Public Disclosure Authorized

Public Disclosure Authorized

Who Learns What in Basic Education? Evidence from Indonesia*

April 2018

Abstract

Indonesia consistently fares poorly on the international tests like PISA and TIMSS. To help diagnosing the causes of poor learning we track student achievement across 9 years in basic education. We find that 40% of students do not learn the basics in the early grades of primary school (recognizing 2-digit numbers by 2nd grade, and ordering 4-digit numbers by 4th grade). We also find that schools do not cover the complete Indonesian curriculum. Only few students learn how to calculate the surface area of a triangle by 5th grade, the rules about the order of operations, and to complete exercises embedded in stories. Poor and incomplete coverage of the primary curriculum helps explain the low levels of student achievement we observed in secondary school. Our analysis also provides directions for future research. We observed a large catching-up effect in learning in 6th grade. A plausible explanation for this pattern is the increased pressure on schools, teachers and students to perform well on the high-stakes national exams. The fact that the system can produce learning once (all) actors are sufficiently motivated, suggests that 1.) performance pressure might help, and 2.) that low levels of teacher's knowledge and skills are currently not a major binding constraint to learning in Indonesia.

Keywords: Basic education, Indonesia, learning crisis

Australian Government

* *This paper is prepared by Rythia Afkar, Joppe de Ree, and Noviandri Khairina (World Bank, Indonesia).*

Financial support for this paper was generously provided by the Government of Australia's Department of Foreign Affairs and Trade (DFAT), through the Supporting 12 Years Quality Education for All (ID-TEMAN) Trust Fund.
Disclaimer: *The views expressed in this publication are the author's alone and are not necessarily the views of the Australian Government or the World Bank.*

1 Introduction

Based on extensive micro-level evidence about learning (and the lack thereof) from across the developing world, the 2018 World Development Report (World Bank 2018) issued a warning about a learning crisis.[1] In many developing countries, too many students learn too little in school. While the World Bank (2018) and many others have highlighted the importance of rigorous evidence about "what works" in education,[2] they have also taken a step back to look at the big picture. They argue that, "Innovation in classrooms won't have much impact if technical and political barriers at the system level prevent a focus on learning at the school level. This is the case in many countries stuck in low-learning traps; extricating them requires focused attention on the deeper causes." Countries stuck in a low-learning trap need to "align actors to make the entire system work for learning." (World Bank, 2018).

Indonesia may be stuck in such a low-learning trap. Indonesia has one of the largest education systems in the world, with a school-aged population (under 15 years old) just slightly smaller than that of the European Union.[3] However, this large group of children is not learning much. International studies of learning achievement have shown that Indonesian students have performed poorly overall with no stable trend towards recovery. Today's 15 year olds, for example, do no better on PISA's science component than the 15 year olds who were tested more than a decade ago in 2006. The lack of progress is striking given that the education budget roughly doubled in real terms between 2000 and 2010,[4] and this has led to a renewed sense of urgency among policymakers in Indonesia about

[1]ASER (2014), PASEC (2014), Singh (2017), Kaffenberger and Pritchett (2017), and World Bank (2018)
[2]Evans and Popova (2016) summarize much of the recent experimental evidence in the field of education.
[3]The EU has an under-15 population of 786,000 compared with 719,000 in Indonesia (United Nations population statistics https://esa.un.org/unpd/wpp/Download/Standard/Population/).
[4]World Bank (2013). One key driver of the expansion of the education budget was the 2005 teacher certification program, which mandated that all teachers had to be certified by 2015. Certified teachers in turn received a professional allowance (also called the certification allowance) equal to a teacher's base pay level. The teacher certification program, therefore, effectively doubled teacher salaries. The research that has been done on this program has found that the certification had no learning effects in the short and medium terms, while potentially stronger effects (due to attracting better candidates to the profession) may only materialize far in the future (De Ree et al, 2018). Whether or not the certification program improves teaching and learning quality in the long term, it has certainly put heavy pressure on the budget and has probably crowded out other investments.

the current state of the education system.[5] This is exemplified by a recent article in Indonesia's leading newspaper *Kompas* in which the finance minister commented on the lack of return to investments in education.[6]

Before system-wide reforms can be attempted, there needs to be a clear and accurate diagnosis of the nature of the problem that they aim to address. This paper contributes to such a diagnosis by presenting evidence on the amount of learning that takes place in Indonesian primary and junior secondary schools and answering the following question: which specific skills are acquired by how many students, and in which grade? In our research, we compared high-output schools with low-output schools[7] and low socioeconomic status (SES) students with high SES students.[8] High-output schools are schools with the highest mean test scores in the last grades of primary (6th grade) and junior secondary (9th grade) schools, and low-output schools are those with the lowest mean test scores in the same grades. In particular, we analyzed student performance on a set of 10 mathematics questions that were administered simultaneously to students in different grades. (For example, we looked at students' performance on precisely the same math question that was administered simultaneously to 4th graders and 5th graders.) Based on this data, we were able to make clear and direct comparisons of differences in the levels of achievement of different students at a single point in time, as well as assessing how much knowledge they gain during a year in school (given some assumptions).

We found that a substantial group of students do not learn basic numeracy skills in the early years of primary school. About 40 percent of students could not recognize two-digit numbers by the end of 2nd grade, and about 50 percent could not order four-digit numbers from big to small by the end

[5] Note that in that the past two decades enrollment rates, especially in secondary school, also increased. The stable performance of those in school may be explained by an increase in teaching quality, while serving more students from lower socioeconomic backgrounds (who tend to score lower on average). But, as enrollment rates were already quite high in the year 2000 (95% of 7-12 year olds and 80% of 13-15 year olds were in school) we find that the stability of PISA scores generally supports the claim that education quality has not much improved in the recent decade (and a half).

[6] *Kompas*, February 24 2016, page 12 (Title: Anggaran Pendidikan Digugat)

[7] High-output schools are schools with higher mean test scores in the last grades of primary (6th grade) and junior secondary (9th grade) schools, and low-output schools are those with the lowest mean test scores in the same grades.

[8] The socioeconomic status of students was determined by survey questions about household assets, administered to students before they did the test.

of 4th grade (see column 9 of Table 2.1 in Annex 3).[9] Towards the end of 6th grade, however, we observed a notable catching-up effect when students (and teachers) were preparing for the national exams. For example, 60 percent of students knew how to calculate the volume of a rectangular cuboid at the end of 6th grade.[10] As only 50 percent of the 4th graders were able to arrange four-digit numbers by the end of 4th grade, this indicates that 20 percent of them[11] had caught up tremendously, learning some geometry and some complex multiplication[12] from scratch in the last one or two years of primary school. We observed this 6th grade catching-up effect in low-output schools as well as in high-output schools.

In our view, the most plausible explanation for this result is that schools, teachers, and students feel the pressure of the national exams as they get closer to the end of the six-year primary cycle. The Indonesian national exams are high-stakes and determine students' eligibility to enroll in the most attractive secondary schools. The 6th grade catch-up shows that when sufficiently incentivized, all schools (even those at the bottom end of the achievement distribution) can produce more learning. This corroborates findings of Muralidharan and Sundararaman (2011) who showed that, in India, monetary incentives based on student-learning results motivated teachers to improve their students' test scores. It also suggests to us that a general lack of ability or skill on the part of teachers is currently not a binding constraint to learning. At least not at the current low levels of student achievement in Indonesia. Indeed, the majority of primary teachers has at least two years of post-secondary training, and the median class sizes in primary schools are well below 30. These learning conditions should be sufficient for the vast majority of students to learn one-digit multiplication by the end of 3rd grade, for example, instead of the mere 57 percent who can actually do so (see column 9 of Table 2.1 in Annex 3).

[9] Not surprisingly, these problems are bigger in low-output schools. In the bottom 30 percent of schools, roughly two-thirds of students are essentially mathematically illiterate at the end of 4th grade.
[10] Forty percent of students learned this in 6th grade. This is half of the 80 percent who could not do it by the end of 5th grade (see column 8-10 of Table 2.1).
[11] Twenty percent of the 50 percent who could not arrange four-digit numbers at the end of 4th grade.
[12] To calculate this volume, you need to be able to do the following multiplication exercise, $8 \times 18 \times 12 = \cdots$

Not surprisingly, the slow progress in the early years of primary and the subsequent catch-up have some clear and measurable downsides. First, the 6th grade catch-up effect started way too late for most students to make up for lost time. Not everybody has the home support or the mental ability to learn in two years what is normally learned in six years of school. Second, we found that the curriculum was not being covered in its entirety. This applied to all students, not just to those who were severely behind in 4th grade. For example, some mathematical concepts were hardly taught at all, even in high-output schools, including some aspects of geometry, rules about the order of operations (first adding and subtraction, then multiplication and division), and "story problems." The incomplete coverage of the primary curriculum was a likely cause of some of the poor performance that we observed in secondary schools where completing exercises requires a mastery of multiple skills.

A final conclusion we drew is that the differences in learning outcomes that we found between socioeconomic groups is mainly (but not exclusively) a between-school phenomenon. This means that students in the same classroom from different socioeconomic groups are learning at about the same speed. Low-SES students in high-output schools, therefore, do much better than high-SES students in low-output schools. Nevertheless, across the entire sample, there is still a strong socioeconomic gradient in learning outcomes because low-SES students are overrepresented in low-output schools.

The rest of this report is organized as follows. In Section 2, we describe the data and the empirical approach that we used. In Section 3, we present and discuss our results, and in Section 4, we suggest some issues for future policymaking and research.

2 Data and Empirical Approach

We used data from a near-representative dataset of primary and junior secondary schools in Indonesia. The data has been used before, for example, in De Ree et al (2018) and World Bank

(2016). We selected 240 primary and 120 junior secondary schools at random from 20 randomly selected districts in Indonesia. Prior to sampling these 20 districts, some districts were excluded from the sampling frame. Some districts were considered too small (46), too dangerous to visit (5) or were already selected for a study that was executed in parallel to the study on the effects of the teacher certification program (20).[13] The excluded areas however represent only a small percentage of the Indonesian population. From the 383 remaining districts in the sampling frame, we randomly selected 20 districts, stratified across five major regions in Indonesia. We selected more districts in regions with larger populations. A list of the selected districts and the strata they were sampled from as well as a map of the sampled districts are shown in Annex 1.

Field teams visited the schools three times, once in November 2009, a second time in May 2011, and a third time in May 2012. In this paper, we have used data from the May 2011 and the May 2012 rounds of the survey.[14] A major component of the field visits was the testing of all 40,000 students in the selected schools. The tests that were administered in each round were developed by the testing agency Puspendik, which is part of the Indonesian Ministry of Education and Culture. An effort was made to match the content of these tests with the Indonesian curriculum. As a result, the raw scores (in other words, the percentage that got the correct answers) give a sense of the extent to which students are mastering the learning goals that are laid out in the curriculum. However, because the level of difficulty of tests can differ from year to year and between grades, directly comparing raw test scores over time or between grades can be misleading.

Therefore, what we do in this paper is to present evidence on student performance on specific test items that were anchored (remained unchanged) between adjacent grades. For example, the question $9 \times 7 = \cdots$ was part of both the 2nd and 3rd grade tests in the May 2011 field visit. By comparing the students' performance on these anchored test items across grades, we obtained information on

[13] The study that was executed and planned simultaneously with the study on the effects of the teacher certification program was a study into the effects of teacher working groups.

[14] The May surveys were held at the end of the school year. This makes them better suited for a mapping against the Indonesian curriculum than the November survey.

levels of achievement in the 2nd and 3rd grades as well as (under assumptions) on the amount of learning that takes place during a year in school.[15] [16]

As mentioned above, the design of the tests adhered closely to the Indonesian curriculum. The items in both the grade *g* and *g +1* tests were usually components of the grade *g* curriculum. In other words, students are supposed to know the tested concept by the end of grade *g* and certainly by the end of grade *g+1*. By analyzing these anchored questions, we therefore got a good sense of the extent to which students are learning at the same pace as the curriculum or with a one-year delay.

We have studied the Indonesian curriculum in detail and compared it to curricula in other countries, including Singapore, the U.K., Finland, and Australia[17]. These countries tend to score substantially higher than Indonesia on internationally comparable tests like PISA or TIMSS, but their curricula are generally quite similar to Indonesia's although there are some differences. For example, $9 \times 7 = \cdots$ is part of the 2nd or 3rd grade curriculum of all the above mentioned countries. This overlap between curricula is not reflected in outcomes. The differences in student achievement between Indonesia and Singapore, for example, are striking. Singapore's students scored roughly 2.5 Indonesian standard deviations above the Indonesian average on the latest PISA math test in 2015 or almost 6 years of schooling ahead.[18]

These results already indicate that a substantial percentage of Indonesian students do not learn at curriculum pace. This phenomenon, where the curriculum out-paces actual learning, can be a cause of poor learning. The idea is that teachers follow the curriculum and effectively teach to those who are on track. It then becomes pretty much impossible for those running behind to catch up. Pritchett

[15] World Bank (2016) also used the anchoring structure in the test items to gather evidence on absolute learning in Indonesia. The present paper, however, goes into much more detail about the kind of skills that students learn.

[16] We interpret a difference performance between, for example, 3rd and 2nd graders at one point in time, as evidence for learning. The approach depends on assumptions. Loosely speaking, we assume that 2nd graders and 3rd graders are similar on average, except for the fact that 3rd graders are a year older than 2nd graders and spent one more year in school. In Annex 1 we test a prediction of this assumption. We test equality of asset levels of 3rd and 2nd graders and usually do not reject. See Annex 1, Table 1.1.

[17] Detailed curriculum comparison is provided in the supplementary document.

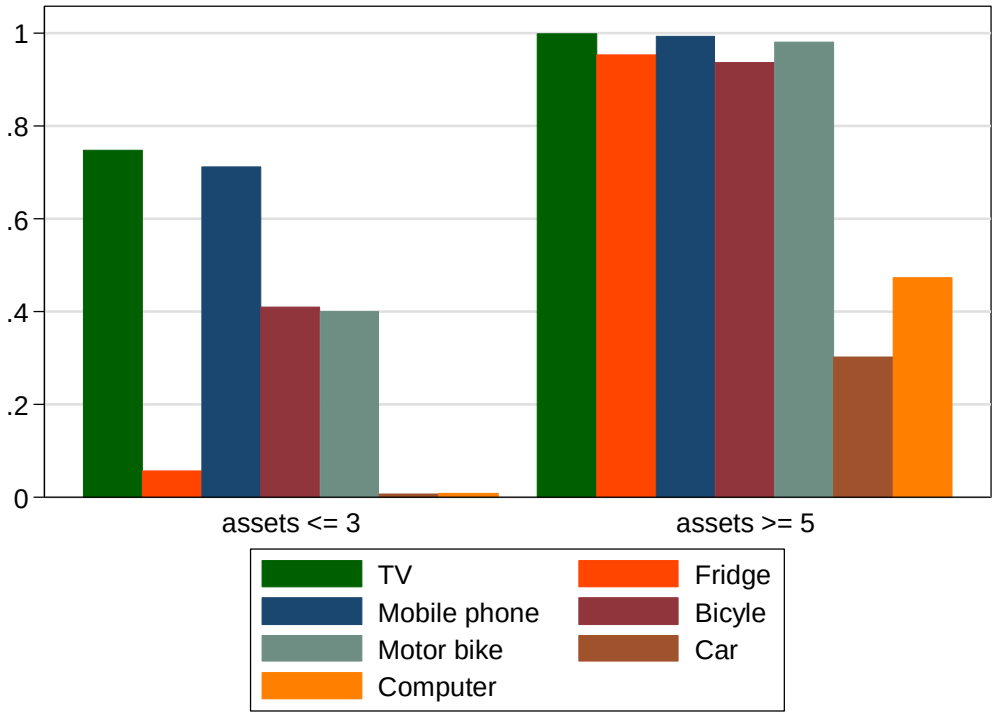
[18] Singapore scored 564 and Indonesia scored 386 on average. At an Indonesian standard deviation of about 70, the difference between the averages is $\frac{564-386}{70} \approx 2.5$ standard deviations (https://data.oecd.org/pisa/science-performance-pisa.htm).

and Beatty (2012) made this point and analyzed this concept theoretically. Their model showed that "paradoxically, learning could go faster if curricula and teachers were to slow down."

In this study, we show that many students lose sight of the curriculum very early on in their learning career. But we attempt to go further in describing this phenomenon. Who actually are the students who are learning? We proposed to break down the data along two dimensions, by the socioeconomic status of students and by the schools' final outputs (in other words, the average test scores of students graduating from 6th grade in primary school and 9th grade in junior secondary school). To construct a measure of socioeconomic background, we used an interesting feature of our data. After doing the test, students had to fill out a set of additional questions on household assets. They were asked whether their household owned a TV, a fridge, a mobile phone, a bicycle, a motor bike, a car, and a computer (by ticking 7 boxes into yes or no). The total number of assets available to the households was our asset index (up to a maximum of 7).

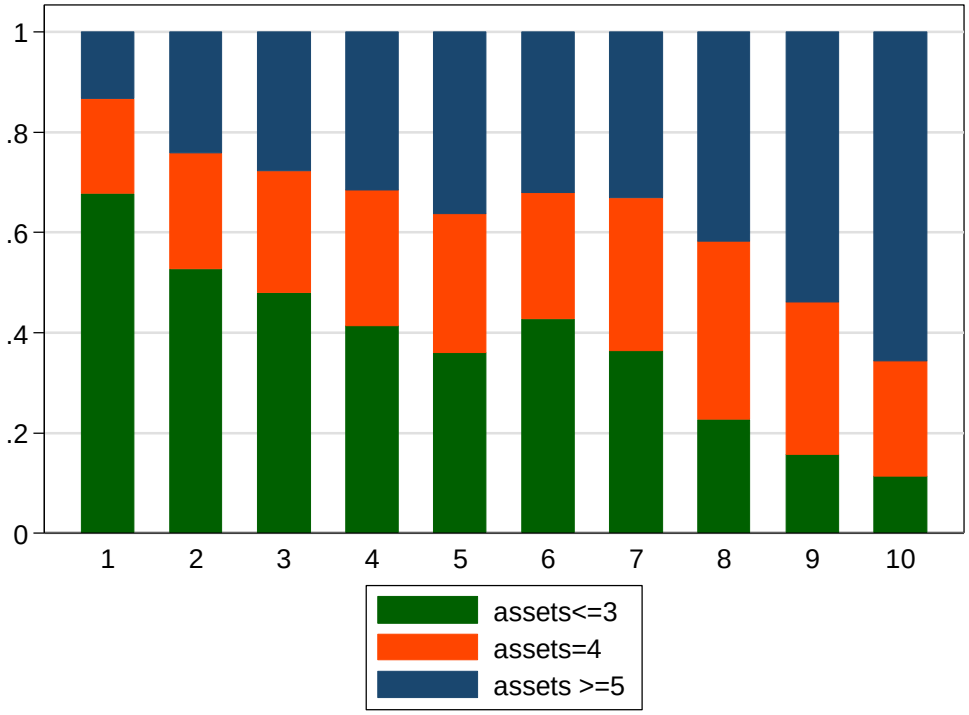
Based on this, we compared the performance of those with few assets (three or less) with those with "many" assets (five or more). Figure 1 shows that this breakdown makes intuitive sense. Households with three of these assets or less tended to have TVs and mobile phones and occasionally own a bicycle or a motorcycle. None of these "low-SES" households had a fridge at home, which appears to set this group apart from "high-SES" households. Households with five assets are indeed substantially better-off.

Figure 1: High-low Asset Level Breakdown



We also distinguished between schools. On the basis of the average 6th grade test scores for primary schools and average 9th grade test scores for junior secondary schools, we sorted schools into (school-output) deciles. It was no surprise that low-SES students were more common in schools in the bottom deciles of the test score distribution. However, a non-negligible fraction of low-SES students was enrolled in high-output schools. Figure 2 presents the shares of households by socioeconomic group and by school-output decile.

Figure 2: Shares of Households in Each Asset Category by School-level Output Decile



Notes: The school output decile was determined by final test scores.

In the next section, we present our results on learning graphically, showing learning gains as arrows, where the length of the arrow is the (estimated) fraction of students who learned a skill in a year in school. We used 20 arrows per graph to distinguish low-SES and high-SES students, and the 10 output deciles.

Note that we did not actually observe the same students learning a skill but instead compared students in grade *g* with students in grade *g + 1* in the same schools. That said, although it is possible to compare grade 2 in 2011 with grade 3 in 2012, i.e. the same students but a year older, we did not use this comparison because *the same* students were not asked *the same* questions twice. Instead, we made certain assumptions, for example, that students in grade *g + 1* would have had similar scores a year previously as grade *g* students had at the time the tests were administered. We tested an implication of this assumption in Table 1.1 in Annex 2. While at times we found differences in the background characteristics of grade *g* and grade *g + 1* students, these differences tended to be

relatively small and not systematic. However, these results do indicate that we must be careful not to focus on specific results for specific school-output deciles (which might depend on occasional changes in population characteristics).

3 Learning in Indonesia's Primary and Secondary Schools

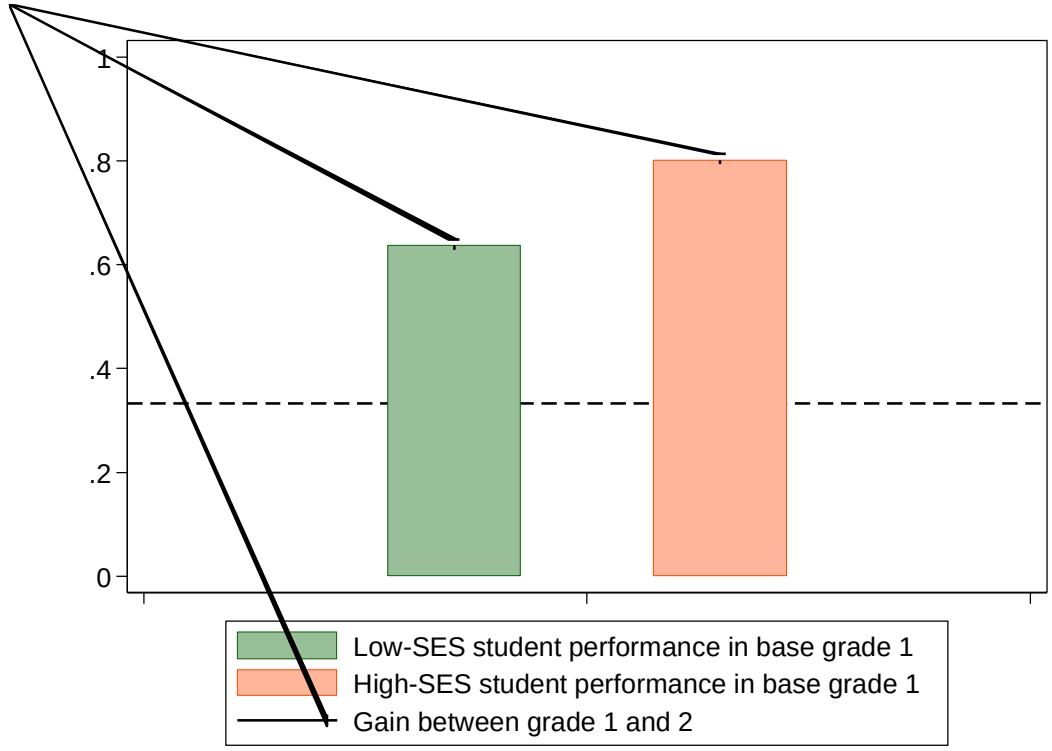
In this section, we discuss the amount and the type of learning that takes place in primary and junior secondary schools in Indonesia. We present our findings graphically and chronologically, starting with 1st grade in primary school and ending with 9th grade in junior secondary school. By comparing all of our results together, we get a sense of what happens across a student's entire career in primary and secondary school. The results on the 10 test questions as well as some breakdowns of these results are presented in Annex 3.[19]

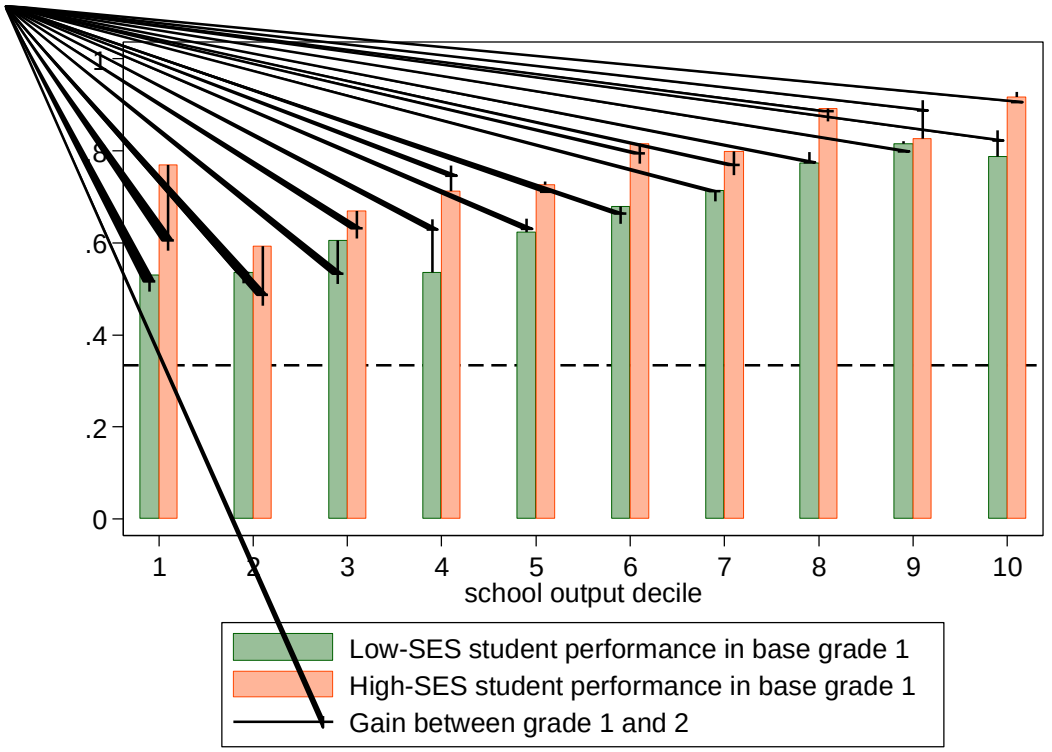
3.1 Primary Schools

The first anchored item was administered to 1st and 2nd grade students in the May 2011 test. Students were asked to circle the number corresponding to the word "seventeen." There were three options -- a. 17, b. 27, or c. 71. The item tested basic reading and numeracy skills. Students need to be able to read the word "seventeen" and link it to the number "17." Figure 3 shows the performance on this question of the low-wealth and the high-wealth students in the top panel with a breakdown across schools in the bottom panel.

Figure 3: What is "seventeen"? Answer: a. 17, b. 27, or c. 71

[19] Table 2.1 (all schools combined), Table 2.4 (selection of schools in the bottom three deciles based on school final output), Table 2.34 (selection of schools in the middle four deciles based on school final output), and Table 2.2 (selection of schools in the top three deciles based on school final output).





It is difficult to distinguish some of the arrows in Figure 3 above. The tiny arrows mean that grade 1 students performed as well on the test item as grade 2 students on average, whereas we find

a difference in the outcomes of low-SES and high-SES students. About 80 percent of the high-SES students could read and recognize two-digit numbers, while only about 65 percent of the low-SES students were able to do so. Notice that the dotted line represents the probability of answering correctly by random guessing among the three options a, b, and c.

Figure 3 suggest that, by the end of 1st grade, some students do not have enough literacy and numeracy to answer this question. What is even more problematic, however, is that none of them seem to acquire these skills in 2nd grade either. This is a first piece of evidence that there is a group of students who do not learn in school. After two years in primary school, these students are already well behind the curriculum pace.

How big is this group of students who struggles with the basics? The top panel of Figure 3 suggests that 20 percent of the high-SES students and 35 percent of the low-SES students do not learn in second grade. However, because these are the results of multiple choice questions, we should consider the fact that students can sometimes answer correctly simply by guessing randomly among the three options a, b, and c. After correcting for this, we estimated that 30 percent of high-SES students and about 50 percent of the low-SES students did not know the correct answer.[20] Therefore, it is clear that the group that did not know the basics by the end of 2nd grade is big.

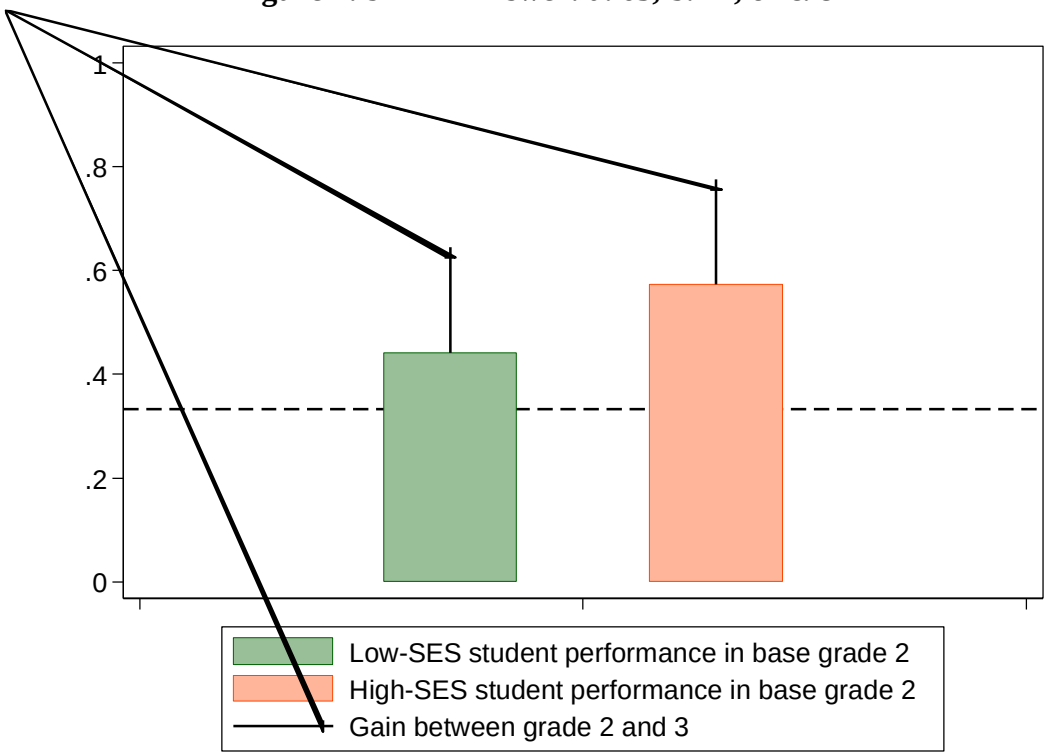
The bottom panel of Figure 3 shows that low performance in 1st and 2nd grade is strongly associated with poor performance at the end of primary school (average performance at the end of 6th grade). This is not surprising perhaps, but it is still worth recognizing that lack of learning in earlier grades strongly predicts bad outcomes at the end. The bottom panel also indicates that some really poor learning takes place in some of the low-output schools. In the bottom three deciles, for

[20] In order to define the share of students who knew the answer to the question, we defined y as the fraction who answered correctly (which is not the same as the percentage of students that knows the answer). The fraction who scored correctly equals $y = \alpha + (1 - \alpha)\frac{1}{K}$, where α is the fraction of students that truly knows the answer, and K is the number of multiple choice options. So the fraction of students who answered correctly equals the fraction α of students that truly knows, added to the fraction $1 - \alpha$ of students that did not know, but guessed correctly with probability $\frac{1}{K}$. With $K = 3$ and $y = 0.65$, we obtained $\alpha = 0.48$. In other words, based on these assumptions, we estimate that 48 percent of the low-SES students knew the answer, despite a total of 65 percent answering correctly.

example, about 60 percent of students answered this question right. Correcting for the fact that random guessing yields a success rate of approximately 33 percent, we concluded that only 40 percent really knew the answer to this question at the end of 1st grade or had learned it in 2nd grade. It is difficult to imagine what happens in these under-performing schools when such large percentages of students do not seem to learn even the basics, but it is clear that, when students cannot read or recognize numbers by the end of second grade, they are not prepared for the next steps in learning.

Figure 4 presents the results from the next anchored question, in which 2nd and 3rd grade students were asked to answer $9 \times 7 = \cdots$. Again, three multiple choice options were provided -- a. 63, b. 72, and c. 81. This question is very straightforward and would appear in almost any 2nd or 3rd grade curriculum in the world.

Figure 4: 9 x 7 = Answer: a. 63, b. 72, or c. 81



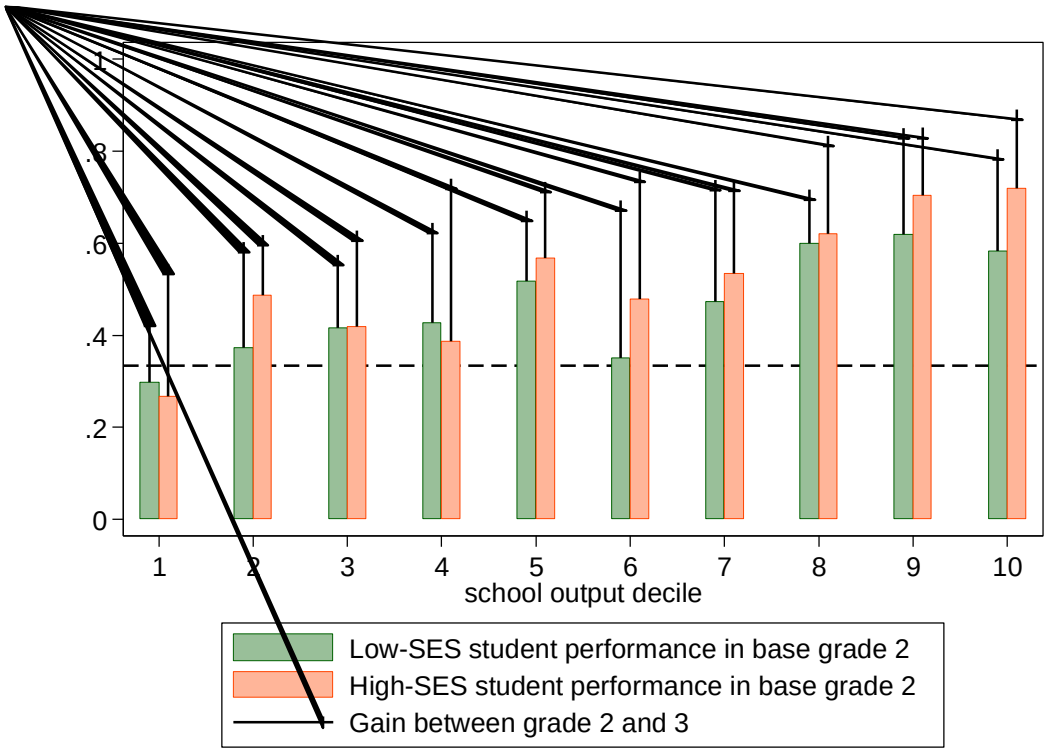


Figure 4 looks quite different than Figure 3 as much more learning has taken place, because grade 3 students tend to score better than grade 2 students. The fraction of students who have learned how to multiply one-digit numbers in 3rd grade is roughly similar across the different output deciles.

The output deciles differ mainly in terms of when students learned to do this. For example, in high-output schools, many students had already learned this skill in (or even before) 2nd grade**.**

The results presented in Figure 4 are consistent with those presented in Figure 3. By the end of 2nd grade, more students could recognize numbers than could do one-digit multiplication. Also, we see that the number of students who could do one-digit multiplication at the end of 3rd grade roughly equals the number of students who could read and recognize numbers by the end of 2nd grade. This suggests that only those who can recognize numbers by the end of 2nd grade are those who will learn one-digit multiplication in 3rd grade. Those who were severely behind in 2nd grade do not seem to catch up.[21] At the same time, however, there was no indication that the group that was behind (in other words, more than two years behind curriculum pace) was significantly increasing in number. To summarize, we observed that some learning takes place in all schools, even in low-output schools, but not all students are learning the curriculum material and those who are too far behind do not seem to catch up. These findings are consistent with the hypothesis put forward by Pritchett and Beatty (2012), which was that students who are on track or not too far behind are learning, and those who were behind stay behind.

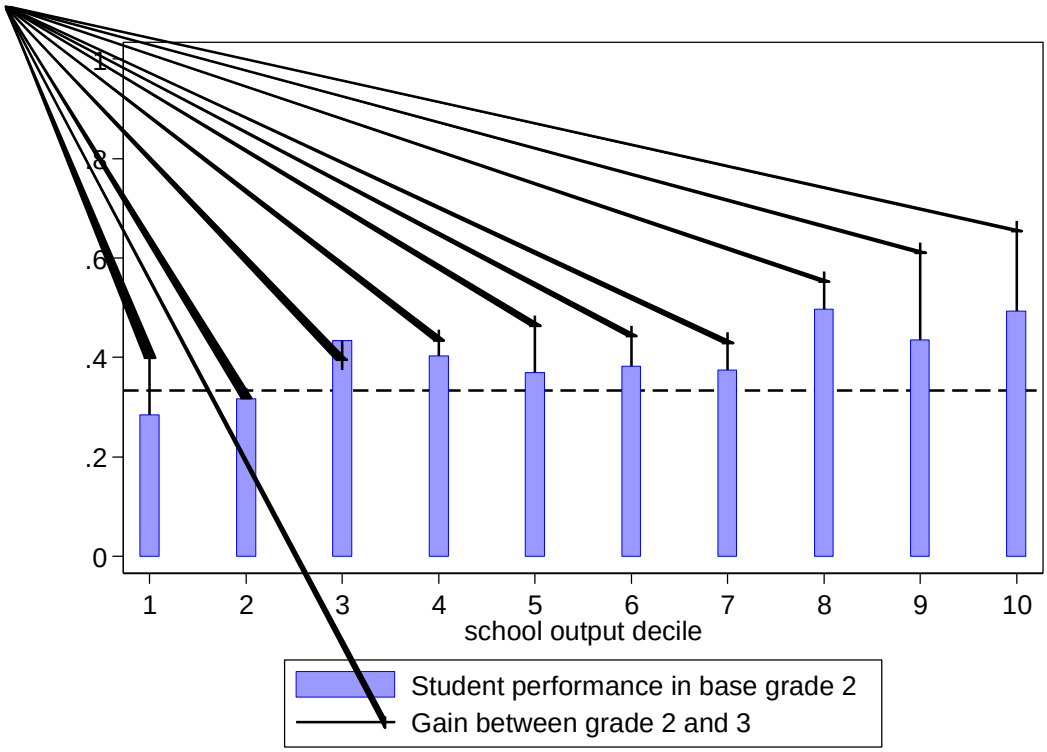
Another interesting phenomenon is shown in Figure 4. It seems that high-SES and low-SES students in the same schools are learning at pretty much the same speed. The top panel shows that the better-off students were more likely than the poor to know one-digit multiplication. The bottom panel shows that most of this is due to the fact that the poor were in different schools than the well-off (see Figure 1), whereas the poor and the more affluent students who attended the same schools were learning at about the same speed. This suggests that it is mainly schools that make the difference in learning outcomes, with differences in student environment, conditional on being in the same school, having much less impact. One caveat to this interpretation is that poor students in high-output

[21] Note that we make some assumptions here. The students who could not recognize numbers in the 2nd grade were not the same students who were asked to answer $9 \times 7 = \cdots$. We essentially made the assumption that the population of 3rd graders is comparable to the population of 2nd graders on average, except for the difference in age.

schools still have different characteristics from poor students in low-output schools. Future analysis is needed to look deeper into the significance of these differences.

Figure 5 also compares students in 2nd and 3rd grades but incorporates another dimension of achievement. In this question, the students were presented with a story. The story went like this: A factory produces 415 sheets, and 252 of these are sold to another factory, with the rest sold at the market. How many sheets are sold at the market? This question involved three-digit subtraction, which is part of the 2nd grade curriculum in Indonesia and of the 2nd and 3rd grade curriculum in the other countries that we studied. The extra element here was the story around it.

Figure 5: A factory produces 415 sheets, 252 of which are sold to another factory, with the rest sold at the market. How many sheets are sold at the market? Answer: a. 163, b. 553, or c. 263

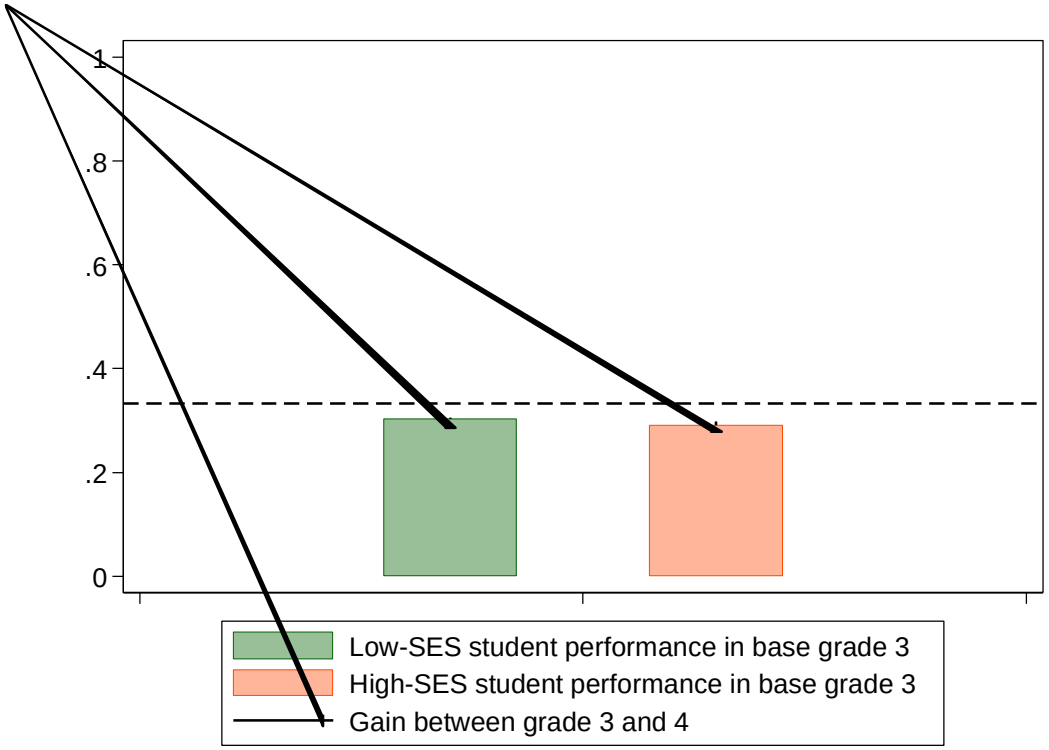


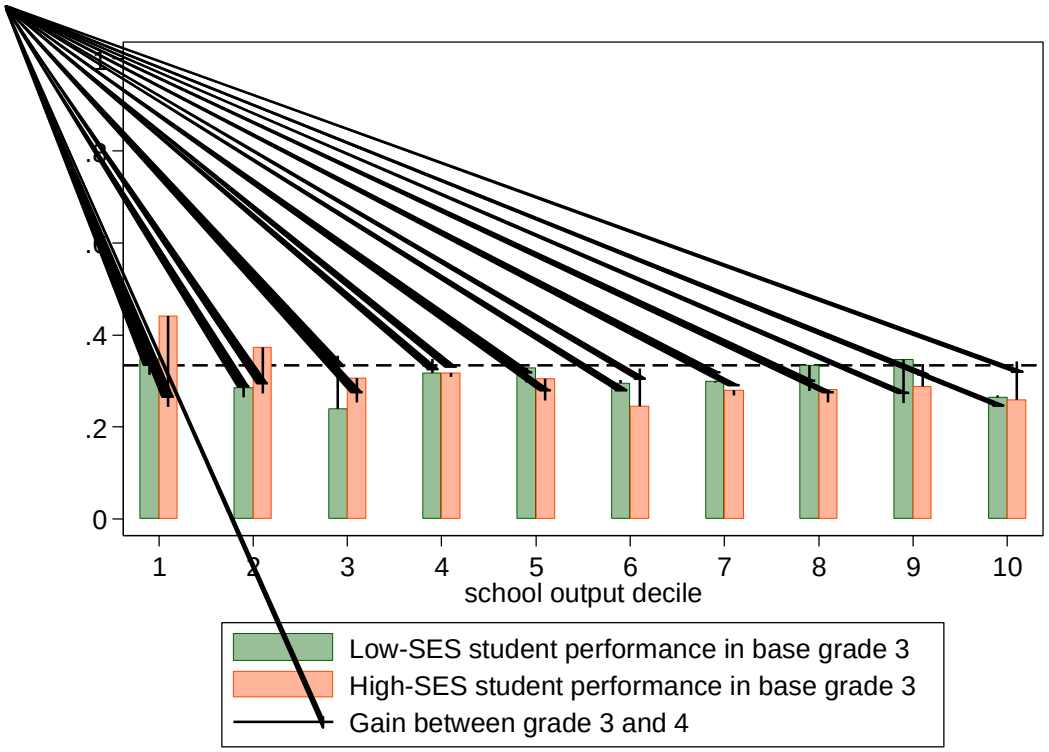
In Figure 5, we cannot present the high-SES/low-SES breakdown as we did not collect this information for 1st and 2nd grade students in the May 2012 survey. However, the overall picture is clear. Only a minority of students in the first seven deciles knew how to approach this question (their

performance is barely above the 33 percent line). Many students in the top three deciles also found this question difficult. This shows that there are many different dimensions of ability and that students do not necessarily develop to the same degree in all of them. The fact that students find it difficult to condense a description of a real-world situation to a mathematical problem indicates that Indonesian schools do not focus enough on this aspect. This is particularly disturbing given that PISA emphasizes that the usefulness of mathematical skills lies in a person's ability to apply them to real-world situations.

Now we move on to comparing 3rd and 4st grade. Figure 6 presents the results of a question about the order of operations in mathematics. The students were asked to answer $216 + 64 - 16 \times 2 = \cdots$. If you know the rules, this question is not difficult: $216 + 64 - 16 \times 2 = \cdots$ is the same as $216 + 64 - 32$ which is the same as $216 + 32$ and the answer is 248. The results, however, clearly show that practically no 3rd and 4th grade student knew the order of operations. The average performance of these students was no greater than the level that would have been achieved by random guessing. We found no differences between high-output and low-output schools or between richer and poorer students.

Figure 6: 216 + 64 – 16 x 2 = Answer: a. 528, b. 312, or c. 248

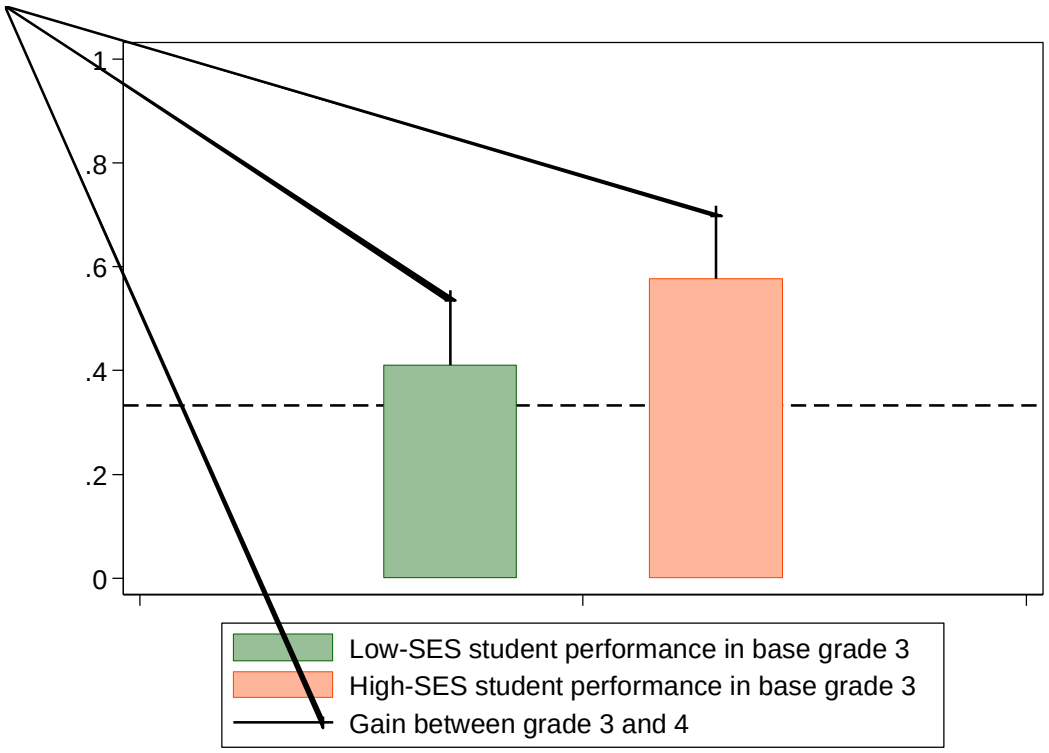


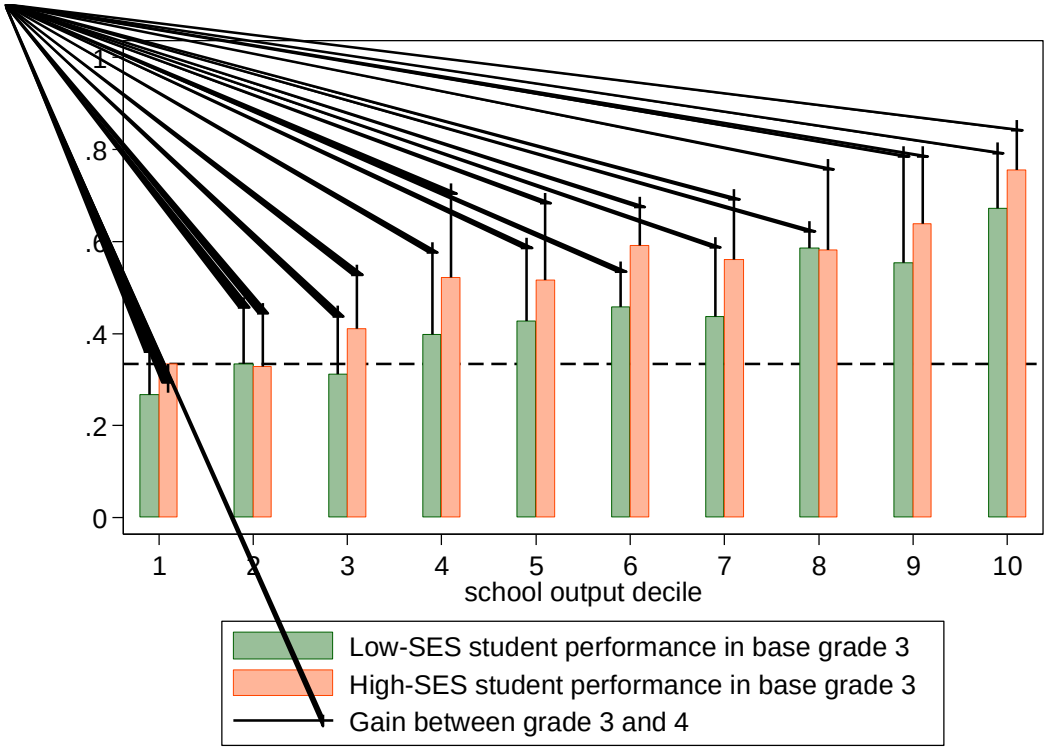


Students in 3rd and 4th grade were also asked to rank the following four-digit numbers from big to small: 2,418, 2,501, 2,470, and 2,465. This question was testing a really basic skill, the ordering and recognition of numbers. Similar to the question that was used to compare 1st and 2nd graders (what is "seventeen"?), this question was designed to be used to benchmark student achievement on very basic competencies. There were three options given without any obvious possibility for misinterpretation. The multiple choice options were: a. 2,501, 2,470, 2,465, and 2.418; b. 2,501, 2,465, 2,470, and 2,418; or c 2,501, 2,418, 2,465, and 2,470.

Again, we found that students in the bottom three school-output deciles struggled greatly, with only a minority of them being able to order four-digit numbers after four years in primary school. We also observed substantial percentages of students struggling with this in mid-range schools, especially the low-SES students. However, as before, students in schools in the top three deciles performed reasonably well. While we did observe some socioeconomic differences within the deciles, most of the variation was still between individual schools and deciles. Thus, it would seem that it is better to be low-SES in a high-output school than high-SES in a low-output school.

Figure 7: Arrange these numbers from biggest to smallest: 2,418, 2,501, 2,470, and 2,465. Answer: a. 2,501, 2,470, 2,465, 2,418; b. 2,501, 2,465, 2,470, and 2,418; or c. 2,501, 2,418, 2,465, and 2,470



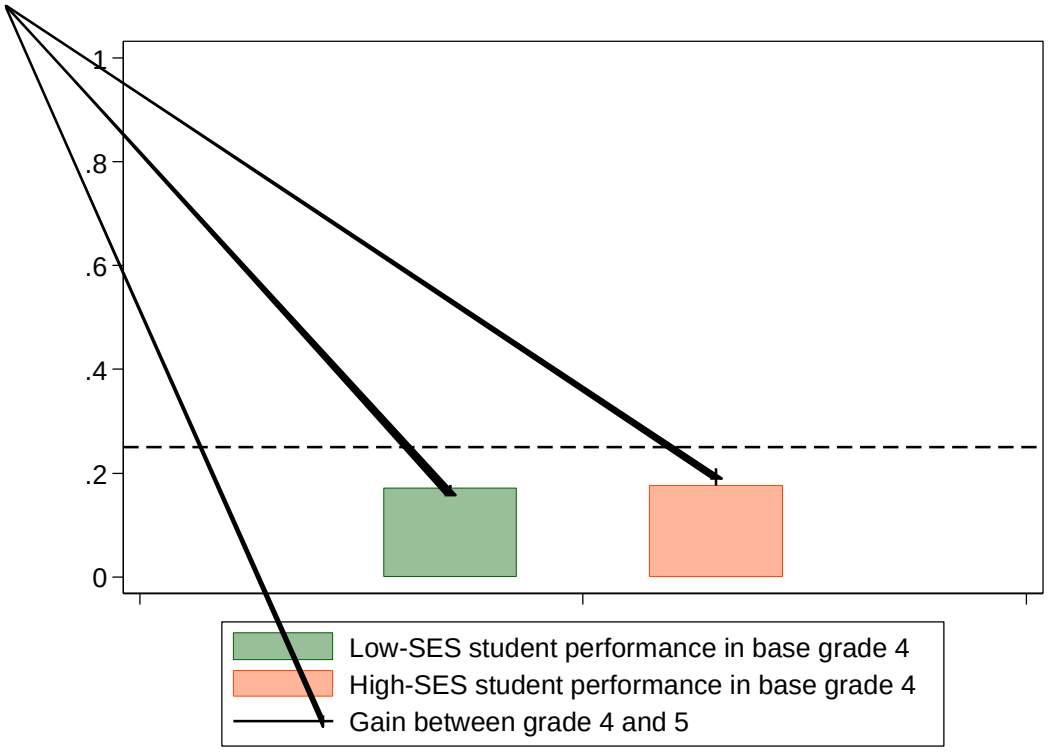


It is interesting to consider Figure 3 and Figure 7 together. We found a significant percentage of students who lacked very basic competencies early on in their schooling careers (in 1st and 2nd

grade) while the same percentage was struggling with basic competencies by the end of 4th grade (though at least the percentage did not go up). As we had already discovered, the size of this group of poor learners was substantial, and poor learning was strongly concentrated in the relatively low-output schools. (This conclusion is not obvious. It might have been that schools had similar average scores and that most of the variation was within schools.)

The test items that we discuss in the remainder of this paper are more complex in the sense that students need multiple skills to answer them correctly. For example, students in 4th and 5th grade were asked to calculate the surface area of a right-angled triangle, of which the two short sides measured 8 centimeters and 12 centimeters with a picture provided. Students needed to know that the surface of this triangle was half of the surface area of a rectangle with sides that measured 8 centimeters and 12 centimeters. Then, they needed to know that the surface of a rectangle is calculated by multiplying the length of the two sides, and finally, they needed to be able to multiply 8 and 12. The solution to this problem is $\frac{1}{2} \times 8cm \times 12cm = 48cm2$. Four possible answers were given: a. 20cm^2, b. 48cm^2, c. 94cm^2, and d. 96cm^2.

Figure 8: Calculate the surface of a right-angled triangle with two short sides of 8 cm and 12 cm (picture of triangle is provided). Answer: a. 20 cm^2, b. 48 cm^2, c. 94 cm^2, or d. 96 cm^2



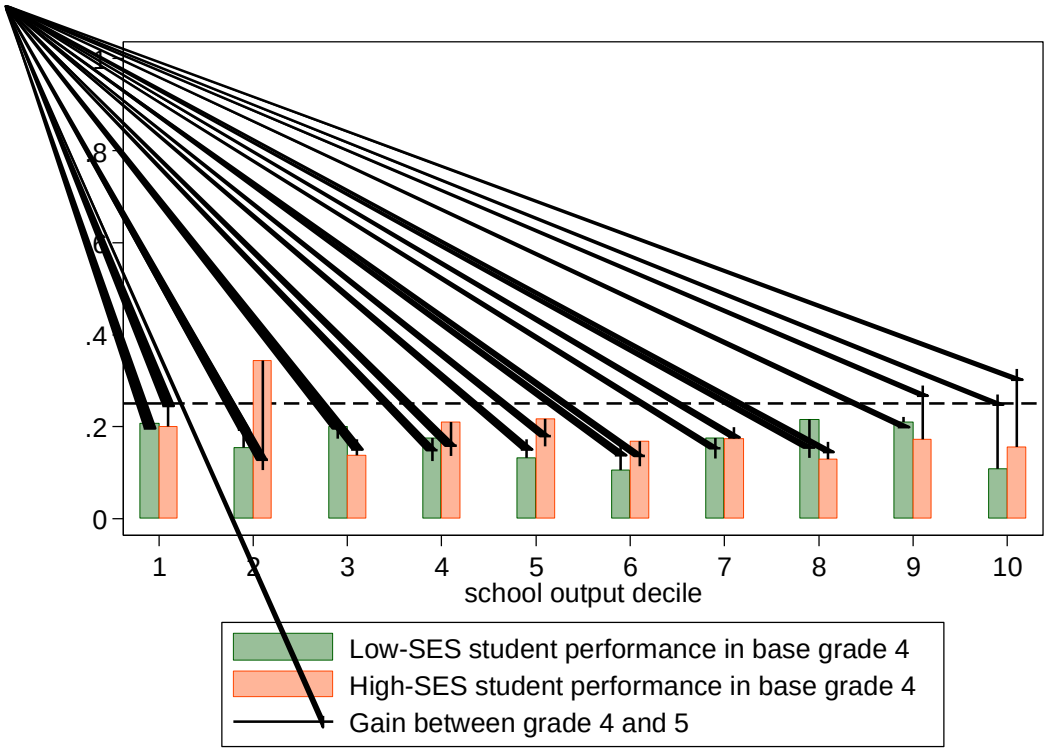
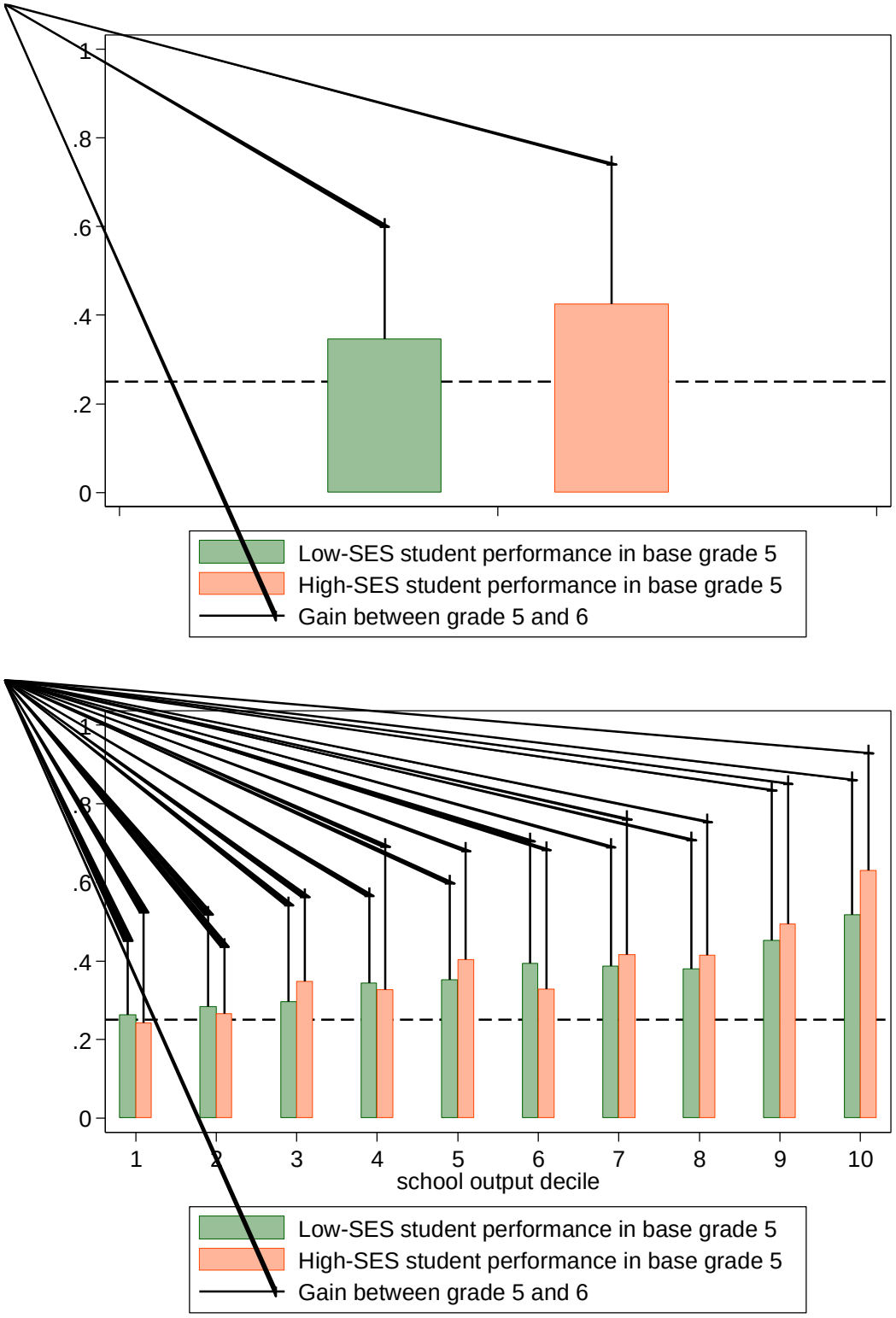


Figure 8 indicates that practically no student knew this by the end of 4th grade, and that hardly any (except a few in the high-output schools) had learned this in 5th grade. Moreover, lots of students performed below the random guessing threshold. This suggests that students are more likely to think of the wrong answer than of the right one. Indeed, we saw that in 4th grade the majority (56 percent) opted for answer a. 20, which is the sum of 8 and 12. In grade 5, we saw a shift towards answer d, but 40 percent still chose a. and 35 percent opted for d. The correct answer b. was the second least popular answer after option c, which only got 6 percent. Fifth graders, therefore, were different from 4th graders. In grade 5, some realized that answers a. and c. were wrong, but even they still tended to select another wrong answer -- d.

Since calculating the surface of a triangle requires multiple skills, we were curious about which skills are lacking.[22] The next question that we discuss shines some light on this. The test asked 5th and 6th graders to calculate the volume of a cuboid with sides that measured 8 centimeters, 18 centimeters, and 12 centimeters. The right answer is $8cm \times 18cm \times 12cm = 1{,}728cm^3$. The options provided were: a. 1,738cm^3, b. 1,728cm^3, c.1,638cm^3, or d. 1,628cm^3. The results of this question are presented in Figure 9.

[22] More research is needed into students' relative performance on different dimensions of ability and skill.

Figure 9: Calculate the volume of a rectangular cuboid with sides measuring 8cm, 18cm and 12cm. Answer: a. 1,738cm^3, b. 1,728cm^3, c. 1,638cm^3, or d. 1,628cm^3.



We observed some of the strongest learning gains between grades on this question, but we also observed that 5th graders are better at calculating the volume of a cuboid (some could do it) than

calculating the surface of a right-angled triangle (nobody could do it). This implies that at least some students were able to calculate the surface area of a rectangle (a skill needed to calculate the volume of a cuboid). This in turn implies that nobody knew that the surface of a right-angled triangle is half of the surface of a rectangle, which is the skill that was lacking. Therefore, it seems that the concept of calculating the surface area of right-angled triangles is not being properly discussed in any primary schools in Indonesia.

More interesting perhaps are the strong gains that we observed between the 5th and 6th grades. So many students answered this question correctly by the end of 6th grade that it seems that even a portion of those who could not arrange four-digit numbers by the end of 4th grade had caught up and effectively learned some geometry and could do somewhat complex multiplication by 6th grade. Something seems to be special about 6th grade. These results might indicate that schools and teachers feel the pressure of the national exams approaching at the end of 6th grade and that this pressure is felt even in the low-output schools.[23]

We learn from this that schools, and even the low-output schools, can produce significant learning, but many schools, especially these low-output schools, start way too late to make up for lost time. The national exams might yield the outside pressure that is not normally felt in the earlier years, but the pressure is not strong enough to induce schools to intensify their teaching earlier. This result (in combination with our previous findings) suggests that it is not a lack of teacher ability or skills that is causing the learning crisis because, when pushed hard enough, all schools can produce learning. It seems unlikely to us that the weak learning gains in low-output schools are caused primarily by a lack of teachers' knowledge and skill.

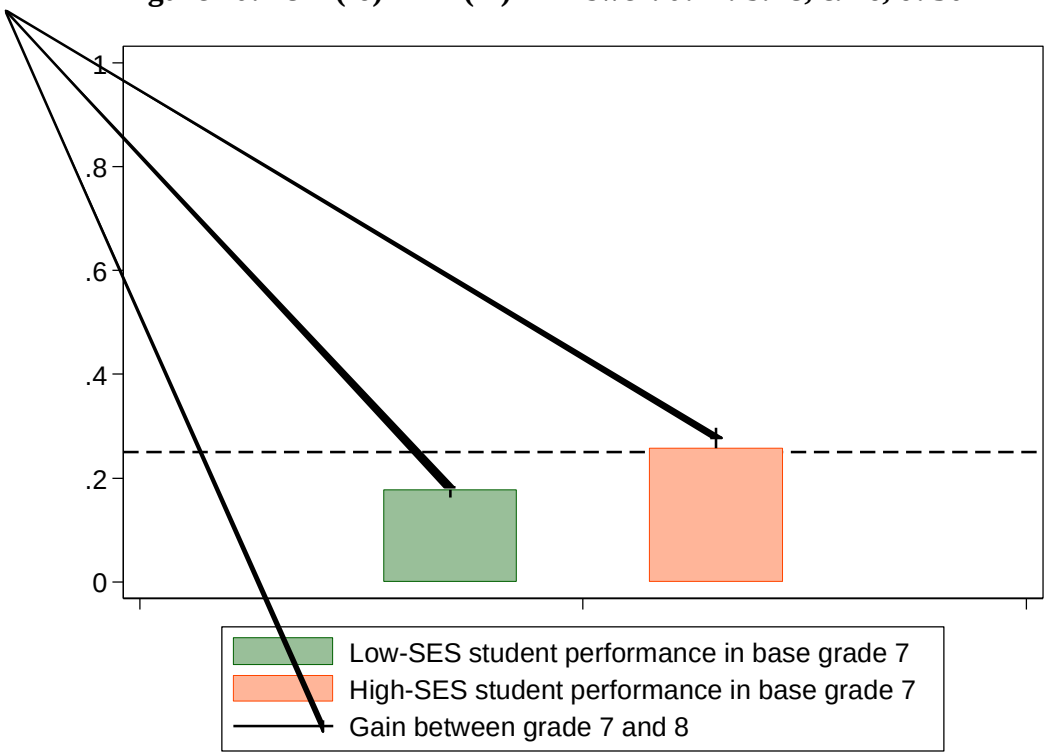
[23] For the breakdown into low-output and high-output schools, we used the test scores of 6th grade students collected in 2011 (the midline data set). Figure 9 uses endline data collected in 2012. This way we did not select on the outcome score for the breakdown

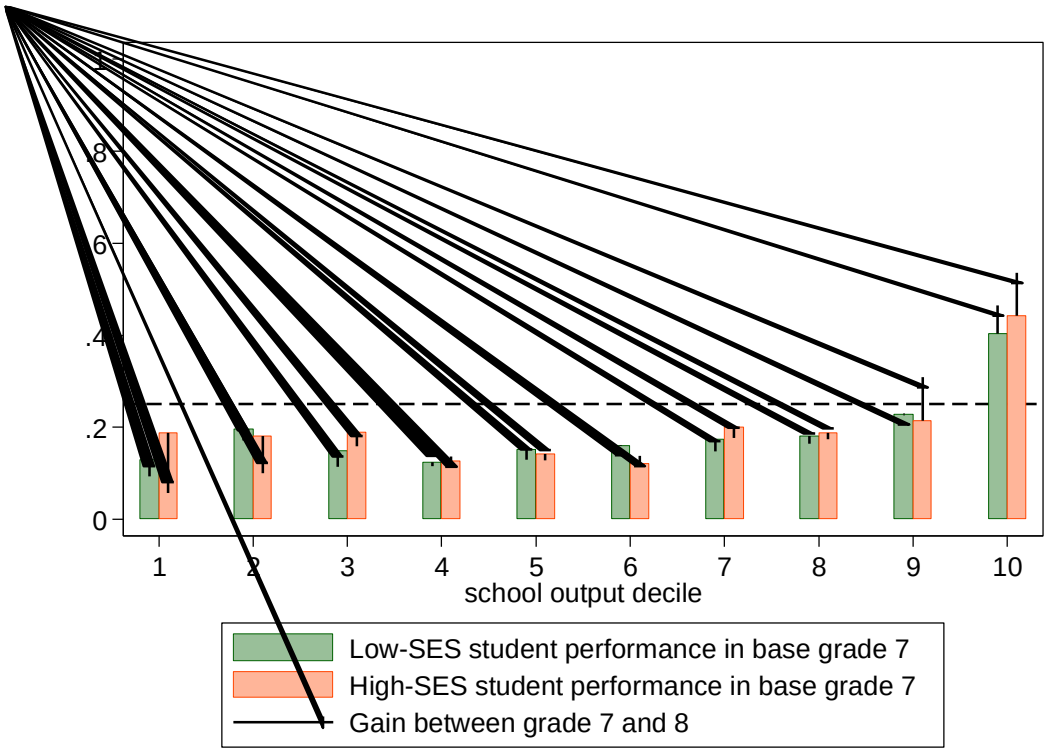
3.2 Junior Secondary School

After primary school, most students in Indonesia continue to junior secondary school. Our analysis so far has shown that students are not well-prepared for the increasingly complex concepts that they will be faced with at the junior secondary level. We can be brief about our findings for junior secondary students as their test scores are particularly poor. In the test administered to 7th and 8th graders, they were asked to solve $18 + (-6) \times 4 - (-2) = \cdots$. This question requires knowledge about the order of operations in mathematics and knowledge about dealing with negative numbers. This material (order of operations and negative numbers) are part of Indonesia's primary school curriculum, but we have seen that 3rd and 4th graders, by and large, do not know the order of operations, particularly that multiplication comes before addition and subtraction.

Students' performance in 7th and 8th grade was particularly poor as can be seen in Figure 10. We observed more or less the same results for 7th and 8th grade as for 3rd and 4th graders. In 80 percent of the junior secondary schools, the average scores were below the benchmark for random guessing. Only in the top 10 percent of schools did we observe that students understood these concepts, albeit still a minority. An interesting discovery was that the differences between low-SES and high-SES were almost entirely a between-school phenomenon. The low-SES students in the high-output schools were learning at the same pace as the high-SES students.

Figure 10: 18 + (-6) x 4 – (-2) = Answer: a. -4. b. -8, c. 46, d. 50

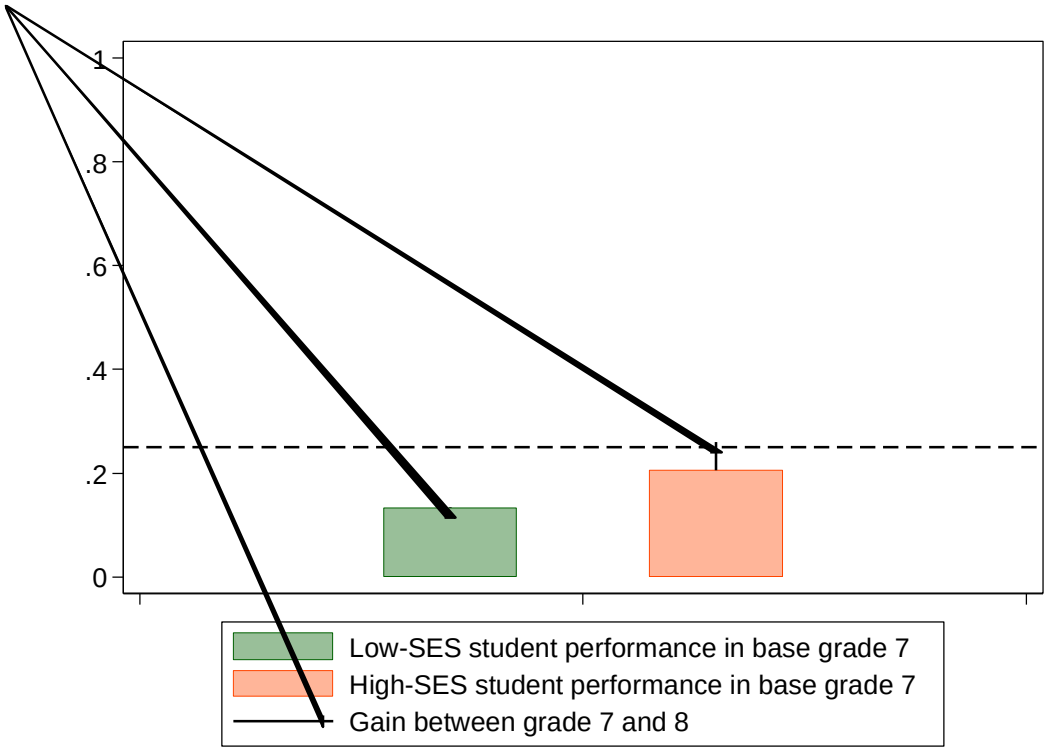


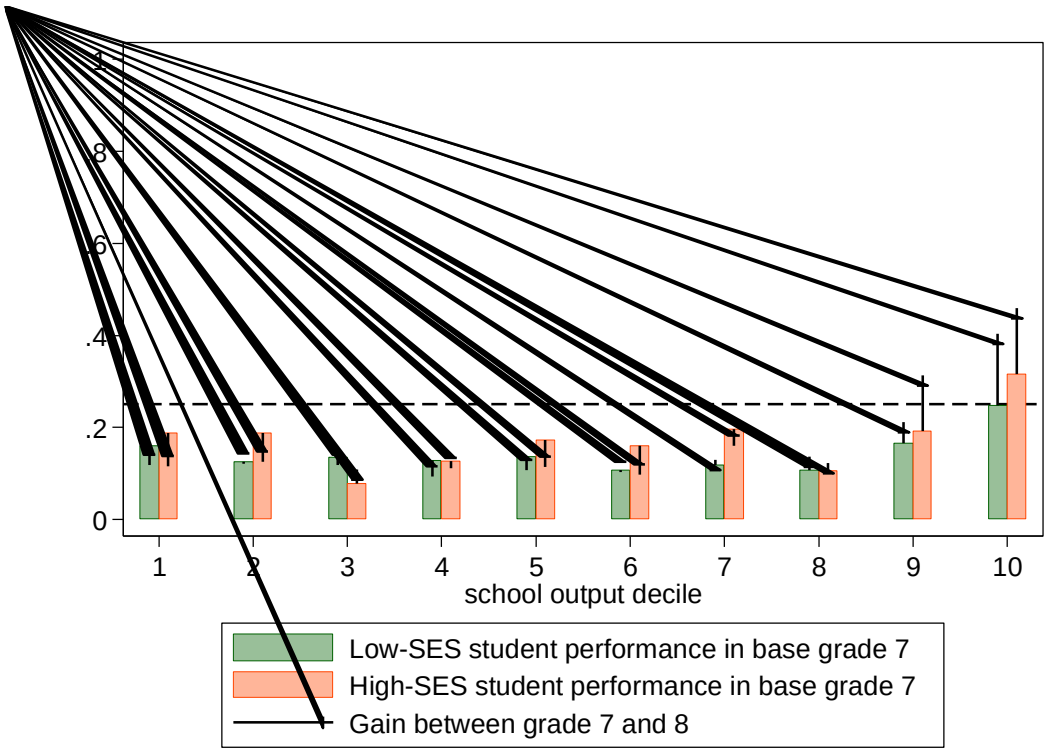


Note: Mean scores below the benchmark for random guessing (the dashed line) indicate that the students who did not know the right answer did not guess randomly but favored one of the wrong answers over a random pick.

The same 7th and 8th grade students were also presented with a "story problem." The story went like this: A merchant sells pens at 2,400 IDR a piece and earns a 20 percent profit. For how much does the merchant buy pens per dozen? The mathematics behind the exercise was again part of the primary curriculum. As the merchant makes a 20 percent profit, he buys each pen for 2,000 IDR. A dozen of them, then, cost 24,000 IDR. However, it is possible that the students did not know what a profit or a dozen was, even though units for quantity are also part of the primary curriculum. On this exercise, we observed very poor performance overall, and only some learning in the top schools. Just as with the previous question, we found that students in the bottom 80 percent scored below random guessing on average.

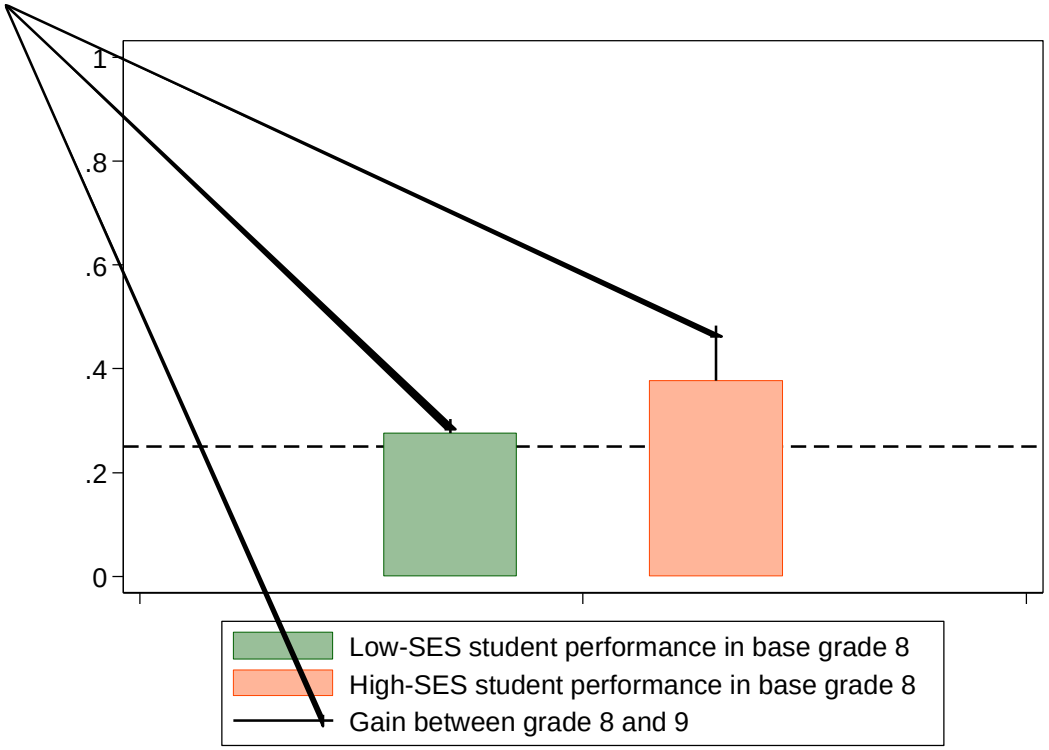
Figure 11: A merchant sells pens at 2,400 IDR a piece and earns a 20 percent profit. For how much does the merchant buy pens per dozen? Answer: a. 2,000 IDR, b. 2,880 IDR, c. 24,000 IDR, or d. 34,560 IDR

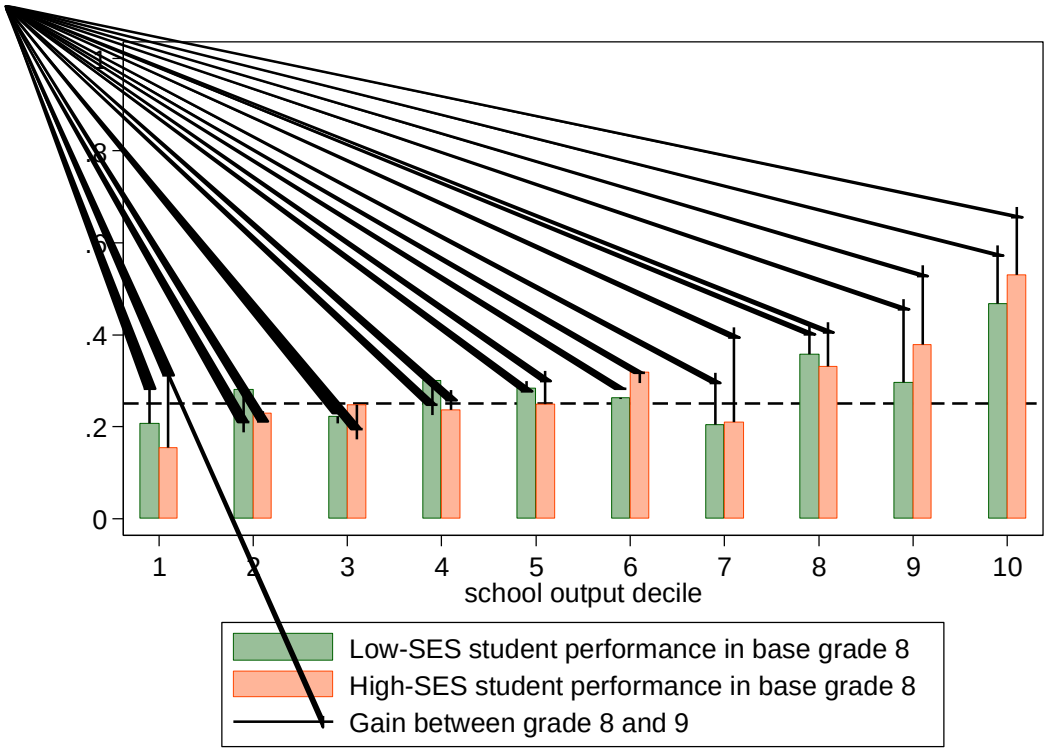




The last question compared performance between students in 8th grade and 9th grade, the last grade of junior secondary school. The question that they were asked was particularly technical and required them to combine different mathematical concepts. The students were presented with a picture of a circle with two points A and B on the edge of the circle highlighted, as well as the center of the circle O. The radius of the circle was 14 cm, and the angle $\angle AOB$ was 120 degrees. The students were then asked to calculate the distance between A and B along the edge of the circle. Furthermore, they were told that the π could be approximated by $\frac{22}{7}$. Four multiple choice options were provided: a. 14.67cm, b. 29.33cm, c. 88cm, or d. 205.33cm. The right answer could be arrived at by calculating the circumference of the circle, $2\pi r = 2 \times \frac{22}{7} \times 14cm = 88cm$ and then dividing by 3 to arrive at 29.33, answer b. The results are presented in Figure 12.

Figure 12: Calculate the distance between two points A and B on the edge of a circle with a 14cm radius. The two points make an angle $\angle AOB$ of 120 degrees with the circle's center O, i.e. (a picture was provided, also $\pi = \frac{22}{7}$). Answer: a. 14.67cm, b. 29.33cm, c. 88cm, or d. 205.33cm.





As with the earlier questions we found very poor performance by students overall, and only some reasonable (but not great) performance from those in the top 20 schools. Here also, we found that any variations in performance between students with different socioeconomic statuses were mainly a between-school issue. The low-SES students in the top schools did better than most of the high-SES students in Indonesia.

The questions were not particularly easy, especially considering the low skill levels with which students tend to leave primary school, but the questions generally reflected what is specified in the Indonesian curriculum. Determining elements, sections, and sizes of circles, for example, is a standard competency for 8th grade students. The government's intention, therefore, is for a large share of Indonesian 9th graders to be able to calculate the distance between two points A and B on the edge of a circle.

One concern about the results for junior secondary schools is that the low stakes nature of these tests might give the students' little motivation to perform well on them. Gneezy et al (2017) have shown that motivation matters on low stakes tests and that they can be context dependent. For example, motivation to do well on low stakes tests may decrease with age. However, this explanation does not seem to be the whole story because the reliability scores of the entire May 2012 9th grade math test is around 0.7,[24] which is not great but also not just measuring noise.

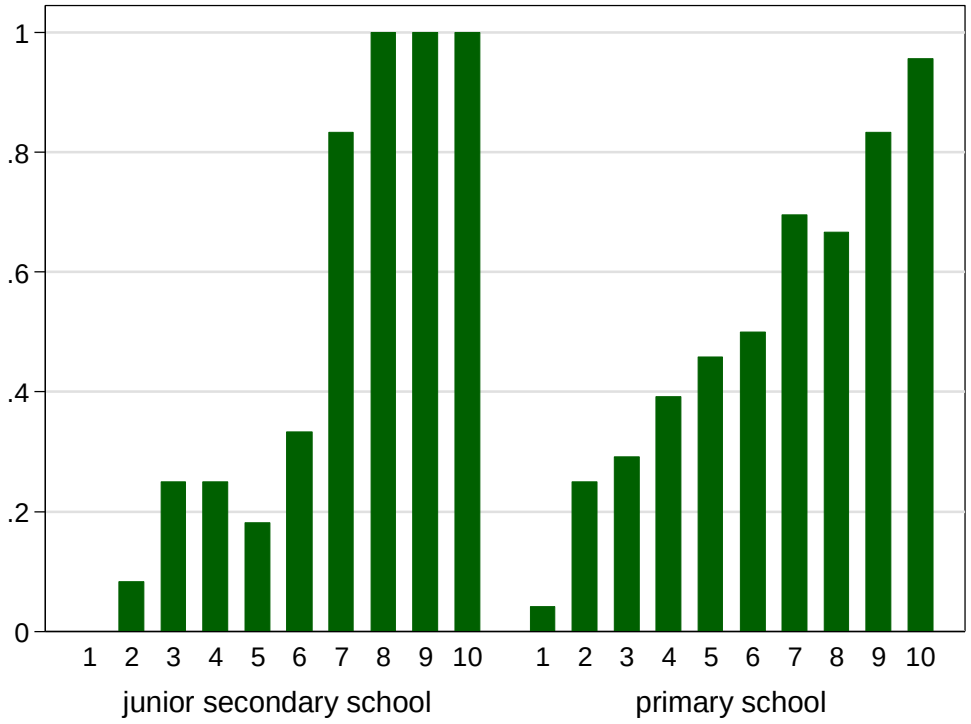
3.3 School Characteristics and More Empirical Results

How are low-output schools different from high-output schools other than having different test scores? The main economic center of Indonesia is the island of Java. Java is home to about half of

[24] We calculated "split-half" reliability by grouping all test items into even-numbered items (items 2, 4, 6, etc.) and odd-numbered items (items 1, 3, 5, etc.). For the two sets of results, we calculated two raw scores (in other words, fraction correct). The correlation between these two raw scores was a reliability estimate (high correlation means highly reliable). We used the Spearman-Brown Prophecy formula to rescale the split-half estimate to reflect the reliability of the entire test (even and odd numbered items combined).

Indonesia's population. Figure 13 presents the fraction of schools per output decile that are located on Java.

Figure 13: Schools on Java by School Output Decile



The figure shows quite strikingly that high-output schools are heavily concentrated on Java, especially junior secondary schools. This puts some of our earlier results in perspective. By far the most learning takes place on Java compared with the peripheral islands.

We also looked at gender differences in learning. Table 3.1 in Annex 4 compares the performance of boys and girls, both overall and within schools. We found that girls tended to do significantly better than boys, although not always, and whether they did depended on what concepts were being tested.

4. Summary and Implications for Policy

This paper presents the first systematic analysis into the learning profiles of Indonesian primary and junior secondary school students. It provides an insight into what Indonesian students learn in a year in school and what they do not. Some patterns have emerged from our analysis.

There is a substantial group of students (about 40 percent to 50 percent) who did not master the most basic skills specified in the Indonesian curriculum (recognizing and ordering numbers). These skills are essential for future learning. However, we did find that some learning takes place in almost all schools, even in the low-output schools, though in those schools, only a very few students (fewer than two-thirds) learn anything at curriculum pace or with a one-year delay.

Some concepts that appear in the curriculum seem not to have been discussed. Very few students in Indonesia had learned to calculate the surface area of a triangle by the end of 5th grade, basic rules about the order of operations (for an exercise using addition, subtraction, and multiplication) by the end of 4th grade, or an exercise using addition, subtraction, multiplication, and negative numbers by the end of 8th grade. Also, we found that most Indonesian students had difficulties with exercises in which the mathematics was embedded in short stories.

There was a notable catching-up effect towards 6th grade in primary school when schools, teachers, and students were preparing for the national exams[25]. While about 50 percent of the 4th graders could not arrange four-digit numbers from big to small, 60 percent of the 6th graders could calculate the volume of a rectangular cuboid (a box). This involved some knowledge of geometry, and a multiplication exercise $8 \times 18 \times 12 = \cdots$, the result of which was a four-digit number. This finding suggests that about 20 percent of the 50 percent of students who could not arrange numbers at the end of 4th grade had caught up tremendously in just two years' time.

[25] The possibility of teachers to allow cheating to increase student's scores during the test is very low as the data is directly collected by the study team. Therefore, systematic cheating (across all regions and field teams) can be ruled out with high levels of certainty.

We found that high-output schools were generally located on Java and that low-output schools were (generally) on the other islands. There is substantial heterogeneity in learning across Indonesia and, especially outside Java, there seems to be a major learning crisis. A related finding was that the differences in learning outcomes between students from different socioeconomic groups were mainly a between-school phenomenon. Within schools, the low-SES and high-SES students had similar levels of performance. The low-SES students in the top schools did much better than the high-SES students in the low performing schools, and vice versa. Also, in most dimensions, girls performed better than boys. And these differences are often quite sizable (up to 9 percentage points).

Based on our results, we have one potentially effective policy recommendation. Poor achievement levels overall in combination with the 6th grade catching up effect suggests that schools, teachers, and students (in any particular order) are sensitive to outside pressure. The pressure provided by the national exams, however, only affects behavior in the last years of primary school, so much of the catching up effect starts too late for most students. Perhaps due to the absence of any performance pressure in the early years, schools appear to be slow-starting. Therefore, outcomes might be improved by monitoring students' performance in the early grades as well.[26] We think that expanding the testing and monitoring system by creating incentives for teachers and schools to reach learning goals *in each grade* could have substantial learning effects.

It would obviously be a challenge to implement a system that incentivizes performance in a way that is practical, cost-effective, and in accordance with the Indonesian system and Indonesian culture. Further study will be needed and perhaps some pilot experiments to find the best way to do this. The most low-cost solution would be for education policymakers to communicate clear intermediate learning goals to schools, teachers, and parents. Better communication might be necessary because, at this point, it is not clear whether all stakeholders know, for example, that one-

[26] Pritchett (2013) has also argued that education systems need to be performance pressured. Also, Muralidharan and Sundararaman (2011) and Woessman (2011) have shown more rigorously that performance incentives can improve learning outcomes.

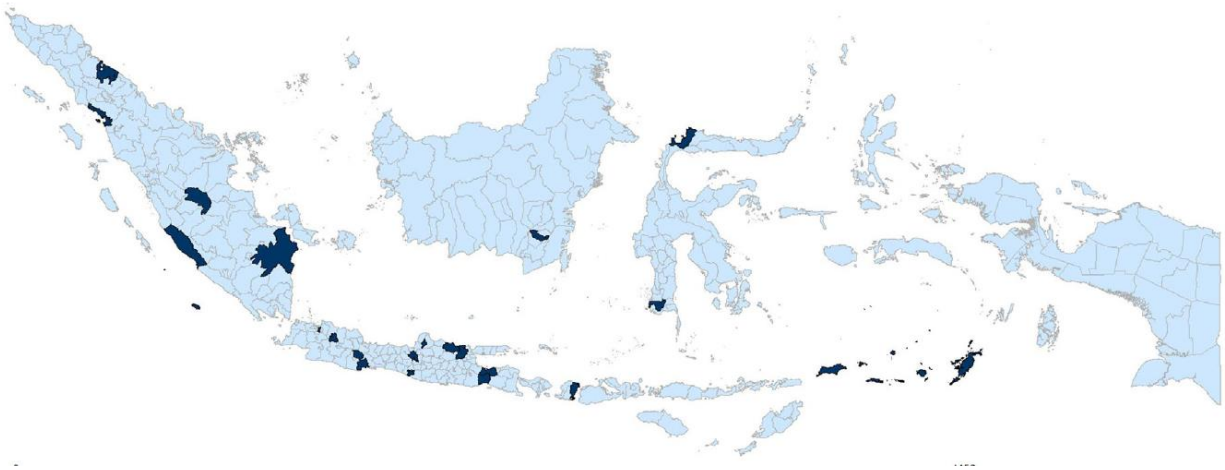
digit multiplication is a learning goal for 2nd and 3rd graders. In addition, teachers may need to keep learning logs for all of their students in order to document their progress.

Annex 1:

Table.1.0 Strata and Sampled Districts

Strata	Sampled Districts
Eastern Indonesia (Maluku and Papua)	Maluku Tenggara Barat
Nusa Tenggara	Lombok Timur
Western Java	Ciamis, Jakarta Timur, and Purwakarta
Central Java	Bantul, Kudus, and Semarang
Eastern Java & Bali	Lamongan, Lumajang, Probolinggo, Tuban
Kalimantan	Hulu Sungai Selatan
Sulawesi	Gowa, Toli-toli
Northern Sumatera	Deli Serdang, Tapanuli Tengah
Western Sumatera	Tebo
Southern Sumatera	Bengkulu Utara, Ogan Ilir

Figure. 1.0. Map of sampled districts



Annex 2: Validity Check

We compared students' performance on tests between grade levels and generally interpreted any positive differences between grades as evidence of learning. This inference is not appropriate if the grade *g* and the grade *g + 1* populations differ in any other ways than just age. In Table 1.1 in Annex 1, we compare asset holdings between the grade *g* and grade *g + 1* populations of students. The data are broken down by grade level (first row, all deciles) and by grade level and school-output decile (rows 2 to 11). There are occasional differences in asset levels on average for the breakdown based on school output, but these differences are not very systematic and relatively small. However, they do suggest that we should be careful when interpreting learning gains for particular output deciles and that we should rather look at the big picture across these output deciles (which is what we do in the main text).

Table 1.1: Testing Similarity of Population Characteristics between Grade Levels

	grade 1 - 2 (midline)		grade 2 - 3 (endline)		grade 2 - 3 (midline)		grade 3 - 4 (endline)		grade 3 - 4 (endline)	
	grade 1	Δ	grade 2	Δ	grade 2	Δ	grade 3	Δ	grade 3	Δ
all deciles	3.95	0.02	4.11	.	3.98	-0.03	4.08	0.01	4.14	-0.05
1	2.61	-0.03	2.83	.	2.50	0.15	2.90	-0.07	2.76	0.08
2	3.41	-0.16	3.46	.	3.33	0.10	3.46	0.21	3.46	0.21
3	3.61	0.00	3.65	.	3.61	-0.13	3.64	0.05	3.65	0.04
4	3.71	-0.11	3.76	.	3.57	0.19	3.69	0.14	3.82	0.01
5	3.79	0.08	3.97	.	3.97	-0.06	3.88	0.16*	4.07	-0.02
6	4.14	0.20**	4.33	.	4.35	-0.34***	4.30	-0.11	4.36	-0.16*
7	3.91	0.01	4.11	.	3.96	0.00	4.11	0.02	4.12	0.02
8	4.24	-0.07	4.40	.	4.19	0.01	4.44	-0.12	4.36	-0.05
9	4.66	0.12	4.97	.	4.79	-0.05	4.95	-0.05	4.99	-0.09
10	4.57	0.06	4.83	.	4.67	-0.09	4.71	0.02	4.95	-0.23**
	grade 4 - 5 (midline)		grade 5 - 6 (endline)		grade 7 - 8 (midline)		grade 7 - 8 (midline)		grade 8 - 9 (endline)	
	grade 5	Δ	grade 5	Δ	grade 7	Δ	grade 7	Δ	grade 8	Δ
all deciles	3.97	0.05*	4.13	0.01	4.10	-0.02	4.10	-0.02	4.16	0.02
1	2.76	-0.13	2.98	-0.08	2.51	0.09	2.51	0.09	2.63	0.28
2	3.26	0.08	3.52	0.06	3.19	0.07	3.19	0.07	3.20	0.17*
3	3.60	-0.01	3.81	-0.04	3.28	0.07	3.28	0.07	3.25	0.22***
4	3.67	0.02	3.81	-0.07	3.67	-0.05	3.67	-0.05	3.73	0.01
5	3.74	0.26***	3.89	0.22**	3.82	0.06	3.82	0.06	3.90	0.10
6	4.16	0.03	4.35	-0.04	3.38	-0.04	3.38	-0.04	3.48	-0.05
7	4.09	-0.03	4.25	-0.11	4.03	-0.19**	4.03	-0.19**	3.95	-0.19**
8	4.38	-0.18**	4.39	-0.06	4.31	-0.05	4.31	-0.05	4.44	-0.09
9	4.77	0.22***	4.93	0.12	4.54	0.02	4.54	0.02	4.64	-0.03
10	4.60	0.17**	4.74	0.13*	5.26	0.03	5.26	0.03	5.29	0.06

Notes: *p < 0:10, p < 0:05, p < 0:01. Table 1.1 reports test of equality of the asset index between adjacent grade levels, and for the selected set of data that was used to make the figures in the main text. There are 10 sets of tests, matching the 10 math questions we look at in the main text. The asset index was constructed as the sum of seven dummy variables indicating asset holdings of the following assets: tv, fridge, mobile phone, bicycle, motorcycle, car, and computer. The fourth column in the top panel does not report estimates, because for endline data we did not collect the asset information for 1st and 2nd grade students.

Annex 3: Results for Reference

Table 2.1: Average Student Performance on the Linked Items

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				raw scores on the linked items				corrected scores		
	# multiple choice options	grade g	grade $g+1$	grade g score	grade $g+1$ score	Δ	p-value	corr. grade g score	corr. grade $g+1$ score	corr. Δ
what is seventeen?	3	1	2	0.72	0.72	0.00	0.94	0.58	0.58	0.00
factory makes sheets	3	2	3	0.41	0.50	0.09	0.00	0.11	0.24	0.13
9 * 7 =	3	2	3	0.52	0.72	0.20	0.00	0.27	0.57	0.30
216 + 64 - 16 * 2 =	3	3	4	0.30	0.30	-0.00	0.77	-0.05	-0.06	-0.01
ordering 4-digit numbers from big to small	3	3	4	0.50	0.64	0.14	0.00	0.25	0.46	0.21
surface of a triangle	4	4	5	0.17	0.19	0.02	0.05	-0.11	-0.08	0.03
volume of a rectangular cuboid	4	5	6	0.39	0.70	0.32	0.00	0.18	0.61	0.42
18 + (-6) * 4 - (-2) =	4	7	8	0.21	0.23	0.02	0.17	-0.05	-0.03	0.02
merchant sells pen with profit	4	7	8	0.17	0.19	0.02	0.13	-0.11	-0.08	0.03
length of the arc A-B	4	8	9	0.33	0.39	0.07	0.00	0.11	0.19	0.09

Corrected scores in columns (8)-(10) take into account that students have a chance of guessing correctly when they do not know the answer. It assumes that there is a fraction α of students that knows the answer (answering correctly with 100% probability) and a fraction $1-\alpha$ not knowing the answer, guessing randomly across the multiple choice items. The observed fraction correct $y(\alpha) = \alpha + (1-\alpha)\frac{1}{K}$, where K is the number of multiple choice options. Columns (8-10) reports $\alpha(y) = \frac{Ky-1}{K-1}$.

Table 2.2: Average Student Performance on the Linked Items in the Top Three School-level Deciles

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				raw scores on the linked items				corrected scores		
	# multiple choice options	grade g	grade $g+1$	grade g score	grade $g+1$ score	Δ	p-value	corr. grade g score	corr. grade $g+1$ score	corr. Δ
what is seventeen?	3	1	2	0.84	0.88	0.04	0.01	0.76	0.82	0.06
factory makes sheets	3	2	3	0.47	0.63	0.15	0.00	0.21	0.44	0.23
9 * 7 =	3	2	3	0.65	0.84	0.18	0.00	0.48	0.75	0.28
216 + 64 - 16 * 2 =	3	3	4	0.28	0.30	0.02	0.38	-0.08	-0.06	0.02
ordering 4-digit numbers from	3	3	4	0.65	0.79	0.14	0.00	0.47	0.68	0.21
surface of a triangle	4	4	5	0.17	0.24	0.08	0.00	-0.11	-0.01	0.10
volume of a rectangular cuboid	4	5	6	0.50	0.85	0.35	0.00	0.33	0.81	0.47
18 + (-6) * 4 - (-2) =	4	7	8	0.27	0.33	0.06	0.01	0.03	0.11	0.07
merchant sells pen with profit	4	7	8	0.20	0.28	0.08	0.00	-0.07	0.04	0.10
length of the arc A-B	4	8	9	0.41	0.54	0.13	0.00	0.21	0.38	0.17

Corrected scores in columns (8)-(10) take into account that students have a chance of guessing correctly when they do not know the answer. It assumes that there is a fraction α of students that knows the answer (answering correctly with 100% probability) and a fraction $1-\alpha$ not knowing the answer, guessing randomly across the multiple choice items. The observed fraction correct $y(\alpha) = \alpha + (1-\alpha)\frac{1}{K}$, where K is the number of multiple choice options. Columns (8-10) reports $\alpha(y) = \frac{Ky-1}{K-1}$.

Table 2.3: Average Student Performance on the Linked Items in the Middle Four School-level Deciles

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				raw scores on the linked items				corrected scores		
	# multiple choice options	grade g	grade $g+1$	grade g score	grade $g+1$ score	Δ	p-value	corr. grade g score	corr. grade $g+1$ score	corr. Δ
what is seventeen?	3	1	2	0.70	0.71	0.01	0.57	0.55	0.57	0.02
factory makes sheets	3	2	3	0.38	0.46	0.08	0.00	0.07	0.20	0.12
9 * 7 =	3	2	3	0.47	0.71	0.24	0.00	0.21	0.57	0.36
216 + 64 - 16 * 2 =	3	3	4	0.31	0.29	-0.01	0.58	-0.04	-0.06	-0.02
ordering 4-digit numbers from	3	3	4	0.48	0.65	0.17	0.00	0.22	0.48	0.25
surface of a triangle	4	4	5	0.16	0.15	-0.01	0.56	-0.12	-0.14	-0.01
volume of a rectangular cuboid	4	5	6	0.36	0.69	0.33	0.00	0.14	0.59	0.45
18 + (-6) * 4 - (-2) =	4	7	8	0.15	0.14	-0.01	0.19	-0.13	-0.14	-0.01
merchant sells pen with profit	4	7	8	0.15	0.11	-0.04	0.00	-0.13	-0.18	-0.05
length of the arc A-B	4	8	9	0.26	0.29	0.03	0.25	0.02	0.05	0.04

Corrected scores in columns (8)-(10) take into account that students have a chance of guessing correctly when they do not know the answer. It assumes that there is a fraction α of students that knows the answer (answering correctly with 100% probability) and a fraction $1-\alpha$ not knowing the answer, guessing randomly across the multiple choice items. The observed fraction correct $y(\alpha) = \alpha + (1-\alpha)\frac{1}{K}$, where K is the number of multiple choice options. Columns (8-10) reports $\alpha(y) = \frac{Ky-1}{K-1}$.

Table 2.4: Average Student Performance on the Linked Items in the Bottom Three School-level Deciles

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				raw scores on the linked items				corrected scores		
	# multiple choice options	grade g	grade $g+1$	grade g score	grade $g+1$ score	Δ	p-value	corr. grade g score	corr. grade $g+1$ score	corr. Δ
what is seventeen?	3	1	2	0.60	0.53	-0.06	0.03	0.39	0.30	-0.10
factory makes sheets	3	2	3	0.36	0.38	0.02	0.57	0.04	0.06	0.03
9 * 7 =	3	2	3	0.40	0.57	0.17	0.00	0.10	0.35	0.25
216 + 64 - 16 * 2 =	3	3	4	0.31	0.30	-0.01	0.58	-0.03	-0.05	-0.02
ordering 4-digit numbers from	3	3	4	0.33	0.45	0.11	0.00	0.00	0.17	0.17
surface of a triangle	4	4	5	0.19	0.19	0.00	0.96	-0.08	-0.08	0.00
volume of a rectangular cuboid	4	5	6	0.28	0.53	0.25	0.00	0.04	0.38	0.33
18 + (-6) * 4 - (-2) =	4	7	8	0.17	0.13	-0.04	0.05	-0.10	-0.15	-0.05
merchant sells pen with profit	4	7	8	0.14	0.12	-0.02	0.16	-0.15	-0.18	-0.03
length of the arc A-B	4	8	9	0.24	0.21	-0.02	0.38	-0.02	-0.05	-0.03

Corrected scores in columns (8)-(10) take into account that students have a chance of guessing correctly when they do not know the answer. It assumes that there is a fraction α of students that knows the answer (answering correctly with 100% probability) and a fraction $1-\alpha$ not knowing the answer, guessing randomly across the multiple choice items. The observed fraction correct $y(\alpha) = \alpha + (1-\alpha)\frac{1}{K}$, where K is the number of multiple choice options. Columns (8-10) reports $\alpha(y) = \frac{Ky-1}{K-1}$.

Annex 4: Differences in Learning by Gender

Table 3.1 compares learning outcomes on the 10 math questions between boys (column 1) and girls (column 2).

Table 3.1: Performance by Gender

	(1) raw score, boy	(2) raw score, girl	(3) difference	(4) difference (school fixed effects)
what is seventeen?	0.71	0.74	-0.04***	-0.03***
factory makes sheets	0.45	0.54	-0.09***	-0.08***
9 * 7 =	0.67	0.76	-0.09***	-0.09***
216 + 64 - 16 * 2 =	0.30	0.29	0.01	0.01
ordering 4-digit numbers from	0.63	0.66	-0.03***	-0.02
surface of a triangle	0.21	0.18	0.03***	0.02***
volume of a rectangular cuboid	0.66	0.75	-0.08***	-0.08***
18 + (-6) * 4 - (-2) =	0.22	0.23	-0.01	-0.01
merchant sells pen with profit	0.20	0.18	0.02***	0.03***
length of the arc A-B	0.39	0.39	0.00	0.00

*p < 0:10, p < 0:05, p < 0:01. Raw scores in columns 1 and 2 are the fractions of students answering correctly.

References

ASER (2014). “Annual Status of Education Report,” Pratham.

De Ree, J., K. Muralidharan, M. Pradhan, and H. Rogers (2018). “Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia,” *Quarterly Journal of Economics.*

Evans, D., and A. Popova (2016). “What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews,” *World Bank Research Observer*, 31(2).

Gneezy, U., J. A. List, J. A. Livingston, S. Sadoff, X. Qin, and Y. Xu (2017). “Measuring Success in Education: The Role of Effort on the Test Itself,” NBER Working Paper 24004.

Government of Indonesia (2016). “Laporan Rapid Assessment: Beban Administrasi: Per-spectif Guru, Kepalah Sekolah dan Pangawas,” Government of Indonesia, Jakarta.

Kaffenberger, M., and L. Pritchett (2017). “More School or More Learning? Evidence from Learning Profiles from the Financial Inclusion Insights Data,” RISE Working Paper 17/012.

Kompas, February 24, 2016

Muralidharan, K. and V. Sundararaman (2011). “Teacher Performance Pay: Experimental Evidence from India,” *Journal of Political Economy*, 119 (1).

PASEC (2014). “Education System Performance in Francophone Sub-Saharan Africa.” Programme d’analyse des systemes educatifs de la confemen,

Pritchett, L. (2013). “The Rebirth of Education: Schooling Ain’t Learning.”

Pritchett, L., and A. Beatty (2012). “The Negative Consequences of Overambitious Curricula in Developing Countries,” CDG Working Paper 293.

Singh, A. (2017). “Learning More with Every Year: School Year Productivity and International Learning Divergence.”

Woessman, L. (2011): “Cross-country evidence on teacher performance pay,” *Economics of Education Review*, 30 (3).

World Bank (2013). “Spending More or Spending Better: Improving education financing in Indonesia,” World Bank, Washington D.C.

_____ 2015): “Indonesia: A Video Study of Teaching Practices in TIMSS Eighth Grade Mathematics Classrooms,” World Bank, Washington, D.C.

_____ (2016): “Indonesia: Teacher Certification and Beyond: An Empirical Evaluation of the Teacher Certification Program and Education Quality Improvements in Indonesia,” World Bank, Washington D.C.

_____ (2018): “World Development Report 2018: Learning to Realize Education’s Promise,” World Bank, Washington D.C.