102430. https://doi.org/https://doi.org/10.1016/j.ijedudev.2021.102430

Beg, S. A., Lucas, A. M., Halim, W., & Saif, U. (2020). Engaging teachers with technology increased achievement, bypassing teachers did not. *American Economic Journal: Economic Policy, 14*(2), 61-90.

Berlinski, S., & Galiani, S. (2007). The effect of a large expansion of pre-primary school facilities on preschool attendance and maternal employment. *Labour economics, 14*(3), 665-680.

Berlinski, S., Galiani, S., & Gertler, P. J. (2009). The effect of pre-primary education on primary school performance. *Journal of Public Economics, 93*, 219-234. https://doi.org/10.1016/j.jpubeco.2008.09.002

Berlinski, S., Galiani, S., & Manacorda, M. (2008). Giving children a better start: Preschool attendance and school-age profiles. *Journal of Public Economics, 92*(5-6), 1416-1440.

Bernal, R., Attanasio, O., Peña, X., & Vera-Hernández, M. (2019). The effects of the transition from home-based childcare to childcare centers on children's health and development in Colombia. *Early Childhood Research Quarterly*, 418-431.

Bettinger, E., Fairlie, R. W., Kapuza, A., Kardanova, E., Loyalka, P., & Zakharov, A. (2020). *Does edtech substitute for traditional learning? Experimental estimates of the educational production function*. Stanford Institute for Economic Policy Research (SIEPR).

Beuermann, D. W., Cristia, J., Cruz-Aguayo, Y., Cueto, S., & Malamud, O. (2015). Home computers and child outcomes: Short-term impacts from a randomized experiment in Peru. *American Economic Journal: Applied Economics, 7*(2), 53-80.

Bhattacharjea, S., Wadhwa, W., & Banerji, R. (2011). *Inside primary schools: A study of teaching and learning in rural India*. ASER.

Bietenbeck, J., Piopiunik, M., & Wiederhold, S. (2018). Africa's skill tragedy: Does teachers' lack of knowledge lead to low student performance? *Journal of Human Resources, 53*(3), 553-578.

Blimpo, M. P., Carneiro, P., Jervis, P., & Pugatch, T. (2019). *Improving access and quality in early childhood development programs: Experimental evidence from The Gambia*. The World Bank.

Bold, T., Filmer, D. P., Martin, G., Molina, E., Stacy, B., Rockmore, C., Svensson, J., & Wane, W. (2017). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa. *Journal of Economic Perspectives, 31*(4), 185-204.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics, 168*, 1-20.

Bouguen, A., Filmer, D., Macours, K., & Naudeau, S. (2018). Preschool and parental response in a second best world: Evidence from a school construction experiment. *Journal of Human Resources, 53*(2), 474-512.

Bowne, J. B., Yoshikawa, H., & Snow, C. E. (2016). Experimental impacts of a teacher professional development program in early childhood on explicit vocabulary instruction across the curriculum. *Early Childhood Research Quarterly, 34*, 27-39.

Bruns, B., Costa, L., & Cunha, N. (2018). Through the looking glass: Can classroom observation and coaching improve teacher performance in Brazil? *Economics of Education Review, 64*(1), 214-250.

Büchel, K., Jakob, M., Kühnhanss, C., Steffen, D., & Brunetti, A. (2022). The relative effectiveness of teachers and learning software: Evidence from a field experiment in El Salvador. *Journal of Labor Economics, 40*(3), 737-777.

Cabezas, V., Cuesta, J. I., & Gallego, F. A. (2021). *Does short-term school tutoring have medium-term effects? Experimental evidence from Chile*. Ponticia Universidad Católica de Chile (PUC).

Camargo, B., Camelo, R., Firpo, S., & Ponczek, V. (2018). Information, market incentives, and student performance: Evidence from a regression discontinuity design in Brazil. *The Journal of Human Resources, 53*(2), 414-444.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006). Missing in action: Teacher and health worker absence in developing countries. *The Journal of Economic Perspectives, 20*(1), 91-116.

Chin, A. (2005). Can redistributing teachers across schools raise educational attainment? Evidence from Operation Blackboard in India. *Journal of Development Economics, 78*(2), 384-405. https://doi.org/10.1016/j.jdeveco.2004.09.004

Chiplunkar, G., Dhar, D., & Nagesh, R. (2022). *Too little, too late: Improving post-primary learning outcomes in India*. University of Oxford.

Cilliers, J., Fleisch, B., Kotze, J., Mohohlwane, N., Taylor, S., & Thulare, T. (2022). Can virtual replace in-person coaching? Experimental evidence on teacher professional development and student learning. *Journal of Development Economics, 155*, 102815.

Cilliers, J., Fleisch, B., Prinsloo, C., & Taylor, S. (2018). How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. *The Journal of Human Resources*. https://doi.org/10.3368/jhr.55.3.0618-9538R1

Conn, K. M. (2017). Identifying effective education interventions in Sub-Saharan Africa: A meta-analysis of rigorous impact evaluations. *Review of Educational Research, 87*(5), 863-898.

Corrales, J. (1999). *The politics of education reform: Bolstering the supply and demand; overcoming institutional blocks*. The World Bank.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2017). Technology and child development: Evidence from the One Laptop per Child program. *American Economic Journal: Applied Economics, 9*(3), 295-320.

Danon, A., Das, J., de Barros, A., & Filmer, D. P. (2022). *Cognitive and socio-emotional skills in low-income countries: Construct and predictive validity*. Learning and Educational Achievement in Pakistan Schools (LEAPS).

Das, J., Singh, A., & Chang, A. Y. (2022). Test scores and educational opportunities: Panel evidence from five low- and middle-income countries. *Journal of Public Economics, 205*, 104570.

Das, J., & Zajonc, T. (2010). India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement. *Journal of Development Economics, 92*(2), 175-187.

de Barros, A. (2020). *Do students benefit from blended instruction? Experimental evidence from public schools in India*. Harvard Graduate School of Education (HGSE).

de Barros, A., Fajardo-Gonzalez, J., Glewwe, P., & Sankar, A. (2021). *Can public servant performance be increased? Experimetnal evidence on efforts to improve teaching in India*. Department of Economics, Massachusetts Institute of Technology (MIT).

de Hoop, T., Ring, H., Siwach, G., Dias, P., Tembo, G., Rothbard, V., & Toungui, A. (2020). *Impact of e-learning technology and activity-based learning on learning outcomes: Experimental evidence from community schools in rural Zambia*. American Institutes for Research (AIR).

de Hoyos, R., Ganimian, A. J., & Holland, P. A. (2021). Teaching with the test: Experimental evidence on diagnostic feedback and capacity-building for schools in Argentina. *World Bank Economic Review, 35*(2), 499-520.

de Hoyos, R., Ganimian, A. J., Holland, P. A., & Djaker, S. (2022). *Great things come to those who wait: Experimental evidence on performance management tools and training for public schools in Argentina*. The World Bank.

de Hoyos, R., García-Moreno, V. A., & Patrinos, H. A. (2017). The impact of an accountability intervention with diagnostic feedback: Evidence from Mexico. *Economics of Education Review, 58*, 123-140.

Dean, J. T., & Jayachandran, S. (2020). *Attending kindergarten improves cognitive development in India, but all kindergartens are not equal*. University of Chicago Booth School of Business.

Djaker, S., Ganimian, A. J., & Sabarwal, S. (2021). *Out of sight, out of mind? The gap between students' performance and teachers' estimations in Bangladesh and its implications for instruction*. Steinhardt School of Culture, Education, and Human Development, New York University.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237-251.

Duflo, A., Kiessel, J., & Lucas, A. M. (2021). *Existing education systems can deliver more learning*. Department of Economics, University of Delaware.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review, 101*(5), 1739-1774. https://doi.org/10.1257/aer.101.5.1739

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics, 123*, 92–110. https://doi.org/10.1016/j.jpubeco.2014.11.008

Eble, A., Frost, C., Camara, A., Bouy, B., Bah, M., Sivaraman, M., Hsieh, P.-T. J., Jayanty, C., Brady, T., & Gawron, P. (2021). How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in The Gambia. *Journal of Development Economics, 148*, 102539.

Evans, D. K., & Popova, A. (2017). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *The World Bank Research Observer, 31*, 242-270.

Fabregas, R. (2019). *Broadcasting human capital? The long term effects of Mexico's telesecundarias*. LBJ School of Public Affairs, University of Texas at Austin.

Farfan Bertran, M. G., Holla, A., & Vakis, R. (2021). *Poor expectations: Experimental evidence on teachers' stereotypes and student assessment*. The World Bank.

Ferman, B., Finamor, L., & Lima, L. (2019). *Are public schools ready to integrate ed-tech classes with Khan Academy?* Munich Personal RePEc Archive (MPRA).

Gallego, F. A., Naslund-Hadley, E., & Alfonso, M. (2021). Changing pedagogy to improve skills in preschools: Experimental evidence from Peru. *The World Bank Economic Review, 35*(1), 261-286.

Ganimian, A. J., Hess, F. M., & Vegas, E. (2020). *Realizing the promise: How can education technology improve learning for all?* Brookings Institution.

Ganimian, A. J., Mbiti, I. M., & Mishra, A. (2022a). *Improving STEM learning: Experimental evidence from urban primary schools in India*. Steinhardt School of Culture, Education, and Human Development, New Yor University

Ganimian, A. J., Muralidharan, K., & Walters, C. R. (2022b). *Augmenting state capacity for child development: Experimental evidence from India*. National Bureau of Economic Research (NBER).

Ganimian, A. J., & Murnane, R. J. (2016). Improving educational outcomes in developing countries: Lessons from rigorous evaluations. *Review of Educational Research, 86*(3), 719-755.

Glewwe, P., & Muralidharan, K. (2016). *Improving school education outcomes in developing countries: Evidence, knowledge gaps, and policy implications*.

Global Education Evidence Advisory Panel. (2020). *Cost-effective approaches to improving global learning: What does recent evidence tell us are "smart buys" for improving learning in low- and middle-income countries?* World Bank; Foreign, Commonwealth & Development Office; Building Evidence in Education.

Gray-Lobe, G., Keats, A., Kremer, M., Mbiti, I., & Ozier, O. (2022). *Can education be standardized? Evidence from Kenya*. Development Innovation Lab (DIL), University of Chicago.

Grindle, M. S. (2004). *Despite the odds: The contentious politics of education reform*. Princeton University Press.

Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy, 4*(4), 146-168.

He, F., Linden, L. L., & MacLeod, M. (2008). *How to teach English in India: Testing the relative productivity of instruction methods within the Pratham English language education program*. Columbia University.

He, F., Linden, L. L., & MacLeod, M. (2009). *A better way to teach children to read? Evidence from a randomized controlled trial*. Abdul Latif Jameel Poverty Action Lab (J-PAL).

Heller, S., Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. A. (2017). Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago. *The Quarterly Journal of Economics, 132*, 1-54.

Johnston, J., & Ksoll, C. (2017). *Effectiveness of interactive satellite-transmitted instruction: Experimental evidence from Ghanaian primary schools*. Stanford Center for Education Policy Analysis (CEPA).

Kane, T. J., Kerr, K., & Pianta, R. C. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. John Wiley & Sons.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains*.

Kaufman, R. R., & Nelson, J. M. (2004). *Crucial needs, weak incentives: Social sector reform, democratization, and globalization in Latin America*. Woodrow Wilson Center Press.

Kerwin, J. T., & Thornton, R. (2021). Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures. *Review of Economics and Statistics, 103*(2), 251-264.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547-588.

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

Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. (2015). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education Review, 47*, 34-48. https://doi.org/10.1016/j.econedurev.2015.03.005

Lai, F., Zhang, L., Hu, X., Qu, Q., Shi, Y., Qiao, Y., Boswell, M., & Rozelle, S. (2013). Computer assisted learning as extracurricular tutor? Evidence from a randomised experiment in rural boarding schools in Shaanxi. *Journal of development effectiveness, 5*(2), 208-231. https://doi.org/10.1080/19439342.2013.780089

Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M., & Rozelle, S. (2012). *Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in public schools in rural minority areas in Qinghai, China*. Rural Education Action Program (REAP).

Linden, L. L. (2008). *Complement or substitute? The effect of technology on student achievement in India*. Abdul Latif Jameel Poverty Action Lab (J-PAL).

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda. *Journal of Policy Analysis and Management, 33*, 950-976. https://doi.org/10.1002/pam.21782

Ma, Y., Fairlie, R. W., Loyalka, P., & Rozelle, S. (2020). *Isolating the "tech" from edtech: Experimental evidence on computer assisted learning in China*. National Bureau of Economic Research (NBER).

Majerowicz, S., & Montero, R. (2020). *Can teaching be taught? Experimental evidence from a teacher coaching program in Peru*. Universidad de los Andes.

Malamud, O., & Pop-Eleches, C. (2011). Home computer use and the development of human capital. *The Quarterly Journal of Economics, 126*, 987-1027. https://doi.org/10.1093/qje/qjr008

Martinez, S., Naudeau, S., & Pereira, V. (2013). *The promise of preschool in Africa: A randomized impact evaluation of early childhood development in rural Mozambique*. International Initiative for Impact Evaluation (3ie).

Masino, S., & Niño-Zarazúa, M. (2015). *What works to improve the quality of student learning in developing countries?* World Institute for Development Economics Research.

McCoy, D. C., Hanno, E. C., Ponczek, V., Pinto, C., Fonseca, G., & Marchi, N. (2021). Um compasso para aprender: A randomized trial of a social-emotional learning Program in homicide-affected communities in Brazil. *Child development, 92*(5), 1951-1968.

McEwan, P. J. (2014). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research, XX*, 1-42. https://doi.org/10.3102/0034654314553127

Mendive, S., Weiland, C., Yoshikawa, H., & Snow, C. E. (2016). Opening the black box: Intervention fidelity in a randomized trial of a preschool teacher professional development program. *Journal of Educational Psychology, 108*(1), 130-145.

Metzler, J., & Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics, 99*(2), 486-496.

Mo, D., Bai, Y., Shi, Y., Abbey, C., Zhang, L., Rozelle, S., & Loyalka, P. (2020). Institutions, implementation, and program effectiveness: Evidence from a randomized evaluation of computer-assisted learning in rural China. *Journal of Development Economics*.

Mo, D., Swinnen, J., Zhang, L., Yi, H., Qu, Q., Boswell, M., & Rozelle, S. (2013). Can one-to-one computing narrow the digital divide and the educational gap in China? The case of Beijing migrant schools. *World Development, 46*, 14-29. https://doi.org/10.1016/j.worlddev.2012.12.019

Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., Qiao, Y., Boswell, M., & Rozelle, S. (2014). Integrating computer-assisted learning into a regular curriculum: Evidence from a randomised experiment in rural schools in Shaanxi. *Journal of development effectiveness, 6*, 300-323. https://doi.org/10.1080/19439342.2014.911770

Mourshed, M., Chijioke, C., & Barber, M. (2011). *How the world's most improved school systems keep getting better*. McKinsey & Company.

Muralidharan, K., Das, J., Holla, A., & Mohpal, A. (2017). The fiscal cost of weak governance: Evidence from teacher absence in India. *Journal of Public Economics, 145*(C), 116-135.

Muralidharan, K., & Niehaus, P. (2017). Experimentation at scale. *Journal of Economic Perspectives, 31*(4), 103-124.

Muralidharan, K., & Singh, A. (2020). *Improving school productivity through computer-aided instruction: Experimental evidence from Rajasthan*. University of California, San Diego.

Muralidharan, K., & Singh, A. (2022). *Improving public sector management at scale? Experimental evidence on school governance in India*. National Bureau of Economic Research (NBER).

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review, 109*(4), 1426-1460.

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

Muralidharan, K., & Sundararaman, V. (2013). *Contract teachers: Experimental evidence from India*. National Bureau of Economic Research (NBER).

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Naik, G., Chitre, C., Bhalla, M., & Rajan, J. (2019). Impact of use of technology on student learning outcomes: Evidence from a large-scale experiment in India. *World Development, 127*, 104736.

Naslund-Hadley, E., Parker, S. W., & Hernandez-Agramonte, J. M. (2014). Fostering early math comprehension: Experimental evidence from Paraguay. *Global Education Review, 1*, 135-154.

Navarro-Sola, L. (2021). *Secondary schools with televised lessons: The labor market returns of the Mexican Telesecundaria*. Human Capital and Economic Opportunity Global Working Group, University of Chicago.

Ozler, B., Fernald, L. C. H., McConnell, C., Neuman, M., & Fraga, E. (2018). Combining pre-school teacher training with parenting education: A cluster-randomized controlled trial. *Journal of Development Economics, 133*, 448-467.

Piper, B., Sitabkhan, Y., Mejía, J., & Betts, K. (2018). *Effectiveness of teachers' guides in the Global South: Scripting, learning outcomes, and classroom utilization*. RTI Press.

Piper, B., Zuilkowski, S. S., & Mugenda, A. (2014). Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. *International Journal of Educational Development, 37*, 11-21.

Popova, A., Evans, D. K., Breeding, M. E., & Arancibia, V. (2022). Teacher professional development around the world: The gap between evidence and practice. *The World Bank Research Observer, 37*(1), 107–136.

Pritchett, L., & Beatty, A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. *International Journal of Educational Development, 40*, 276-288.

Rodriguez-Segura, D., & Mbiti, I. (2022). *Back to the basics: Curriculum reform and student learning in Tanzania*. Research on Improving Systems of Education (RISE).

Saavedra, J. E., Naslund-Hadley, E., & Alfonso, M. (2017). *Targeted remedial education: Experimental evidence from Peru*. National Bureau of Economic Research (NBER).

Sabarwal, S., Abu-Jawdeh, M., & Kapoor, R. (2022). Teacher beliefs: Why they matter and what they are. *The World Bank Research Observer, 37*(1), 73-106.

Sandefur, J. (2018, 2018/02/01/). Internationally comparable mathematics scores for fourteen african countries. *Economics of Education Review, 62*, 267-286. https://doi.org/https://doi.org/10.1016/j.econedurev.2017.12.003

Santibañez, L. (2006). Why we should care if teachers get A's: Teacher test scores and student achievement in Mexico. *Economics of Education Review, 25*(5), 510-520.

Singh, A. (2019). Learning more with every year: School year productivity and international learning divergence. *Journal of the European Economic Association*, 1-44.

Sinha, S., Banerji, R., & Wadhwa, W. (2016). *Teacher performance in Bihar, India: Implications for education*. The World Bank.

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., Jobse, H., Geelen, M., Pastorello, M. G., & Eyers, J. (2015). *Interventions for improving learning outcomes and access to education in low- and middle-income countries: A systematic review*. International Initiative for Impact Evaluation (3ie).

UNESCO. (2014). *EFA global monitoring report 2013/4: Teaching and learning: Achieving quality for all*. United Nations Educational, Scientific, and Cultural Organization (UNESCO).

UNESCO. (2017). *EFA global monitoring report 2017/8. Accountability in education: Meeting our commitments*. United Nations Educational, Scientific, and Cultural Organization (UNESCO).

UNESCO. (2021). *Global education monitoring report 2021/2: Non-state actors in education. Who chooses? Who loses?* United Nations Educational, Scientific and Cultural Organization (UNESCO).

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural Bolivia. *Review of Economics and Statistics, 88*(1), 171-177. https://doi.org/10.1596/1813-9450-2711

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review, 99*, 179-215. https://doi.org/10.1257/aer.99.1.179

Uwezo. (2015). *2014 Tanzania annual assessment report*. Uwezo East Africa Regional Office.

Uwezo. (2016). *2015 Kenya annual assessment report*. Uwezo East Africa Regional Office.

Uwezo. (2021). *Are our children learning? Illuminating the Covid-19 learning losses and gains in Uganda. Uwezo national learning assessment report*. Uwezo Uganda.

Walton, G. M., & Yeager, D. S. (2020). Seed and soil: Psychological affordances in contexts help to explain where wise interventions succeed or fail. *Current Directions in Psychological Science, 29*(3), 219-226. https://doi.org/10.1177/0963721420904453

Wolf, S., Aber, J. L., Behrman, J. R., & Tsinigo, E. (2019). Experimental impacts of the “Quality Preschool for Ghana” interventions on teacher professional well-being, classroom quality, and children’s school readiness. *Journal of Research on Educational Effectiveness, 12*(1), 10-37.

Wong, H. L., Luo, R., Zhang, L., & Rozelle, S. (2013). The impact of vouchers on preschool attendance and elementary school readiness: A randomized controlled trial in rural China. *Economics of Education Review, 35*, 53-65. https://doi.org/10.1016/j.econedurev.2013.03.004

World Bank. (2016). *Time on task study in secondary schools: Madhya Pradesh & Tamil Nadu, 2015-2016*. The World Bank and Educational Initiatives.

Yeager, D. S. (2021). *Lessons from a decade of mindset interventions*. New York University.

Yeager, D. S., Carroll, J. M., Buontempo, J., Cimpian, A., Woody, S., Crosnoe, R., Muller, C., Murray, J., Mhatre, P., Kersting, N., Hulleman, C. S., Kudym, M., Murphy, M., Duckworth, A. L., Walton, G. M., & Dweck, C. S. (2021, 2022/01/01). Teacher mindsets help explain where a growth-mindset intervention does and doesn't work. *Psychological Science, 33*(1), 18-32. https://doi.org/10.1177/09567976211028984

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hullerman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang Hooper, S., Carvalho, C., Hahn, R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A. L., & Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*, 364-369.

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Arbour, M. C., Barata, M. C., Weiland, C., Gómez, C., Moreno, L., Rolla, A., & D'Sa, N. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Journal of Developmental Psychology, 51*, 309-322. https://doi.org/10.1037/a0038785.supp