Public Disclosure Authorized

“Semua Bisa Pintar”:
Introducing a Growth Mindset to Indonesia’s Youth

A Behavioral Insights Experiment
Activity Completion Report

Updated as of 23 July 2020

EXECUTIVE SUMMARY

A behavioral experiment was conducted between October 2016 and June 2018 to test the impact of a "growth mindset" lesson plan on learning outcomes. The experiment was designed based on a prior World Bank activity in Peru, which had a positive impact on learning outcomes through a simple and minimally-intrusive set of classroom activities. The original Peru intervention was improved with additional materials and delivery methods, based on the latest insights in the behavioral economics literature. Adjustments were also made to fit an Indonesian context.

The intervention was designed to support the Government of Indonesia's "Revolusi Mental" (mental revolution) campaign. The "Revolusi mental" campaign was developed to promote positive social values and norms such as integrity and perseverance. This national campaign manifested in the Ministry of National Education's priorities through the inclusion of a "character building" component of the national education curriculum. This intervention supports the character building curriculum by testing how to best instill beliefs and behaviors such as growth mindset, grit, and perseverance.

The impact evaluation found the intervention to have significant and positive impacts in some, but not all, scenarios. Using a difference-in-difference model, the evaluation found significant and positive impacts of the intervention on UN scores, but not evenly across the board. The impact was only found in schools in Sumatra, with the magnitude of impacts strongest in Mathematics and weakest in Bahasa Indonesia. The impact in Sumatra across all subjects tested in the UN was 0.15 standard deviations; the literature suggests that 0.2 standard deviations is equivalent to an additional year of schooling. This is a substantial impact, especially considering an implementation cost of US$ 0.25 per student. However, no information was collected on other outcomes of interest, such as beliefs and behaviors of students and teachers at baseline and endline, which could have helped to explain the differences between locations and subjects.

This activity has since been followed up with a second experiment that expands the treatment and conducts additional data collection. Following the completion of this activity and presentation of initial results, the Ministry of Education and Culture (MoEC) requested a follow-up experiment that expands the intervention to include additional materials on other socio-emotional skills and extends the implementation time. This follow-up experiment also addressed the challenges faced during this activity's impact evaluation, such as the difficulty in explaining the causes of heterogeneous impacts across island groups, school performance, and academic subjects, by conducting additional data collection to track changes in student beliefs and behaviors as the mechanisms by which the intervention affects learning outcomes.

ACKNOWLEDGEMENTS

This activity was conducted by the WB Indonesia Poverty and Education teams, in partnership with the WB Mind, Behavior, and Development unit (eMBeD). The activity was financed by the Australian Department of Foreign Affairs and Trade (DFAT) through the Partnership for Knowledge-Based Poverty Reduction Trust Fund (PKPR). Activities under PKPR are focused on supporting the Government of Indonesia (GoI) in reducing inequality in Indonesia, and are endorsed by the team's key stakeholders at the National Team for the Acceleration of Poverty Reduction (Tim Nasional Percepatan Penanggulangan Kemiskinan, TNP2K) and the Ministry for National Development Planning (Badan Perencanaan Pembangunan Nasional, Bappenas).

I. INTRODUCTION

I.1. COUNTRY CONTEXT[1]

Though government spending on education is high, student learning outcomes are low. Since a 2002 amendment of the Indonesian constitution, the state has been required to allocate 20 percent of the national budget on education. However, despite this high level of spending, Indonesia's learning outcomes have been persistently poor. Scores in all three subjects measured in the Program for International Student Assessment (PISA)[2] test – science, reading, and mathematics – have not changed much between 2000 and 2015.

Table 1. Indonesia's PISA scores have been persistently poor over the past 15 years. [3]

	SCIENCE	READING	MATHEMATICS
2000	393	371	367
2003	395	382	360
2006	393	393	391
2009	383	402	371
2012	382	396	375
2015	403	397	386
Change since 2012	**21**	1	11

Source: Indonesia PISA 2015 brief

In the international context, Indonesian students performed worse than regional neighbors and other middle-income countries. In science, Indonesian students obtained an average PISA score of 403 which put Indonesia as one of the lowest performers among all 2015 PISA-participant countries (Figure 1). Meanwhile, other Southeast Asian countries (i.e. Thailand, Singapore, and Malaysia) achieved significantly higher PISA scores in all assessed subjects. Indonesia even lags Vietnam, which has a much lower GDP per capita. Vietnam obtained an average PISA score of 502 and ranked in the top 10 of participant countries.

Aside from performance, Indonesian students also lagged the OECD average and regional neighbors on science-related career expectations. Only 15.3 percent of students reported expecting to have a career in a field that required science skills, which was lower than the OECD average (24.5 percent) as well as regional neighbors Singapore (28 percent), Thailand (19.7 percent) and Vietnam (19.6 percent). This is primarily driven by a lower percentage of Indonesian boys with science-related career expectations (8.6 percent) compared to girls (22.1 percent), whose rates tracked closer to the OECD average.

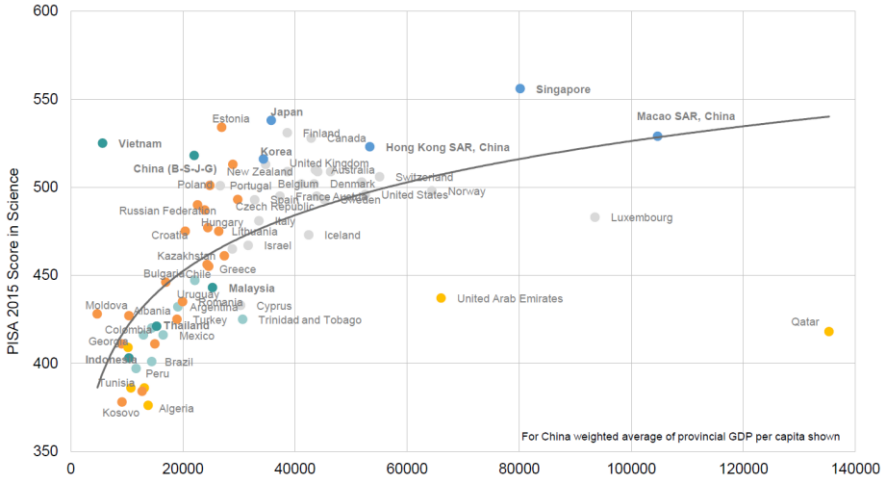
---

[1] This section describes the context in 2017 under which the project was initiated. New information, including from the 2018 PISA round, has become available since then, but is not discussed in this report.

[2] PISA is "a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students". Approximately 540,000 students completed the 2015 PISA assessment, representing 29 million 15-year olds in 72 countries and economies. (http://www.oecd.org/pisa/aboutpisa/)

[3] PISA scores are scaled to have an average of 500 and a standard deviation of 100, with higher scores indicating better performance. Two-thirds of students in OECD countries score between 400 and 600.

Figure 1. Indonesia ranks poorly, even among countries with similar GDP per capita.

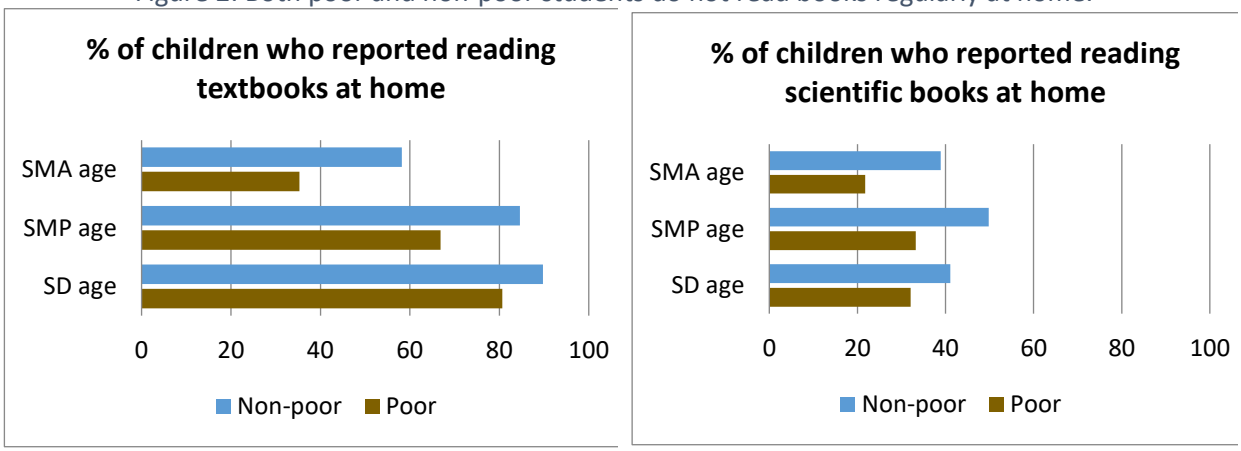


Source: Indonesia PISA 2015 brief

Note: colors are based on region. Gray = OECD, Yellow = Middle East & North Africa, Orange = Europe and Central Asia, Light Green = Latin America and the Caribbean, Dark Green = East Asia and Pacific, Blue = East Asia & Pacific (high income)

Low student motivation is one possible cause of poor academic outcomes. Though the quality of education spending is one possible explanation, it is possible that students are not carrying out the necessary behaviors to maximize their learning outcomes: that is, students are not putting enough effort into learning. According to the Indonesian National Socio-Economic Survey (Susenas) in 2012, only 67 percent of SMP-aged[4] students from poor[5] households reported reading school textbooks at home in the past week, compared to 85 percent of students from non-poor households. The rates drop steeply for SMA-age students, and even further when SMA-age students were asked about scientific books aside from their school textbooks.

Figure 2. Both poor and non-poor students do not read books regularly at home.



Source: Susenas (2012)

[4] The Indonesian system comprises 6 years of Primary School (Sekolah Dasar, SD), 3 of Junior Secondary School (Sekolah Menengah Pertama, SMP), and 3 of Senior Secondary School (Sekolah Menengah Atas, SMA).

[5] The definition of poverty used in this document is based on Indonesia's national poverty line in 2012 of approximately IDR350,000 (US$27) per person per month.

I.2. INTERVENTION CONTEXT

The "Semua Bisa Pintar" (Everyone Can Be Smart, SBP) intervention was developed with the intention to improve student learning outcomes by building a "growth mindset". The "growth mindset" theory (Dweck, 1999)[6], posits that people with "fixed mindsets" – who believe that intelligence is static – will put less effort into learning than those with "growth mindsets", who believe that intelligence can be improved. The SBP intervention would attempt to change student mindsets from "fixed" to "growth" and anticipated that the resulting increase in effort would lead to improved learning outcomes.

The SBP experiment drew on a past World Bank experiment on growth mindset in Peru. The Peru experiment developed a set of classroom activities that were designed to convince students that intelligence is not static and can be improved through enough effort. The intervention, called 'Expande tu Mente!' or Grow Your Mind (GYM)[7], consisted of a 90-minute school session, led by tutor teachers, where students were introduced to neuroplasticity and the malleability of cognitive abilities. In the first 30-minute segment, students were asked to read a two-page text designed to introduce them to the concepts. In the second 30-minute segment, the students were asked to engage in a classroom debate moderated by the teacher. In the final 30-minute segment, students were asked to individually write a reflective letter to a friend or relative describing what they learned in the session.

The GYM results indicated that improving student mindsets could lead to higher effort to learn and persevere through difficult challenges, ultimately contributing to an increase in test scores. The GYM impact evaluation showed that, on average, students experienced increases in test scores of 0.2 standard deviations after being exposed to this intervention. This number is roughly equivalent to having a parent with two to three more years of education. In addition, this intervention successfully reached more than 50,000 students with a relatively low implementation cost of USD 0.20 per student.

The SBP experiment would provide additional evidence of whether this style of "growth mindset" intervention would be effective in other contexts. Peru and Indonesia have many similarities: they are both middle income countries with similar GDP per capita, and they had similarly low scores in the 2015 PISA round. However, the 2015 PISA data shows differences in student beliefs: Peruvian students were more likely to aspire to science-related careers (38.7 percent compared to 15.3 percent). As such, the SBP experiment was a test of whether a GYM-style intervention would be effective outside of the specific context of Peru.

I.3. CLIENT CONTEXT

This was a World Bank-initiated activity that received full endorsement from key Government of Indonesia stakeholders. SBP was initiated by the World Bank Indonesia poverty team as part of a series of Behavioral Interventions projects, focused on supporting the Government of Indonesia to apply the principals of behavioral economics and psychology to induce "vital behaviors", such as exclusive breastfeeding or, in this case, putting more effort into learning at school. This activity received a full

[6] Dweck, C.S. (1999). Self-Theories: Their role in motivation, personality and development. Philadelphia: Taylor and Francis/Psychology Press.

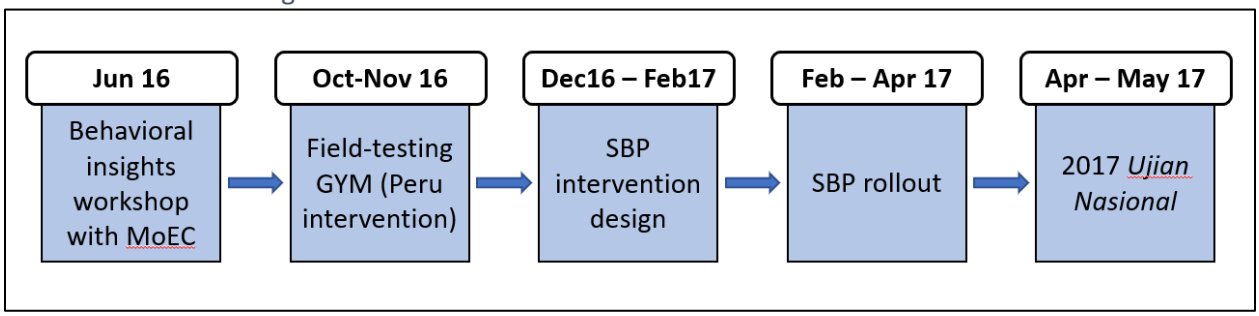
[7] Additional details on the project can be accessed on the World Bank website: http://documents.worldbank.org/curated/en/212351580740956027/pdf/The-Power-of-Believing-You-Can-Get-Smarter-The-Impact-of-a-Growth-Mindset-Intervention-on-Academic-Achievement-in-Peru.pdf

endorsement from the Indonesia Poverty team's Steering Committee and Technical Advisory Committee, which consists of DFAT, TNP2K, and Bappenas.

The intervention was proposed to the Ministry of Education and Culture (MoEC) as a way to support the Government of Indonesia's "revolusi mental" (mental revolution) priority of national character building. The "Revolusi mental" featured heavily in the Joko Widodo administration's first term in office, as a way to promote positive social values and norms such as integrity and perseverance. This national campaign manifested in the Ministry's priorities through the inclusion of a "character building" component of the national education curriculum. The team proposed the SBP package of interventions as a way to support the character building curriculum by testing how to best instill beliefs and behaviors such as growth mindset, grit, and perseverance.

Following an introduction to the behavioral insights field of study, MoEC endorsed the intervention and the RCT format. The team started the engagement by inviting MoEC to a three-day workshop in June 2016 on applying behavioral insights in public policy. In the aftermath of that meeting, a series of follow-up meetings were held to discuss implementing the SBP experiment as the first component in a "curriculum" of sessions designed to develop the socio-emotional skills of students. MoEC fully endorsed the rollout of an experiment involving public SMP schools in Java and Sumatra on a timeline that would conclude with the 2017 Ujian Nasional[8] (National Exams, UN) as detailed in Figure *3*. MoEC agreed with the use of the UN as the metric to assess the improvement in learning outcomes, and technical counterparts at MoEC provided valuable support in convincing schools to participate and accommodating all requests for data.

Figure 3. The intervention timeline took a total of 12 months.



II. INTERVENTION DESIGN

II.1. INITIAL FIELD TESTING

The intervention design started with a field test in October and November 2016. The intervention design started with a process of testing the GYM intervention instruments – the two-page text, classroom discussion, and letter-writing session – in the Indonesian context. Following the delivery of the intervention, the team interviewed students and teachers to get feedback on the implementation. This testing period took place in October and November 2016, in four junior high schools (SMP) in Klaten, Central Java, and four SMP in Malang, East Java.

[8] The Ujian Nasional is a national standardized exam administered to final year students in each schooling level (Grade 6 for primary school, Grade 9 for junior secondary school, and Grade 12 for senior secondary school). More details about the UN are provided in the Impact Evaluation section below (Section V.2.a).

The field test confirmed prior assumptions about the need for the intervention. Student responses to the pre-pilot in Klaten and Malang were positive. In the letters written by students, they reported that they had never been exposed to neuroplasticity and the notion that intelligence is not static. Some reported feeling relieved and motivated because they now knew that they could catch up to the top students in class, whereas previously they felt helpless and "not smart enough" to do it. In the student interviews, many reported believing that intelligence is "something you are born with" and "is not something you can change", with the same applying to athletic or artistic talents. Several interviews with teachers also revealed the prevalence of a "fixed mindset": some teachers felt the intervention would not be useful because "some students just have different IQs, which can't be changed".

However, the field test also revealed a need to modify the intervention. The team's classroom observations found that many students struggled to read and comprehend the long, 1000-word text document, which is consistent with other information about student appetites for reading books in Indonesia. Additionally, even when students were able to read the documents in a reasonable time, they found the concepts too technical and hard to grasp. Finally, the team also became concerned with the ability of teachers, particularly those in rural or remote schools, to deliver the sessions adequately.

II.2. SEMUA BISA PINTAR INTERVENTION

The Theory of Change for this style of intervention is to influence learning outcomes by exposing students to new beliefs. The intervention assumes that classroom exposure to these concepts will influence student beliefs about growth mindset, grit, and perseverance. The addition of concrete examples about actions that students can take will provide them with the tools to implement these new beliefs, causing them to take on improved study habits and behaviors. These changes in beliefs and behaviors are expected to ultimately lead to improved learning outcomes, as measured by test scores.

The initial GYM intervention was modified to better suit the Indonesian context and tackle concerns about delivery by teachers. Based on the results of the field test, the major issues identified were: 1) the two-page text was too difficult for students with constraints in reading pace and comprehension; 2) the concepts were too technical for SMP-level students; and 3) teachers without previous exposure to the concepts would struggle to deliver the sessions well. To address these issues, the team redesigned the intervention to make it simpler for both students and teachers. The key components of the redesign are detailed in Table *2* below.

Table 2. The GYM intervention was modified to simplify the intervention for both students and teachers.

Component	GYM (Peru)	SBP (Indonesia)
Delivery of key concepts	2 page, 1000-word document	Four comic books, 4 pages each
Total session time	A single 90-minute session	Two 40-minute sessions
Session components	Reading text, classroom debate, letter-writing	Reading comics, guided classroom discussions, sharing personal stories, group yells
Teacher preparation materials	None	Detailed scripts for all sessions

All intervention revisions were done in-house, drawing on behavioral insights principles, particularly the MINDSPACE framework. The process of revising the intervention instruments was led by the

Indonesia poverty team, supported by the eMBeD team, and with quality assurance from the Indonesia education team. The revisions were done based on the MINDSPACE framework developed by the Behavioral Insights Team[9]. Examples of how the MINDSPACE framework was used are provided in Table *3* below. The revisions were presented to MoEC over several iterations before finalizing for roll-out.

Table 3. The SBP intervention utilized the MINDSPACE framework extensively

MINDSPACE principle	**Application in SBP intervention**
M – Messenger	The intervention used classroom discussions as a way to show students that their peers are going through and overcoming similar challenges, thus increasing the likelihood that they would believe the message that the challenges could be overcome through a change in mindset and perseverance (as opposed to if a much older teacher communicated the message).
S – Salience	The intervention utilized a series of comic books with simple messages that are easy to remember, in order to draw attention and improve retention. This is the largest change between the initial GYM intervention (a 2-page text) and the SBP intervention. The "group yell" session was also included as a novel way to induce retention of the sessions and key messages.
A – Affect	The plots in the comic books showed situations where the characters were going through emotional reactions to situations that are relatable to the audience, in order to elicit the same emotional reactions and a feeling of relatability to the plot.
C – Commitments	The group photo exercise was included as a form of commitment-building within the classroom, so that students would be more likely to encourage each other to stick to the efforts of changing their mindsets and behaviors.
E - Ego	The letter-writing exercise was included so that students could engage in self-reflection, not only about the materials provided but also about how those changes in mindsets and behaviors might be beneficial for them in the long run.

II.2.a. Comics and videos

The largest redesign was a shift from a long text to a series of comic books to make the messages as simple and attractive as possible. Given the challenges faced by students in the field test, it was clear that the 2-page text needed to be replaced with a simpler medium. The new instrument took the form of four comic books, which featured largely the same key messages about growth mindset, self-belief, and perseverance, but re-packaged them in a way that could be more easily absorbed by students. The comic books used several behavioral principles to convince students of the key messages, such as:

1. Short, simple messages that were repeated several times in large, bold text, e.g.: "otakmu seperti otot" (your brain is like a muscle), "belajar dari bertanya" (learn by asking others);
2. Relatable characters that were designed in the style of modern-day cartoons; and
3. Storylines that put the characters in classroom situations that are familiar to the reader.

Figure *4* below shows a snippet from the comic books that demonstrates these principles. In the snippet, Asri, the female lead, is explaining to Pandu, the male lead, that "intelligence is not a trait of birth, but

[9] https://www.bi.team/publications/mindspace/

something gained through effort". Instead of just saying that line once, this snippet repeats it four times in different ways. First, it delivers the simple messages of "no one is born stupid" in bold text. Secondly, it emphasizes in the same speech bubble that "you are not born smart or stupid, it all depends on effort". Third, there is a visual analogy that compares a person lifting a dumbbell to a brain lifting a dumbbell, complete with the line that "your brain is like a muscle". Fourth, there is an element of humor when Pandu seeks clarification about whether "growing his brain" means that his head will, in fact, get larger.

Figure 4. The comics use simple, repeated messages, relatable characters, and humor.

The four comics explained the growth mindset through a variety of aspects. The first comic, entitled "Otak Itu Seperti Otot" (your brain is like a muscle) conveyed the message that the brain, like a muscle, can get stronger through training, and that no one is born inherently smart or stupid. The second comic, entitled "Belajar Dari Kegagalan" (learning from failure) covered grit and perseverance through the message that failing and trying again is a process of training your brain, so failure is not a negative outcome and should not be feared. The third comic, entitled "Belajar Dengan Bertanya" (learning through questions) initiated a shift in mindset by changing the vocabulary from "tidak bisa" (I can't do it) to "belum bisa" (I can't do it yet), and encouraged the reader to seek help from others and ask questions if they don't understand. The final comic, entitled "Belajar Dengan Menyukai Tantangan" (learning by embracing challenges) explained the science of neural connections as evidence of the benefits of a growth mindset, and encouraged the reader to seek out challenges.

II.2.b. Session structure

The redesign also included a reworking of the classroom sessions to emphasize key messages and simplify delivery for teachers. In order to ensure that the key messages were understood and led to behavioral change, simple classroom discussions would not suffice. Thus, the team designed a lesson plan

that incorporated visual, auditory, and kinesthetic sessions, in order to ensure that students were able to absorb the information in ways that fit them best. The sessions included storytelling, dance movements, and even short "mantras" that the students and teachers could repeat throughout and after the lesson. The letter-writing session from the GYM experiment was also included as part of the second SBP session.

The intervention was split into two sessions implemented one week apart. The decision to split the intervention into two sessions was made for two reasons: 1) to avoid overloading the students with new materials, and 2) to better match the length of classroom periods in Indonesian SMPs and make it easier to implement. Thus, the sessions were split into two 40-minute sessions, which were implemented over two consecutive weeks, for a total of 80 minutes of exposure to the materials, still a similar level of exposure to the GYM experiment in Peru (90 minutes).

Table 4. Two classroom sessions combined a variety of learning methods.

Session 1 Activity	**Length**	**Justification**
Reading Comic 1	5 mins	Introduction to the key concepts through a simple narrative
Key message discussion	2 mins	Guided discussion to ensure students find the key messages
Group discussion about "the brain is like a muscle"	3 mins	Discussing the key concept with peers allows students to reach a shared understanding of the concept
Sharing personal experiences about improving through practice	5 mins	Students engaging in paired discussion to demonstrate to each other (and themselves) that they have gotten smarter and more proficient through practice
Reading Comic 2	5 mins	Introduction to the key concepts through a simple narrative
Key message discussion	2 mins	Guided discussion to ensure students find the key messages
Teacher shares a story of learning from past failure	2 mins	Demonstrating to students that teachers have also learned from failure in the past
Student volunteer shares a story of learning from failure	3 mins	Student volunteer will demonstrate to students that failure is commonplace and can be learned from
Writing exercise: current challenges and frustrations	5 mins	Students are encouraged to think about the challenges they are currently facing through a new lens of "not being afraid of failure"
Group yell	5 mins	Brief song and dance to inspire students

Session 2 Activity	**Length**	**Justification**
Session 1 recap	3 mins	Reminding students of the previous week
Reading Comic 3	5 mins	Introduction to the key concepts through a simple narrative
Key message discussion	2 mins	Guided discussion to ensure students find the key messages
Paired storytelling about asking questions	3 mins	Discussing the key concept with peers allows students to reach a shared understanding of the concept
The "not yet" game	1 min	Students are taught a shift in mindset from "I can't" to "I can't yet"
Reading Comic 2	5 mins	Introduction to the key concepts through a simple narrative
Key message discussion	8 mins	Guided discussion to ensure students find the key messages

Letter writing	10 mins	Students engage in self-reflection about the learnings in the two sessions
Group photo	3 mins	Session is ended on a fun and memorable note, which creates a positive association and improves retention

II.2.c. Teacher instructions

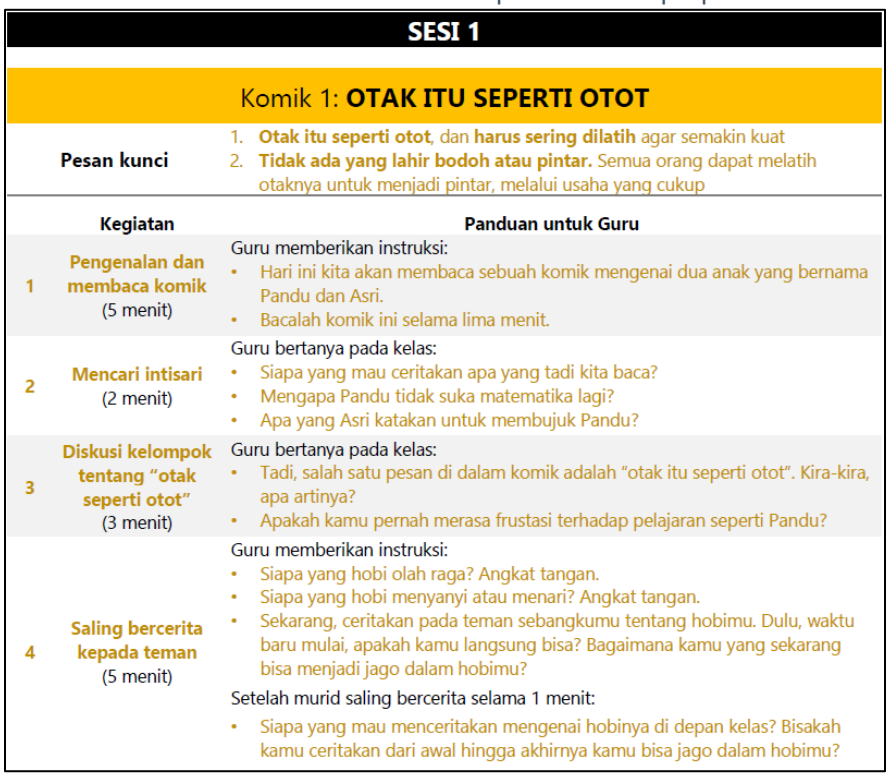
The team recommended that the intervention be implemented by homeroom teachers at a fixed time for each of the two weeks, but allowed schools to make their own decisions about how and when to implement. In order to build a measure of consistency across schools, the intervention package recommended that the sessions be delivered by homeroom teachers for each class, at the same day and time for each of the two weeks (one example suggested was the first lesson on Monday morning). However, the team acknowledged that schools would be constrained in the available timing for delivering this program due to the preparation schedule for the UN exam, and thus gave schools flexibility on when and how the program would be implemented. Specific implementation details were not included in the data collection on compliance, which only focused on collecting whether the intervention was delivered.

A set of simple instruction manuals were designed to simplify the intervention delivery for teachers. The team was concerned with the potential difficulty that teachers would experience when delivering these diverse and complex classroom sessions, especially given that not many teachers would have experience with the messages and the materials, and most would be severely time-constrained. This was evident in the pre-pilots conducted in Malang and Klaten, where teachers had a hard time following the content of the sessions without preparation. Thus, to mitigate the risk, the team prepared a detailed instruction manual with scripts and time stamps, so that teachers could deliver the sessions simply by following the script. The instruction manual was a total of 9 pages long, with the following structure:

- Page 1: cover page and call to action
- Pages 2-3: intervention summary and timeline
- Pages 4-5: session 1 detailed instructions
- Pages 6-7: session 2 detailed instructions
- Pages 8-9: additional information on growth mindset and neuroplasticity

Although preparation time was not completely eliminated, the materials were expected to greatly reduce the total time commitment for teachers. Teachers were asked to spend 30 minutes of preparation time before each session to study the comics and session structure, as well as to think about possible improvisations to the script. Thus, the total time commitment for teachers was 140 minutes for the entire intervention.

Figure 5. Instruction manuals contained detailed scripts to reduce preparation time for teachers.



SESI 1

Komik 1: **OTAK ITU SEPERTI OTOT**

Pesan kunci	1. **Otak itu seperti otot**, dan **harus sering dilatih** agar semakin kuat 2. **Tidak ada yang lahir bodoh atau pintar.** Semua orang dapat melatih otaknya untuk menjadi pintar, melalui usaha yang cukup

	Kegiatan	Panduan untuk Guru
1	**Pengenalan dan membaca komik** (5 menit)	Guru memberikan instruksi: • Hari ini kita akan membaca sebuah komik mengenai dua anak yang bernama Pandu dan Asri. • Bacalah komik ini selama lima menit.
2	**Mencari intisari** (2 menit)	Guru bertanya pada kelas: • Siapa yang mau ceritakan apa yang tadi kita baca? • Mengapa Pandu tidak suka matematika lagi? • Apa yang Asri katakan untuk membujuk Pandu?
3	**Diskusi kelompok tentang "otak seperti otot"** (3 menit)	Guru bertanya pada kelas: • Tadi, salah satu pesan di dalam komik adalah "otak itu seperti otot". Kira-kira, apa artinya? • Apakah kamu pernah merasa frustasi terhadap pelajaran seperti Pandu?
4	**Saling bercerita kepada teman** (5 menit)	Guru memberikan instruksi: • Siapa yang hobi olah raga? Angkat tangan. • Siapa yang hobi menyanyi atau menari? Angkat tangan. • Sekarang, ceritakan pada teman sebangkumu tentang hobimu. Dulu, waktu baru mulai, apakah kamu langsung bisa? Bagaimana kamu yang sekarang bisa menjadi jago dalam hobimu? Setelah murid saling bercerita selama 1 menit: • Siapa yang mau menceritakan mengenai hobinya di depan kelas? Bisakah kamu ceritakan dari awal hingga akhirnya kamu bisa jago dalam hobimu?

III. EXPERIMENT DESIGN

III.1. TREATMENTS AND SAMPLING

The experiment included one treatment group and one control group of Grade 9 students in public SMP schools. Though the team initially considered multiple treatment arms to test certain aspects of implementation, such as which teacher was assigned to deliver the intervention, in the end the team decided to simplify the experiment to a single treatment arm. The sample selection process drew information from multiple administrative data sources on education:

- Dapodik 2016: school level dataset with geographical and institutional information of the educational institutions.
- UN SMP 2015: school level dataset with results of the average scores on the National Examination (*Ujian Nasional*, UN)[10] in 2015 for students in the last grade of SMP, 9th grade.

Additional control variables, such as a "rurality" index which shows the proportion of rural sub-districts, poverty level, and socioeconomic status at the district level were obtained from SUSENAS 2014.

The program selected 1,202 schools for the treatment arm and control group, for a total of 2,404 schools. Only public (*Negeri*) middle schools (*Sekolah Menengah Pertama* – SMP) were targeted, excluding Islamic schools (*Madrasah Tsanawiyah* – MT) and technical or remedial schools (SMPT). The list of schools used for the selection of the sample was obtained by merging the school census (Dapodik) and National

[10] The Indonesian education system features a national standardized exam for all students in the final year of each of the three school levels (SD, SMP, and SMA).

Examination results (UN-SMP) administrative datasets. Once duplicated observations were dropped, this universe of schools consisted of 21,310 units in 34 provinces and 514 districts nationwide.

The population was restricted to the islands of Java and Sumatra due to logistical reasons. Restricting the experiment to Java and Sumatra allowed the team to lower both the time and the cost of shipping. As the main purpose of this experiment was as a proof of concept, the team decided that implementation feasibility was a higher priority than national representativeness. This restriction still resulted in a large population of 15 provinces, 266 districts, and a total of 12,141 of the schools that met the criteria above, roughly 57% of the schools in the country. Because of this large population, it would still allow the team to conduct a second experiment with a nationwide scope in future years, even if the 2,404 schools from this experiment were excluded.

The random sampling was stratified in four categories of schools according to their academic performance, as measured by UN test scores in 2015[11]. On each of the two islands, schools were classified into one of the following groups: low performance and low dispersion schools, low performance and high dispersion, high performance and low dispersion schools, high performance and high dispersion. Performance and dispersion were defined as the average and standard deviation of the within-school UN test scores, and the high and low level groups in each island corresponded to schools over and below the median level of the variables in the island. Table *5* below provides the summary statistics of the two indicators by stratum.

Table 5. Summary statistics by stratum and indicator

		School performance (Average Test Score)			School dispersion (Test Score S.D.)		
	Stratum	**Mean**	**S.D.**	**Obs.**	**Mean**	**S.D.**	**Obs.**
1	Low performance & low dispersion	199	28	2,597	21	5	2,597
2	Low performance & high dispersion	201	21	3,473	36	7	3,473
3	High performance & low dispersion	291	31	3,472	19	6	3,472
4	High performance & high dispersion	275	30	2,599	40	8	2,599
	Total	**242**	**50**	**12,141**	**29**	**11**	**12,141**

Note: high vs. low for both performance and dispersion were defined using the population median.

Due to the limitations of the data provided, the intervention would need to be delivered to all students in Grade 9 of each selected school. Because data was only provided at the school aggregate level, it was not possible to isolate impacts at the student or classroom level. Thus, the experiment had to be delivered to all students in the Grade 9 cohort in order to detect the impact at the cohort level.

III.2. SAMPLE SIZE, STATISTICAL POWER, AND BALANCE

The random sampling and selection were performed at the school level, i.e. the sampling unit was the school within each stratum. However, by stratifying by academic performance categories, the random assignment is guaranteed to be balanced across the four groups. This balance increases the statistical

[11] Several other indicators were tested for the definition of strata, including categories identified by islands, district and municipal rurality, district poverty, school size, and a combination of those. However, the definition by school test score performance was the only category that yielded large differences between strata, which would allow for a more comprehensive heterogeneity analysis.

power of the experiment (Bruhn y McKenzie, 2008), and allows for a heterogeneity analysis between and within strata.

The power calculations assumed a clustered design of the standard errors. The formula to obtain the optimal number of clusters (schools) per treatment arm is shown below:

$$N_c = 1 + \left(t_{1-\alpha/2} + t_{1-\beta}\right)^2 \cdot \frac{\sigma_1^2 + \sigma_2^2}{(\mu_1 - \mu_2)^2} \cdot \left[\frac{1}{m}(1 + (m-1)\rho)\right]$$

Where the significance level is α = 0.05, the power level β =0.8, the test scores variances of the control and treatment groups are standardized and set as $\sigma_1^2 = \sigma_2^2 = 1$, the (standardized) difference in test scores or Minimum Detectable Effect – MDE is $\mu_1 - \mu_2$, the cluster or school size is m = 100 students in 9th grade, and the intra-class correlation – ICC is ρ. Table *6* below provides the sample size per treatment arm obtained by different combinations of MDE and intraclass correlations.

Table 6. Number of clusters (schools) for given values of MDE and intraclass correlation

		Minimum detectable effect size $\lvert\mu_1 - \mu_2\rvert$					
		0.1	**0.11**	**0.12**	**0.13**	**0.14**	**0.15**
Intraclass correlation ρ	**0.45**	717	592	498	425	366	319
	0.50	794	657	552	471	406	354
	0.55	872	721	606	517	446	388
	0.60	950	785	660	563	485	423

Assuming an intraclass correlation between 0.45 and 0.50[12] and a MDE of 0.15, the number of schools necessary to detect a treatment effect on the test scores would be 319 schools per stratum or 1,276 in total for the treatment arm and control group.

Once the number of schools was determined, schools within each stratum were randomly selected and the final sample was proportionally drawn according to strata distribution in the population. Balance checks for treatment and control arms are presented in Table *7* below. As seen in the table, all school and district level variables are balanced between treatment and control and within stratum, as all estimates control for strata group.

Table 7. Balance check between arms

	***t*-test of differences between arms**				
	Treatment (n=1202)		**Control (n=1202)**		***p*-value**
Variable	**Mean**	**S.D.**	**Mean**	**S.D.**	**difference**
Test scores	243.2	51.1	242.44	50.21	0.779
SD test scores	28.79	11.26	28.76	11.09	0.962
Class size	28.47	6.64	28.71	6.57	0.484
School size	135.75	104.34	137.67	109.61	0.738
% Female students	0.49	0.06	0.49	0.06	0.597

[12] ICC values for Indonesia vary between 0.46 and 0.49, according to the latest available PISA test results for Science, Math and Reading at the moment of implementation (2012) https://www.oecd.org/pisa/pisaproducts/PISA%202012%20Technical%20Report_ANNEX%20C%20%E2%80%93%20STANDARD%20ERRORS%20OF%20MEANS,%20SAMPLE%20SIZES,%20SCHOOL%20VARIANCE%20ESTIMATES,.pdf

Internet access	0.54	0.5	0.53	0.5	0.646
PC laboratory	0.46	0.5	0.44	0.5	0.362
Rural district	0.62	0.48	0.61	0.49	0.583
No rural info	0.06	0.24	0.07	0.25	0.588
Java	0.51	0.5	0.49	0.5	0.488

Notes: 1) test scores and SD are from 2015; 2) “No rural info” means the school was in a district that was missing its rural/urban definition.

IV. IMPLEMENTATION

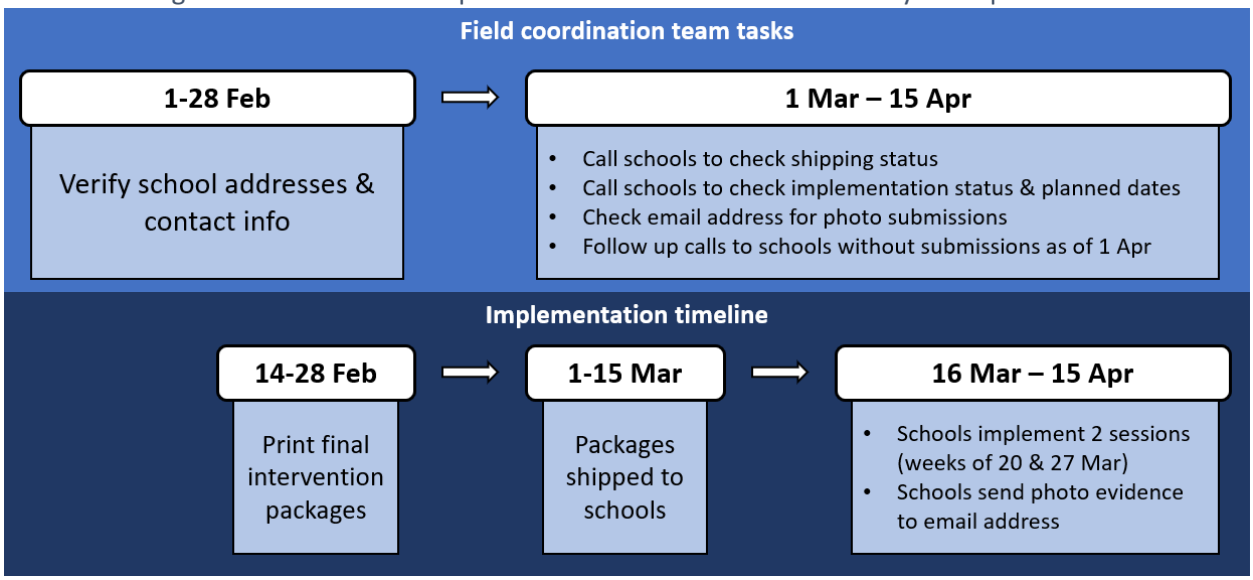
IV.1. FIELD COORDINATION TEAM

A field coordination team was hired to manage the implementation of the experiment. The field coordination team consisted of four consultants. The team was tasked with liaising with the treatment schools throughout the implementation period to both collect and provide information related to the rollout. Each member of the four-person team was responsible for around 300 treatment schools, divided provincially. The team worked for approximately 40 working days per person, between February and April 2017. The field coordination team and their outputs were monitored closely by one member of the Indonesia poverty team and one member of the eMBeD team. The specific tasks of the field coordination team are detailed in Figure *6* below.

IV.2. ROLLOUT TIMELINE

The experiment rollout was done between February and April 2017. The implementation timeline is depicted visually in Figure *6* below, divided into specific tasks of the field coordination team and the overall implementation of the experiment. The experiment was faced with a hard deadline of the end of April 2017, which is when the national exams were set to happen.

Figure 6. Rollout of the experiment was done between February and April 2017.



IV.2.a. Data verification

The first phase of the roll-out was data verification. The team received the Dapodik data set from MoEC, which contained the addresses and contact information of each of the 2404 schools. However, to ensure that packages would be delivered to the right addresses, the field coordination team conducted phone calls to validate the Dapodik data, namely the school addresses and contact information. These initial calls were also used to inform the schools that a package containing a lesson plan from MoEC would soon arrive. Based on data verification, the final tally for the treatment group (all students in all classrooms in Grade 9 of the selected schools) was 1,202 schools, 10,400 classrooms, and 186,000 students.

IV.2.b. Printing and shipping

To increase the likelihood of successful implementation, the team arranged to print all materials and ship them directly to the schools. Another implementation option was to send soft-copies of the materials through e-mail and ask schools to arrange their own printing. However, this could have been perceived as too burdensome by the schools, which would have hindered full compliance. Thus, although printing and shipping incurred significant extra costs, the team believes those costs to be justified.

Shipping was carried out by a courier service to expedite the process. One lower-cost option would have been to arrange delivery through the Indonesian postal service. However, the country office did not have a pre-existing relationship with the postal service, and there were no clear estimates for delivery time. Thus, the team decided to hire a large courier service with a nationwide scope of operations, which could provide detailed estimates of shipping costs and delivery times per location. In addition, the vendor's system allowed for direct tracking of the delivery status of all packages.

IV.2.c. Compliance checks

To control for attrition and to measure implementation quality, the team conducted compliance and quality checks on the delivery of materials. Packages were mailed out in the first two weeks of March 2017, with instructions to implement the intervention over two weeks: the first on the week of 20 March and the second on the week of 27 March. This deadline was set to leave a large enough gap between the implementation of the experiment and the National Examination week, which was scheduled to take place on the last week of April and first week of May 2017. However, to check for potential delays in delivery and implementation, there were a number of compliance checks put in place, including: 1) follow-up phone calls and emails, and 2) requests to schools to submit evidence of completion.

The team sent letters addressed to the headmaster to nudge compliance. As part of the package, the team included a letter addressed to the headmaster informing them of the intervention's importance, asking them to select their best teacher to implement the program, to monitor the sessions, and to send an email to a designated email address with photos after the sessions were completed. To increase the probability of compliance, the letter was signed by the Ministry of Education.

The team also included a "call to action" in the teacher instruction manual to nudge compliance. The front page of the teacher instruction manual included a letter addressed to the teachers that introduced the intervention and its importance. It also included a "call to action" asking them to send an email to the designated address with i) a photo of students holding their letters written during Session 2 and ii) photos of the best 3 letters in the class.

The field coordination team conducted phone calls and emails to track delivery of packages and to remind schools to implement the sessions. The field coordination team were tasked with follow-up correspondence with each school during the roll out: i) to confirm the package delivery; ii) to ask for the proposed date the school will begin the intervention; iii) to follow-up on if the intervention was carried out on that date. Specifically:

- Schools were called immediately after receiving confirmation from the courier service. During this first call, the field coordination team encouraged schools to conduct the session on schedule, and mentioned that a follow-up call will be conducted in a few days' time.

- A second call was made to check whether sessions had taken place. If yes, the field coordination team asked the schools to send a picture. If not, schools were asked to submit a schedule of when the sessions will take place.
- Whatsapp correspondence was also used to send reminders to specific teachers and schools.

Because making hundreds of phone calls takes time, emails were sent to groups of schools using the "Semua Bisa Pintar 2017" Gmail account. The email contained the same content: that they should have received a package, that they should conduct the sessions, and that they should send pictures afterwards. Up to 3 emails were sent to some schools.

Based on these compliance check measures, the experiment was implemented successfully by a clear majority of schools. As of the end of April 2017, 1156 schools (96% of the treatment group) had confirmed successful receipt of materials and 975 schools (81% of the treatment group) had submitted evidence of full implementation, in the form of class photos and scans of letters. This level of compliance was substantially higher than the GYM experiment in Peru, which had 85% confirmed deliveries of materials and only 44% of schools providing photo verification.

Figure 7. Class photo from SMP N 2 Kalimanah, Central Java.

V. EVALUATION ACTIVITIES

V.1. IMPLEMENTATION FEEDBACK

After the completion of the experiment roll-out, the team collected feedback from schools about the quality of the materials and lesson plan. The team conducted field visits to 10 schools across 4 provinces in Java and Sumatera on 12-14 June 2017. In each school, the team interviewed the principal, two teachers that delivered the session, and two students that received the session.

Principals and teachers gave positive feedback on the sessions, agreeing that the program could be very useful to motivate students and encourage them to increase their level of effort in school. Teachers and principals from all schools suggested that the program could be implemented starting from Grade 7, to expose students to the materials as early as possible. Some teachers noted that the sessions could boost the confidence of students who they considered to be struggling academically, and students who were usually too shy to participate in class.

However, teachers requested additional materials to help them deliver the sessions. Teachers complained that the teacher guidelines were long and wordy, which caused them to need a long time to prepare for the sessions. When interviewed, some teachers had difficulty recounting the key messages of the sessions, and reported confusion in how to deliver the sessions. Some teachers also requested supplementary materials to guide them through familiar challenges that could come up when delivering these sessions, such as when students need alternative explanations of complex concepts (e.g., growth mindset, neuroplasticity), or when students are unwilling to participate in classroom discussions. One common request was to supplement the teacher guidebook and reading materials with training videos, to save preparation time and improve knowledge retention.

Student feedback was also positive, as they reported the sessions and comics being enjoyable and motivating. One student reported that the program had motivated him to try harder in class, and he went from struggling academically to achieving one of the highest national exam scores in his school. Students found the characters and plot points in the comics to be highly relatable, but some felt that the plots were too simple and flat. Students also reported that the sessions felt too packed, and that they needed more time to be allocated for the writing exercises and class discussions.

V.2. IMPACT EVALUATION

V.2.a. Descriptive statistics of the UN data

An impact evaluation was conducted using National Exam (*Ujian Nasional*, UN) scores at the school level to measure the impact of the intervention on student performance. The UN is a national standardized exam that takes place at the end of Year 9 and is a requirement for graduation from SMP[13]. The UN at SMP level covers four subjects: Bahasa Indonesia, English, Math, and Science. Tests are scored from a scale of 100. Every student nationwide is given the same test on the same day. Until 2015, all tests

[13] The UN has since been phased out as a graduation requirement (as of the 2019-20 school year), but was consistently used as a graduation requirement during the lifetime of the experiment. Due to this consistency, as well as the universal coverage, the regularity of administration, and the fact that it was implemented by the government, the team felt it was appropriate to use UN scores as the main outcome variable.

were paper-based, but a gradual rollout of computer-based tests began in 2016. None of the schools in the sample implemented the computer-based test in 2017.

The UN data showed significantly higher scores for Bahasa Indonesia than for the other three subjects. In both years, the average scores for Bahasa Indonesia across the sample were much higher than scores for the other three subjects (as seen in Figure *8* and Figure *9* below). This also holds across islands and provinces.

The data also showed significantly higher performance for schools in Java than Sumatra. The sampled schools in Java outperformed their counterparts Sumatra in the 2017 national exams, scoring 8 points higher in Bahasa Indonesia (out of a scale of 100) and 6 points higher in the other three subjects (Figure 5 below). Within Sumatra, Sumatra Utara had the highest overall scores (average of 54.1 per subject) and Sumatra Selatan had the lowest overall scores (average of 47.9 per subject). In Java, Yogyakarta had the highest scores (68.7 per subject, over 6 points higher than second best province DKI Jakarta with 62.0) and Banten had the lowest (45.3 per subject, the lowest across all provinces). These differences still held, though were less pronounced, in the 2016 scores.

Figure 8. Average 2016 UN test score in sample, by island.

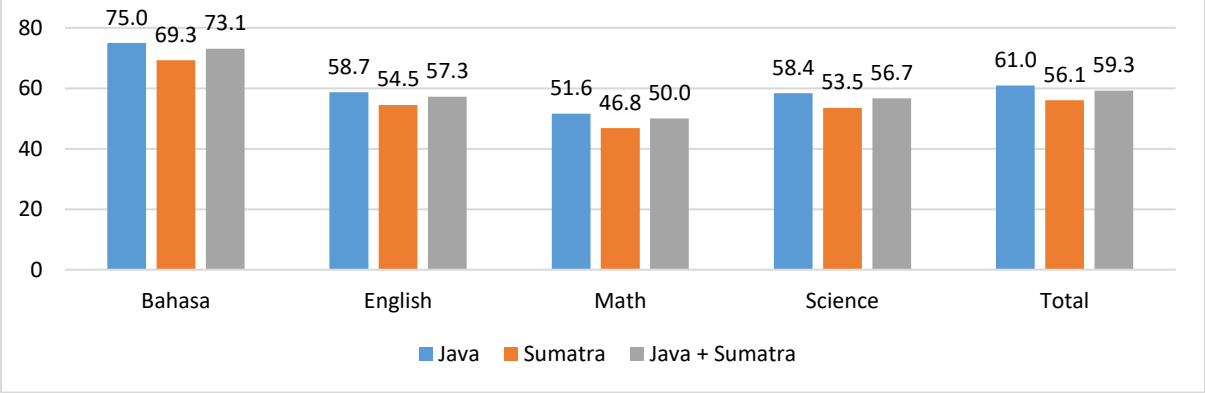
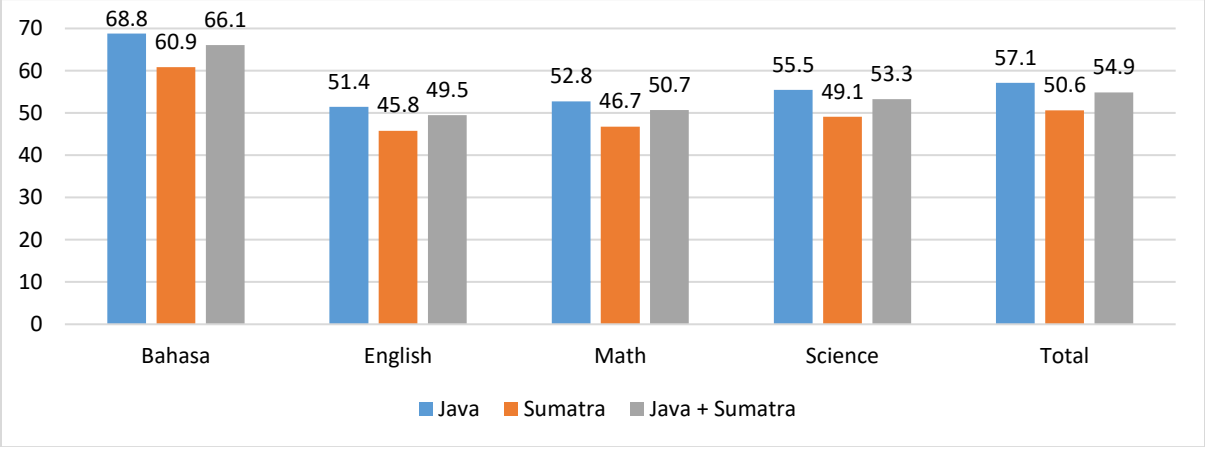


Figure 9. Average 2017 UN test score in sample, by island



V.2.b. Outcome analysis: National Exam scores

The impact evaluation tested a variety of model specifications to extract results from a limited set of data. The impact evaluation consisted of 1) an *ex-post* difference analysis after the intervention, comparing the 2017 school-level UN scores between control and treatment schools, and 2) a difference in differences analysis comparing the change of UN scores from 2016 to 2017 and between control and treatment groups. Theoretically, the difference in difference method is the preferred method, as it is 1) better able to correct for differences in school quality and is thus better able to consistently estimate causal effects, and 2) it is able to control for factors that are constant over time as well as those that are time-variant (e.g., shifts in the curriculum between school years). School level scores were used because the Ministry of Education and Culture are prohibited from sharing individual-level scores by confidentiality laws. So, rather than testing impacts at the individual level, the impact evaluation assessed the means and dispersion of scores at the school level, which is an average of the entire test-taking cohort in that school. As the 2016 UN assessed a different cohort to the 2017 exam, the difference in differences analysis assumes that the quality of schools, teachers, and student cohorts remain similar between treatment and control groups over time. There was no analysis on other outcome indicators, such as changes in beliefs and behaviors, which would have required additional data collection in the form of a baseline and endline survey in control and treatment schools; the team did not have the resources to conduct that additional data collection at the time.

In order to account for the compliance issues observed during the implementation, several alternative analyses were produced. As discussed in Section IV.3, 81 percent compliance was achieved by April 2017, which was the start of the national exam period. In order to account for the differing impacts of treatment schools that either did not implement the intervention or implemented them late, the impact evaluation used three different specifications: 1) intention to treat (full sample included, assuming all treated schools implemented the intervention), 2) treatment on the treated (dropping treatment schools that did not report compliance), and 3) Instrumental Variable (full sample included, but using an instrumental variable estimation to represent schools that did not submit proof of compliance).

The analysis also assessed the impact of the intervention on sub-samples that were considered relevant for policymaking. Aside from the whole-of-sample analysis, the impact evaluation was also conducted on the following sub-samples: 1) dividing schools into low-performing and high-performing schools based on the average 2015 UN scores used in the sampling process, and 2) examining the sub-sample of schools in Java and schools in Sumatra only to assess the impacts on provinces with differing cultures and levels of economic development.

V.2.c. Whole of sample results

An evaluation of the simple difference in 2017 UN scores showed no impact on the overall sample, regardless of model specification. The analysis found no statistically significant impacts of the treatment on the overall sample, in all three models (ITT, compliers, and instrumental variable), meaning that in a snapshot of the 2017 UN results, the control and treatment schools were indistinguishable. This applied to both the aggregate test scores as well as the scores by subject. There are some discernible differences when the sample is split by islands or by high/low performing schools, but none of the results are significant at the 5% level.

Table 8. Whole of sample, simple difference of 2017 scores

	Bahasa	English	Math	Science	Total
ITT (n = 2,397)	-0.01 (0.02)	0.02 (0.03)	0.04 (0.03)	0.02 (0.03)	0.02 (0.03)
Compliers (n = 2,199)	0.00 (0.02)	0.03 (0.03)	0.04 (0.03)	0.03 (0.03)	0.03 (0.03)
IV (n = 2,397)	-0.01 (0.02)	0.02 (0.04)	0.04 (0.04)	0.02 (0.04)	0.02 (0.03)

Notes: Coefficients are in s.d. rather than scores. Regressions control for sample stratum and province fixed effects. SE clustered at the district level. Regressions weighted by school size.
*p<0.05, ** p<0.01, *** p<0.001

However, treatment schools had significantly larger improvements from the previous cohorts (2017 compared to 2016), suggesting that the intervention had a positive impact. When comparing the change in scores between the 2016 and 2017 cohorts (difference-in-difference analysis), significant impacts were found across the overall sample. As the difference in difference method is better able to correct for differences in school quality, this suggests that the SBP intervention was effective in improving student performance. In terms of magnitude, the intervention's impact on the overall test scores was 0.07 standard deviations in the "compliers" specification, with similar effect sizes in English, Math, and Science. The existing education literature suggests that test score differences of 0.2 standard deviations is roughly equivalent to one year of additional schooling. The effect size was stronger in the "compliers" specification than in the "ITT" specification. The impact was strongest, in terms of magnitude and statistical significance, for scores in Math.

Table 9. Whole of sample, difference in differences

	Bahasa	English	Math	Science	Total
ITT (n = 4,794)	0.01 (0.02)	0.05 (0.03)	**0.08** (0.03)**	0.05 (0.03)	**0.06** (0.03)**
Compliers (n = 4,398)	0.02 (0.02)	**0.06** (0.03)**	**0.09** (0.03)**	**0.06** (0.03)**	**0.07** (0.03)**

Notes: Coefficients are in s.d. rather than scores. Regressions control for previous year test scores, sample stratum, and province fixed effects. SE clustered at the district level. Regressions weighted by school size.
*p<0.05, ** p<0.01, *** p<0.001

V.2.d. Impacts by school performance

The intervention contributed to Math score improvements in high performing schools. The statistically significant impacts found in the whole of sample analysis were lost for English and Science scores after disaggregating the sample to low and high performing schools[14]. Some impact was found for English in low performing schools, but with a low overall magnitude. For Math scores, it was found that the intervention was effective in high performing schools but not in low performing schools, with a magnitude of 0.13 standard deviations in both the ITT and compliers analysis.

[14] As mentioned in Section III.1. above, performance levels were defined at the island level using the island-level median scores in 2015 as the threshold. As such, there were an equal number of high and low performance schools in each island and between treatment and control.

Table 10. Low and high performing schools, simple difference

	Bahasa		English		Math		Science		Total	
	Low	High	Low	High	Low	High	Low	High	Low	High
ITT n_{low} = 1,200 n_{high} = 1,197	0.01 (0.02)	-0.04 (0.03)	**0.06*** **(0.03)**	-0.01 (0.05)	0.02 (0.03)	0.05 (0.05)	0.03 (0.03)	0.02 (0.05)	0.03 (0.03)	0.01 (0.05)
Compliers n_{low} = 1,100 n_{high} = 1,099	0.03 (0.02)	-0.03 (0.03)	**0.06*** **(0.03)**	0.00 (0.05)	0.02 (0.03)	0.05 (0.06)	0.04 (0.03)	0.02 (0.05)	0.04 (0.03)	0.02 (0.05)
IV n_{low} = 1,200 n_{high} = 1,197	0.02 (0.02)	-0.04 (0.04)	**0.07*** **(0.03)**	-0.01 (0.05)	0.02 (0.03)	0.06 (0.06)	0.03 (0.04)	0.02 (0.06)	0.04 (0.03)	0.02 (0.05)

Notes: Coefficients are in s.d. rather than scores. Regressions control for sample stratum and province fixed effects. SE clustered at the district level. Regressions weighted by school size.
*p<0.05, ** p<0.01, *** p<0.001

Table 11. Low and high performing schools, difference in differences

	Bahasa		English		Math		Science		Total	
	Low	High	Low	High	Low	High	Low	High	Low	High
ITT n_{low} = 2,400 n_{high} = 2,394	0.02 (0.02)	0.01 (0.03)	**0.06*** **(0.04)**	0.03 (0.05)	0.01 (0.03)	**0.13**** **(0.05)**	0.01 (0.03)	0.07 (0.05)	0.02 (0.03)	0.07 (0.04)
Compliers n_{low} = 2,200 n_{high} = 2,198	0.01 (0.03)	0.02 (0.03)	0.06 (0.04)	0.05 (0.05)	0.01 (0.04)	**0.13**** **(0.05)**	0.02 (0.04)	0.08 (0.05)	0.03 (0.03)	**0.08*** **(0.05)**

Notes: Coefficients are in s.d. rather than scores. Regressions control for previous year test scores, sample stratum, and province fixed effects. SE clustered at the district level. Regressions weighted by school size.
*p<0.05, ** p<0.01, *** p<0.001

V.2.e. Impacts by island group

The data showed that all aggregate results seem to be driven by Sumatra. When conducting the analysis separately on Sumatra and Java, the impact of the treatment in Sumatra was statistically significant across most subjects as well as on the overall aggregate score, whereas the impact of the treatment in Java was not statistically significant for any of the score categories. The impact on Math in Sumatra was highest (0.18 standard deviations in the "compliers" specification), consistent with analysis in previous sections, but the effect size in English, Science, and the total aggregate score was also high. Scores in Bahasa Indonesia were not affected to the same extent by the intervention in Sumatra.

Table 12. Sumatra vs. Java, simple difference

	Bahasa		English		Math		Science		Total	
	S	J	S	J	S	J	S	J	S	J
ITT n_S = 1,184 n_J = 1,213	0.01 (0.04)	-0.03 (0.02)	0.04 (0.05)	0.00 (0.04)	**0.10*** **(0.05)**	0.00 (0.04)	0.09 (0.05)	-0.01 (0.04)	0.07 (0.05)	-0.01 (0.04)
Compliers n_S = 1,052 n_J = 1,147	0.04 (0.04)	-0.03 (0.02)	0.07 (0.05)	0.00 (0.04)	**0.12*** **(0.05)**	0.00 (0.04)	**0.11*** **(0.06)**	-0.01 (0.04)	**0.10*** **(0.05)**	-0.01 (0.04)
IV n_S = 1,184 n_J = 1,213	0.02 (0.05)	-0.03 (0.03)	0.05 (0.07)	0.00 (0.04)	**0.13*** **(0.06)**	0.00 (0.05)	0.10 (0.06)	-0.01 (0.04)	0.09 (0.06)	-0.01 (0.04)

Notes: Coefficients are in s.d. rather than scores. Regressions control for sample stratum and province fixed effects. SE clustered at the district level. Regressions weighted by school size.
*p<0.05, ** p<0.01, *** p<0.001

Table 13. Sumatra vs. Java, difference in differences

	Bahasa		English		Math		Science		Total	
	S	J	S	J	S	J	S	J	S	J
ITT n_S = 2,368 n_J = 2,426	0.06 (0.04)	-0.02 (0.02)	**0.10**** **(0.05)**	0.02 (0.04)	**0.17***** **(0.05)**	0.03 (0.04)	**0.13**** **(0.05)**	0.00 (0.04)	**0.13***** **(0.05)**	0.01 (0.04)
Compliers n_S = 2,104 n_J = 2,294	**0.08*** **(0.04)**	-0.02 (0.02)	**0.13***** **(0.05)**	0.01 (0.04)	**0.18***** **(0.05)**	0.03 (0.04)	**0.15***** **(0.05)**	0.00 (0.04)	**0.15***** **(0.05)**	0.01 (0.04)

Notes: Coefficients are in s.d. rather than scores. Regressions control for previous year test scores, sample stratum, and province fixed effects. SE clustered at the district level. Regressions weighted by school size.
*p<0.05, ** p<0.01, *** p<0.001

Additional work is needed to assess why the intervention's effectiveness varied so greatly between islands. Not enough information was collected to determine why the intervention was highly effective in Sumatra but not effective at all in Java. Although schools in Sumatra had substantially lower average scores than schools in Java (as detailed in section V.2.a. above), the analysis on low vs. high performing schools does not seem to support the narrative that the intervention helps low performers more than high performers. To determine this for sure would require additional estimates that disaggregate the sample across both school performance and island, but statistical power limitations prevent that analysis in this sample. Further experiments are needed, with more comprehensive data collection on beliefs and behaviors, to determine the main causes.

V.2.f. Comparison to Peru

This experiment showed lower effect sizes than those found in Peru. The GYM impact evaluation had effect sizes of 0.2 standard deviations after a similar level of total exposure (90 minutes in Peru vs. 80 minutes in Indonesia). Indonesia's effect sizes at the whole of sample level were comparatively much smaller (0.06 for ITT and 0.07 for compliers), though the impacts in Sumatra were of similar magnitude (0.13 for ITT and 0.15 for compliers in the 4-subject average, and 0.17 and 0.18 in Math). Due to the many differences in implementation (individual vs. school-level test scores, intervention format) and intangible

differences in context between Peru and Indonesia, it is difficult to identify why the intervention worked better in Peru than Indonesia.

V.2.g. Impact on Math

A greater magnitude of impacts was found on Math than in other subjects. As shown in previous sections, the impact of the intervention on improved Math scores was consistently highest across all model specifications. The whole of sample difference in difference results show Math as the only subject with statistically significant impacts at $p<0.01$; similar results were found when the analysis was broken down by school performance and by island group.

The greater impact on Math scores could be caused by the specific format of the intervention. The first comic book in the lesson plan, which first exposed students to the idea of a growth mindset, featured a story where the protagonist faced great difficulty in learning math at the SMP level. It is possible that this specific presentation of materials influenced the students to apply their newly acquired growth mindset on Math more than other subjects. Other comic books later in the lesson plan used examples of English and Physics subjects being the main challenges, which may also have contributed to the magnitude of impacts. Future experiments could test this theory by differentiating the subject matter used in the comic books to the different subjects that are tested in the National Exams.

The impact could also be linked to student perceptions towards the specific subjects. The intervention materials included a call to action for students to think of subjects they found challenging and work towards overcoming the challenge. Students may have considered Math to be the most challenging subject, thus dedicating more effort and persistence towards studying Math over other subjects. This would be consistent with Math being the second lowest scoring subject in the UN and the lowest of Indonesia's PISA scores in all past rounds. Future experiments could test this by conducting a baseline survey of student beliefs and perceptions towards specific subjects that are tested in the National Exams.

The impact on Math scores could also be linked to the nature of the questions tested in the exams. It is possible that the beliefs and behaviors involved in having a growth mindset – e.g., exercising the brain like a muscle, asking others for help – are more effective in improving performance for some types of subjects over others. For example, the types of questions tested in SMP-level Math exams require both an understanding of mathematical concepts, practice in applying the formulas, and mental arithmetic. Although other subjects such as English and Bahasa Indonesia may also test reading comprehension and the application of linguistic concepts, many parts of the exam could also entail rote memorization, which may not be improved as much through applying a growth mindset. Due to the limited literature on how growth mindset affects specific types of skills and learning, no further conclusions can be drawn about this potential linkage at this time.

V.3. COST ASSESSMENT

The low per-student logistical cost of the intervention makes replication and scale-up feasible. Assuming the use of the same method for distribution of materials (centralized printing in Jakarta and shipping to schools via private courier services), the average cost of printing and shipping of materials was US$0.25 per student at the time of implementation (approximately IDR3,500 per student).[15] This is a similar cost

[15] This average cost per student intentionally excludes costs that are specific to this project's experimental design, e.g. monitoring implementation compliance through the field coordination team, intervention and experiment

level to the previous World Bank experiment in Peru of US$0.20 per student. Alternative shipping options could be explored, for example utilizing the Indonesian Postal Service (PT. Pos Indonesia), which comes at a much lower cost than private couriers. However, tracking of packages and delivery timelines would need to be managed more closely when using the Indonesian Postal Service.

Shipping and printing costs could be reduced through localized printing and distribution, but it may reduce the quality of materials and is more dependent on the availability of local firms. The printing and shipping costs of the intervention could be further reduced with localized printing at the provincial or district level, which would reduce the cost of distribution to schools. This would be more important for provinces in eastern Indonesia, where shipping from Jakarta would be prohibitively expensive. However, localized printing relies on the availability of local printing firms and high-quality materials. Poor-quality printouts may affect the effectiveness of the intervention, as the materials are of a highly visual nature (i.e. comic books).

In the long term, costs could be further reduced by printing at the school level, or to use existing channels by which MoEC distributes regular school textbooks and materials. Distribution costs could be further reduced by sending soft-copy materials to schools and having them arrange their own printing. This would reduce distribution costs, but would be constrained by the availability of local printing firms near the schools, or by the school's ability to print materials in-house. Alternatively, MoEC could leverage its existing channels and procedures to distribute regular school textbooks and other materials.

Beyond a simple replication of this paper-based intervention, the costs and consequences become less clear. A transition from paper-based comic books and student worksheets to a digital platform would eliminate the need for printing and shipping. However, this alternative would necessarily lead to a "mode effect", where the intervention may not have the same impact when delivered in a different format. Also, a transition to digital materials relies on schools having access to enough computers or tablets for all students to access materials individually during the session.

The cost calculations above do not include the costs and workload of developing the intervention and the experimental design, which in both cases could be absorbed by existing MoEC teams. The in-house method of developing the intervention instruments entailed four WB staff with a total estimated workload of 50 person-weeks over a six-month period. The intervention roll-out entailed a further 10-person weeks from the core team, as well as the field coordination team of four people working for 8 weeks (40 days) each. These costs are not included in the calculations, even if MoEC were to continue the development of the intervention and run additional experiments, as they would be absorbed by existing MoEC resources as part of the development of a character building curriculum.

Overall, the low cost, combined with the low time commitment and degree of intrusiveness on the core curriculum, makes replication or scale-up feasible at either the national or local government levels.

---

design, impact evaluation, and client engagement. It is assumed that if the activity is scaled up nationwide under MoEC, linked activities such as adapting the design of materials, rollout, and M&E will be undertaken by existing MoEC staff/functions and will not incur extra costs.

VI. LEARNING AND FOLLOW-UP STEPS

VI.1. KEY LEARNINGS

The first key learning from the activity is the need to conduct additional data collection to measure the intermediate impacts of behavioral interventions, e.g. changes in the beliefs and behaviors of students, as well as of teachers. The impact evaluation conducted for this experiment relies on some large assumptions: 1) that teachers would be able to deliver the intervention as designed, 2) that the intervention as designed could successfully teach growth mindset to students, 2) that students who developed growth mindsets would then demonstrate improvements in their ability to learn, and 3) that these demonstrated improvements would manifest in improved UN scores. In addition, although the SBP experiment was found to have an impact on test scores even using school aggregate scores, there were several findings that could not be explained using the limited data available, such as why impacts were stronger for certain subjects and why SBP was only effective in Sumatra and not Java. The impact evaluation could be much improved with additional data collection to monitor: 1) changes in the beliefs and attitudes of teachers, 2) changes in the beliefs of participating students (e.g. whether they acquired a growth mindset), as well as 3) changes in their study behaviors (e.g. whether they persevered for longer in response to challenges or applied the behaviors taught in the lesson plan). This would provide evidence of the success of the intervention and would allow for a clear attribution of final outcomes.

The second key learning from the activity is that experiments involving students and schools should be better timed to ensure greater and longer exposure. One of the main challenges in implementation was that the timing of the experiment roll-out (March 2017) was too close to the dates of the UN (end of April 2017). This meant that schools received the materials and the request to conduct the sessions while the schools were in the middle of preparation activities for the exams. Although the compliance checks described in Section IV.3. indicated a high level of overall compliance, it would not have left enough time for students to fully digest and apply the growth mindset concepts to improve their perceptions and study habits. Implementing the lessons at more compatible times for schools, such as the start of the semester or school year, could increase the likelihood of having the exposure to the intervention translate to improved outcomes.

Finally, further testing of alternative intervention designs could provide more insight into the effectiveness of different methods of delivering the materials. In the SBP project, as well as in the work done in Peru and other contexts, the delivery methods were similar: only a minimal amount of session time (between 60 to 80 minutes total) and only implemented once. However, future opportunities could be used to test alternative intervention designs, such as having more sessions, covering more materials, delivering it in different ways, and also testing the use of reinforcing strategies beyond the core sessions.

VI.2. FURTHER DEVELOPMENT

In response to a MoEC request for a more extensive "socio-emotional curriculum", the team continued the development of additional modules in a follow-up experiment called "Semua Bisa Sukses" (Everyone Can Succeed). MoEC's response to the "Growth Mindset" lesson plan was positive, with the client requesting similar modules to be prepared that could cover the entire course of a semester or school year. The team has since developed a follow-up experiment that includes additional lessons in: 1) developing student aspirations, 2) grit and goal-directed persistence, 3) planning and time management, and 4) task initiation (particularly, fighting procrastination). This follow-up experiment was conducted between February 2018 and June 2019 and incorporates the key learnings discussed in section VI.1.