INDONESIA: Teacher certification and beyond

An empirical evaluation of the teacher certification program and education quality improvements in Indonesia

Kingdom of the Netherlands

DESP - TF
Dutch Education Support Program Trust Fund

THE WORLD BANK
IBRD • IDA

The Word Bank Office Jakarta
Indonesia Stock Exchange Building, Tower II, 12th floor
Jl. Jend. Sudirman Kav. 52-53
Jakarta 12190
Tel : (+6221) 5299-3000
Fax : (+6221) 5299-3111
http://www.worldbank.org/id

Published in December 2015

Cover photo courtesy by The World Bank – Jakarta

Teacher certification and beyond: An empirical evaluation of the teacher certification program and education quality improvements in Indonesia is a product of the staff of the World Bank. The findings, interpretation, and conclusions expressed herein do not necessarily reflect the view of the World Bank, its board of Executive Directors, officers or any of its member countries.

The World Bank does not guarantee the accuracy of the data included in this work. The boundaries, colors, denominations, and other information shown on any map of this work do no imply any judgment on the part of the World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

The work is funded by the Government of the Kingdom of the Netherlands under the Dutch Education Support Program (DESP) Trust Fund. DESP provides support to the Government of Indonesia through the World Bank for the purpose of developing policies, studies, and programs that help the government achieve its education strategic plan.

Report No. 94019-ID

INDONESIA:
Teacher certification and beyond

An empirical evaluation of the teacher certification program and education quality improvements in Indonesia

Education Global Practice
East Asia and Pacific Region

Kingdom of the Netherlands
DESP - TF
Dutch Education Support Program Trust Fund
THE WORLD BANK
IBRD • IDA

Regional Vice President	Axel van Trotsenburg
Country Director	Rodrigo A. Chaves
Senior Practice Director	Claudia Costin
Practice Manager	Harry Patrinos
Task Team Leader	Andrew Ragatz

Contents

Figures

Tables

Acknowledgements

This report was written by Joppe de Ree and most of the research was executed under his lead, while working for the World Bank office in Indonesia. The report, however, could not have been written without the help of a large group of people, who, at different times, have assumed important roles and responsibilities all contributing towards the successful completion of this work. The research was set up in 2008, as one of the research projects under the Better Education through Reformed Management and Universal Teacher Upgrading (BERMUTU) program. In this, the World Bank worked in a unique and longstanding partnership with Indonesia's Ministry of Education and Culture. The research was generously supported by the Government of the Netherlands, under the Dutch Education Support Program.

The research project was from the design planned as a randomized controlled trial, with the intention of rigorously evaluating the process of certification, and the impact of paying teachers double on their behavior and on student-learning outcomes. The randomized controlled trial was initiated by Menno Pradhan (then World Bank) partnering with Karthik Muralidharan (University of California, San Diego) and Halsey Rogers (World Bank). Subsequently, the project management was taken over by Amanda Beatty, with the continuing support of Menno Pradhan, Karthik Muralidharan and Halsey Rogers. In 2010 Joppe de Ree took over the project management and the analytical work, until the completion of this document. Various colleagues in the World Bank have often played a pivotal role in the success of the project. Susiana Iskandar, Titie Hadiyati (and team) and Dedy Junaedi (and his teams of enumerators throughout the country, data managers, data cleaners, people he affectionately calls "our friends") have been indispensable and, without them, this work would not have been possible. The government teams were equally important for the success of this project. Ibu Dian Wahyuni and Ibu Santi Ambarukmi (both *Pusat Pengembangan Profesi Pendidik, Pusbangprodik*) managed the implementation of the "the intervention" of the randomized trial, by making sure that prequalified teachers in the randomly selected treatment schools received accelerated access to the certification process. In *Pusat Penelitian Kebijakan (Puslitjak)*, particular thanks go to Ibu Yendri Wirda Burhan, Pak Simon Silisabon (and team); and in *Pusat Penilaian Pendidikan (Puspendik)*, particular thanks go to Ibu Dhani Nugaan, Pak Bastari, Pak Hari Setiadi, Ibu Rahmawati and Ibu Yani Sumarno (and team). Together with various others over the years they have taken on exceptional responsibilities in the design and implementation of this research project. Subsequent to collecting and organizing the data, the research activities have also benefited from senior government consultants, such as Pak Jiyono and Pak Yahya Umar.

The content of this report is the culmination of years of researching the data. Various people have played a key role in this process. Co-authors on different projects related to this work, such as Mae Chu Chang, Sheldon Shaeffer, Samer Al-Samarrai, Andy Ragatz, Ritchie Stevenson, Susiana Iskandar, Karthik Muralidharan, Halsey Rogers and Menno Pradhan, have helped shape ideas for this report. In-depth discussions also took place on the content of various earlier drafts of this report, notably with Samer Al-Samarrai, Andrew Ragatz, Susiana Iskandar, and Ratna Kesuma. Over the years, the research has been supported by research assistants Ai Li Ang, Husnul Rizal and Laura Wijaya. In the final stages, this report has benefited substantially from editorial assistance of Peter Kjaer Milne, graphic design and production support from Yvonne Armanto and Santi Santobri, and last but not least, thoughtful and generous peer reviewer comments from Lant Pritchett, Tazeen Fasih and Daniel Suryadarma. Thank you all!

Acronyms and Abbreviations

AEC	ASEAN Economic Community	
ASEAN	Association of Southeast Asian Nations	
Balitbang	Research and Development Board	Badan Penelitian dan Pengembangan
BERTUMU	Better Education through Reformed Management and Universal Teacher Upgrading	
BOS	School Operational Assistance	Bantuan Operasional Sekolah
BPS	BPS-Statistics Indonesia	Badan Pusat Statistik
CPD	Continuous Professional Development	
CTL	Contextual Teaching and Learning	
Edstats	World Bank Education Statistics Database	
GDP	Gross Domestic Product	
GTT	School-hired temporary teachers	Guru Tidak Tetap
IDR	Indonesian Rupiah	
INPRES	Presidential Instruction	Instruksi Presiden
LPTK	Teacher Education Institution	Lembaga Pendidikan Tenaga Kependidikan
MenPAN	Minister of Administrative and Bureaucratic Reform	Menteri Negara Pendayagunaan Administrasi Negara dan Reformasi Birokrasi
MoEC	Ministry of Education and Culture	
MoRA	Ministry of Religious Affairs	
NUPTK	Unique Identification Number for Teachers and Teaching Personnel	Nomor Unik Pendidik dan Tenaga Kependidikan
OECD	Organisation for Economic Co-operation and Development	
PISA	Program for International Student Assessment	
PLPG	Education and Training for the Teaching Profession	Pendidikan dan Latihan Profesi Guru
Puslitjak	Center for Policy Research, Ministry of Education and Culture	Pusat Penelitian Kebijakan
Puspendik	Educational Assessment Center, Ministry of Education and Culture	Pusat Penilaian Pendidikan
RCT	Randomized Controlled Trial	
SD	Primary School	Sekolah Dasar
SMP	Junior Secondary School	Sekolah Menengah Pertama
SMA	Senior Secondary School	Sekolah Menengah Atas
SMK	Vocational Secondary School	Sekolah Menengah Kejuruan
TIMSS	Trends in International Mathematics and Science Study	
UGK	Teacher Competency Exam	Ujian Kompetensi Guru
USD	United States Dollar	
UT	Open University	Universitas Terbuka

Preface

Nearly a decade ago in 2005, Indonesia passed Law No. 14/2005 on Teachers and Lecturers. By far the largest component of the law, at least in terms of its fiscal implications, was the teacher certification program. The certification program aimed to certify all teachers by 2015. The program was rolled out at a rate of approximately 200,000 teachers each year. With more than 2 million teachers to certify, the process was meant to take 10 years.[1] The rollout is now well advanced and government statistics show that 1.15 million teachers were certified in 2012. As teachers completed the certification process, they were eligible for a certification allowance, equal to their base salary in the civil service. In many cases therefore certification **doubled a teacher's take-home pay**. Certification comes with a serious price tag: if the program is fully implemented it would cost more than USD 5 billion each year, roughly a quarter of the overall education budget. In order to prequalify for certification, teachers must have a university bachelor's degree or equivalent, although exemptions to this rule were introduced for senior teachers.

This report evaluates the certification program as it was implemented, in terms of its impact on student-learning outcomes. Results of the analysis are sobering: despite its massive fiscal implications, the certification program has not led to substantial improvements in student-learning outcomes so far. However, the report provides leads into how to gradually transform the system into one that can yield higher returns in educational performance going forward. It emphasizes the importance of a system that rewards useful demonstrated competencies, such as minimum levels of subject-matter knowledge, rather than loose proxies for quality such as bachelor's degrees or seniority alone (which is essentially what the current certification program does). The report also highlights the need for reforms in the pre-service system of teacher training and teacher hiring.

The findings and conclusions of this report are drawn from 6 years of collecting and researching micro level data in a unique partnership between the Government of Indonesia and The World Bank. A geographically representative sample of 360 Indonesian public primary and junior secondary schools were visited three times in a period of 2.5 years, and the academic development of tens of thousands of students were tracked throughout this period. The academic performance of students across time was then linked to survey information and subject-matter test scores of their teachers. Based on this unusually rich, matched student-to-teacher data base we analyze the impact of certification on student learning outcomes (in Chapter 2), but we also look much deeper into "how much students learn in a year in school" and the role of their teachers in supporting their learning process (in Chapter 3).

Four main, general conclusions are drawn from the data analysis presented in this report:

- **Conclusion #1.** Paying teachers more does not make them teach better.
- **Conclusion #2.** Teachers with bachelor's degrees are only moderately better teachers than teachers without bachelor's degrees. This is especially true for primary school teachers.
- **Conclusion #3.** Teachers with a reasonable level of subject-matter knowledge are much better teachers than teachers who have difficulties with even the most basic mathematical exercises. Hundreds of thousands of primary teachers in Indonesia have difficulties with even the most basic mathematical exercises.
- **Conclusion #4**. Teacher training colleges produce 250,000 university trained teachers each year, whereas the school system needs only 50,000 – 100,000.

[1] Most teachers in the early years of certification were certified through a successful portfolio assessment of past training and experiences. Later, most teachers followed a 90-hour training program.

Based on these general conclusions we analyze the potential effect of the certification program in the short, medium and longer term (Chapter 2 of this report). The "effect" of certification is a combination of three forces working together. The three complementary forces are the behavioral mechanism, the academic upgrading mechanism, and the attraction mechanism.

The behavioral mechanism. Teachers who are certified are paid more. Some theories predict that higher salaries could increase effort, and/or allow teachers to spend more time on lesson preparation for example (by spending less on outside employment).[2] This behavioral mechanism applies to all teachers who are certified and receive the certification allowance. We learn from ***conclusion #1*** however that the behavioral mechanism does not stand up to empirical scrutiny: paying teachers more decreases their reliance on outside employment, it also decreases (self-reported) financial stress, but it does not make teachers teach better.[3]

The academic upgrading mechanism. About half of all teachers did not prequalify for certification when the program was introduced nearly a decade ago. These teachers had to obtain a bachelor's degree first, in order to prequalify. In other words, the certification allowance provided a massive financial incentive for half of the teaching force to "go back to school". And indeed many in-service teachers did exactly that, most of them through remote learning courses. In 2012, 55 percent of all primary teachers had a bachelor's degree or equivalent, up from only 17 percent in 2005/06. But from ***conclusion #2*** we conclude that this wave of academic upgrading led to only moderate improvements in education quality. With limited supervision on the competency levels teachers attain in the process of obtaining their degrees, minimum quality standards are not guaranteed. Competency "on paper" (bachelor's degrees) are not necessarily on a par with competency "in reality" (skills that are useful for teaching). Based on the analysis presented in Annex C and Section 2.2 we forecast that when all teachers in the system have obtained a university bachelor's degree, student-learning outcomes in the aggregate will only increase by 0.16 of a standard deviation, or, if we rely on the scaling used by the Programme for International Student Assessment PISA, by roughly 11 PISA points. These effects are not nearly enough to close half of the gap in educational performance between Indonesia and, say, Thailand and Malaysia.

The attraction mechanism. A partial aim of the certification program was to provide incentives for the brightest high-school graduates to pursue a career in teaching, by making the profession much more attractive financially. While the base salary of a junior teacher in the civil service compares roughly to the salary level of the median worker with a university bachelor's degree, the base salary *plus* the certification allowance compares to the 90th percentile. Today, we cannot yet precisely assess the efficacy of this mechanism, as the impact of attraction should only become measurable in increased educational performance over the next 10 to 20 years. But the report argues that it is questionable that attraction will successfully increase education quality in the medium term, unless today's pre-service system is carefully amended. From ***conclusion #4*** we learn that teacher training colleges currently produce too many university trained teachers. At the same time, districts and schools do not appear to hire the best, or even the best trained candidates out of this large pool of job-seekers. Government statistics show that approximately 50 percent of junior primary school teachers in the civil service *do not have a university bachelor's degree*. With massive overproduction of teachers, combined with unclear and potentially unfair hiring rules, the chances of finding a secure and well-paid teaching position are slim, even for the high-caliber candidates. When high-school graduates internalize the poor career prospects in teaching, the brightest among them should opt out first, because they have more and better outside options. The current system, therefore, potentially achieves the opposite of what it intended: *deterring* rather than *attracting* high-caliber high-school graduates to pursue a career in teaching.

The findings of this report confirm what many skeptics of the current program had already anticipated, or feared perhaps, when the certification program was introduced nearly a decade ago. In the mean time, however, competency testing has

[2] See Akerlof & Yellen (1990) for example for background.

[3] See Section 2.1 for a discussion of the empirical results and the design of the randomized controlled trial. Also, see De Ree, Muralidharan, Pradhan, & Rogers (forthcoming) for more details and results on the experiment.

been introduced to teachers entering the certification process, and some have failed these tests. This has forced those who failed initially to embark on a process of professional development to improve competencies, with the purpose to do better in one of the later rounds. In this way, the system started to incentivize improvements in quality. In the report we argue that these are steps in the right direction. But a system that directly rewards useful demonstrated competencies could benefit from a more rigorous implementation (see Chapter 4 for a discussion). Annex C and Section 3.3 of this report show that improvements in the level of teachers' subject-matter knowledge across the board, to reasonable levels, can quickly lead to much improved levels of student achievement in the short to medium term. Today, however, with the majority of prequalified teachers certified, we still observe hundreds of thousands of teachers who demonstrate difficulties with even the most basic mathematical exercises (see Section 3.3.1).

Making Indonesia's education system "future-proof" requires reforms explicitly geared towards quality. As people, also teachers, tend to respond to incentives, such reforms may work best if steps in teachers' career progressions, or professional (re)certification, are tied directly to their ability to demonstrate useful competencies, one of such competencies being minimal levels of subject matter proficiency. In this way, teachers bear some responsibility for the system's quality upgrade, as their monetary interests are aligned with the broader interest of society, i.e., better teachers in Indonesian classrooms. Chapter 4 discusses such policy options as well as options for pre-service reform in more detail. The effective implementation of these reforms and making them work is the next real challenge, especially in an environment where different groups have competing interests. Increasing the quality of education in Indonesia requires broad agreement on the need to improve education quality and full commitment from all stakeholders, politicians, policymakers, unions, teachers and parents.

CHAPTER 1

One decade on for teacher certification in Indonesia

The origins of Indonesia's teacher certification program lie in the 1970s when the then-president, Suharto, ordered the building of tens of thousands of new primary schools under the so-called INPRES[4] primary school construction program. Between 1973 and 1979, Indonesia managed to more than double the number of new entrants into primary school (World Bank, 1989). But, while the school construction program had clear benefits,[5] its rapid implementation also had downsides: schools were built so fast that there were not enough trained teachers to fill them and new teachers were prepared in a rush. This process is said to have diluted the quality of the teacher force in Indonesia. This dilution of quality underpins some of the more recent reform programs in Indonesian education, such as the teacher certification program, the main focus of this report.

Twenty-five years after the start of the INPRES program, in the late 1990s, the Suharto regime fell in the aftermath of the Asian financial crisis. The new Indonesian Government decided to participate in an Organisation of Economic Cooperation and Development (OECD) initiative called PISA, the Programme for International Student Assessment. PISA measures academic achievement in representative samples of 15-year-olds in a number of countries around the world. Because PISA relies on comparable tests for its evaluation, it is able to measure how countries compare with each other in terms of their main educational outcome, namely student academic achievement. Hence, at least to some extent, PISA sheds light on how well countries' education systems are functioning.

PISA 2000, the first round of PISA and the first in which Indonesia participated, confirmed what many had long suspected: the quality of education in Indonesia was low by international standards. The results placed Indonesia in 38th place among 41 participating countries, significantly lower than Thailand, a regional peer. More importantly perhaps, PISA classified 70 percent of Indonesian 15-year-olds below Level 2, the level of proficiency that PISA deems is "needed to participate effectively and productively in society" (OECD, 2010). What PISA means by "in society" is not clearly articulated and Indonesia's society is significantly different from those of many OECD countries. But it does indicate that Indonesia needs to take significant steps to improve education if it is to maintain its robust rates of economic growth in the longer term.

The PISA findings marked the starting point for massive government investment in Indonesia's education system in the post-Suharto period. Indonesia's new leaders understood that something dramatic was needed to upgrade (or restore) the profile of the teaching profession. Indonesia's answer was the formulation and implementation of Law No. 14/2005 on Teachers and Lecturers (known as the Teacher Law). The flagship program under the new law was a teacher certification program that aimed to reestablish some of the esteem that the teaching profession had lost during

[4] INPRES stands for Instruksi Presiden, or Presidential Instruction.

[5] MIT economist Ester Duflo used the INPRES primary school construction program as a source of information to estimate the effect of schooling on wages. Esther Duflo estimates that the economic rate of return to a year of primary school ranges between 6 and 10 percent (Duflo, 2001). The building of schools, therefore, clearly had benefits.

the rapid expansion of the 1970s and 1980s. The program promised teachers a generous professional allowance—the certification allowance—that was equal to their base salary upon successful completion of the program. Certification, therefore, essentially doubled teachers' take-home pay.[6] The initial design of the program was such that to prequalify for certification teachers first had to obtain a university bachelor's degree and, as additional proof of competency, had to demonstrate their skills through a written competency test, classroom observation, and a portfolio of past training and experience. The idea was that teachers without the right teaching skills would have a clear financial incentive (a doubling of pay through the certification allowance) to upgrade their skills to the standard required.

In the early 2000s, political momentum was building in favor of implementing this design of the certification program. However, the initial enthusiasm for fundamental reform waned once the program arrived in parliament; the previously strict regulations in the initial design were significantly watered down. Under pressure from teacher unions, the requirement to *demonstrate* competency was dropped and only a portfolio assessment of past training and experience was retained. The unions argued that, by obtaining a bachelor's degree from one of the teacher training colleges in the country, teachers had already demonstrated their competency and that an additional test to measure the value of their formal training was not necessary. Under pressure from union lobbyists, parliament agreed in favor of the unions and against the additional rigor of the original program designers.[7]

The new Teacher Law mandated that by 2015 all teachers in the system had to be certified. Today, the implementation of the certification program is well advanced and each year over 200,000 teachers are certified. Observers noted that during the roll-out of the certification program there has been very little if any selectivity. Possibly driven by the need to meet the ambitious annual implementation targets, the vast majority of teachers passed the certification process, either directly after a successful portfolio assessment of past training and experience, or after just nine days of additional training. The most recent 2012 NUPTK teacher census lists around 1.15 million certified teachers and, if the current rate is maintained, most eligible teachers will be certified by 2015.

The fiscal implications of the certification program are enormous. In the past decade, Indonesia has doubled its spending on education. Much of the increase has gone towards teacher salaries in the form of the certification allowance (World Bank, 2013). With roughly 2 million potentially eligible teachers in the country, on full implementation of the program in 2015, certification will cost Indonesia about IDR 65 trillion each year (equivalent to about USD 5.4 billion at the current IDR 12,000/USD exchange rate).[8] The fact that the 2010 education budget was about IDR 200 trillion indicates the huge fiscal implications of the program. In terms of fiscal impact, the teacher certification program is by far the largest reform in recent education history. Not only this, Al-Samarrai & Cerdan-Infantes (2013) argue that at the current rate of implementation, and also assuming that contract staff is "regularized", the program will become financially untenable.

In light of this fiscal expansion, the most recent and highly anticipated PISA 2012 results were particularly disappointing. PISA 2012 ranked math ability of Indonesia's 15-year-olds as second to last among the 65 participating countries, and still well behind Thailand and Malaysia. Meanwhile, Vietnam, an economically comparable country to Indonesia participating in PISA for the first time, entered the hit parade of educational performance at a level well above Indonesia's. The bottom line seems to be that, despite all the extra spending, nothing major has changed in the performance of Indonesia's school children. The combined evidence suggests that past efforts to transform the teacher force, such as making sure that all teachers have a university bachelor's degree, have either been ineffective, or at best not effective enough to be measurable in today's PISA scores. Indonesia's new Minister of Education Anies Baswedan recently made reference to

[6] Some teachers also receive other allowances in addition to their base pay. In that case, the certification allowance does not double their pay, yet it significantly increases it.

[7] A more in-depth discussion on the political economy of this process is provided by Chang, Shaeffer, Al-Samarrai, Ragatz, De Ree, & Stevenson (2013)

[8] The exact amount depends on the average base salary of the certified teachers, as well as the cost of implementation. We arrive at the IDR 65 trillion by multiplying a monthly base salary of IDR 2.7 million with 12 (months) and 2 million potentially eligible teachers (potentially eligible teachers are the 1.7 million in the civil service, and the 0.3 million permanent private school teachers). The equivalent of this in US$ terms is US$5.4 billion at the current IDR 12,000/USD exchange rate.

these PISA findings in *Republika*, one of Indonesia's mainstream daily newspapers, arguing that "Indonesia's education is in a state of emergency" (Novia, 2014).

This report tries to explain what has gone wrong with the teacher certification program and which elements of the system now need urgent political attention if student-learning outcomes are to be improved going forward. Earlier publications on Indonesia's teacher certification program have mainly looked at the effects of certification in the short term (Fahmi, Maulana, & Yusuf (2011), World Bank (2012), Chang, Shaeffer, Al-Samarrai, Ragatz, De Ree, & Stevenson (2013) and De Ree, Muralidharan, Pradhan, & Rogers (forthcoming)). This report recognizes explicitly that some of the impacts of the certification program are longer term and it makes informed assessments of these future impacts. For example, although the law required teachers to obtain a university bachelor's degree, and hundreds of thousands of Indonesian teachers have since obtained their degrees, many others are still in the process of upgrading. Consequently, we estimate that the full impact of the *academic upgrading* process will only be completely incorporated in aggregate statistics on student performance from 2020 onwards. The fact that the most recent PISA 2012 results were disappointing, therefore, does not necessarily mean that we cannot still hope for gains in the medium term, given that the full impacts of academic upgrading are lagging.

However, in Section 2.2 we forecast that the impact of teacher upgrading on student-learning outcomes will be only moderate. This estimated weak-to-moderate impact of the recent wave of academic upgrading indicates that competency "on paper" is not necessarily on a par with competency "in reality". The report shows that there are many good teachers with bachelor's degrees in Indonesia. The issue is, however, that there are also many teachers with bachelor's degrees who are underperforming. One of the reasons why the certification program failed to have the intended impact on student-learning outcomes is that it explicitly rewards bachelor's degrees, whereas a bachelor's degree is only a weak marker for quality. This adds to existing empirical evidence that finds that formal academic qualifications are not particularly strong predictors of teacher quality (see Hanushek & Rivkin (2006) and Glewwe, Hanushek, Humpage, & Ravina (2011) for examples). Based on this finding, we argue that the current certification system in Indonesia lacks incentives for teachers to raise their performance in the classroom. In fact, the certification allowance provides the financial incentives to obtain a bachelor's degree, which is not necessarily proof of being a good teacher. Consequently, new policies should seek to tie allowances to minimum levels of *demonstrated* teaching ability, thereby shifting a larger share of the responsibility for quality education onto teachers themselves.

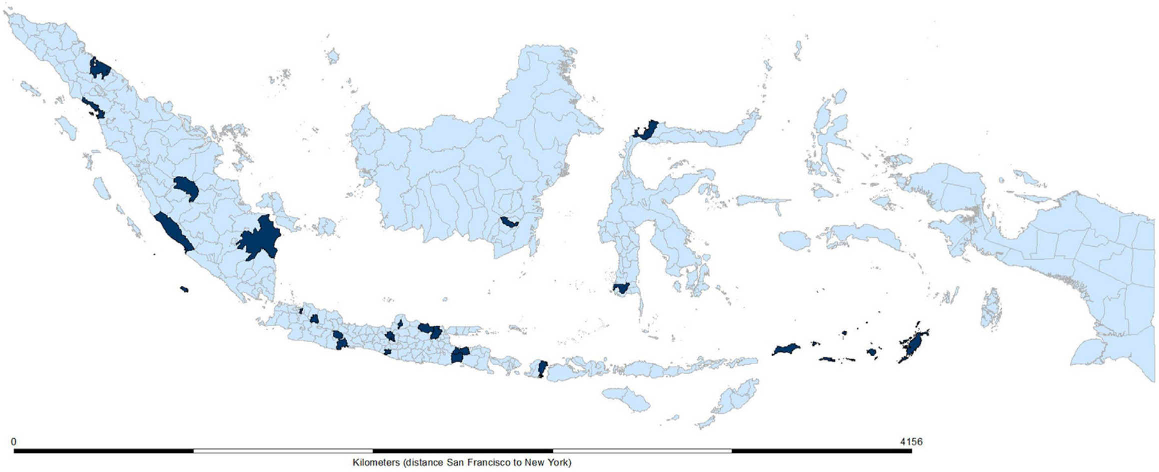
In Section 3.3.1 we show that a worryingly large group of primary school teachers has trouble with even the most elementary mathematical exercises.[9] One feasible policy option, in the spirit of the initially proposed design of the certification program, is to link teacher certification and the certification allowance to demonstrated minimum levels of teacher subject-matter proficiency. Our empirical estimates suggest that reasonable gains in teachers' subject-matter proficiency in primary schools have the potential to generate substantial gains in student achievement in the medium term. We explicitly recognize that subject-matter proficiency is not the only characteristic of teachers that matters—or even the most important one. Some notion of pedagogical ability complements the command of the subject itself. But pedagogical abilities are much less easily measurable or evaluated given the nature and scale of Indonesia's education system. After all, Indonesia employs close to 3 million teachers spread out across several thousand islands.

The analytical work discussed in this report relies for the most part on a unique matched student-to-teacher database, collected specifically for the evaluation of the teacher certification program. The data were collected by the Government in partnership with the World Bank, and financially supported by the Dutch Government, through the Dutch Education Support Program (DESP). Primary and junior secondary students were tracked for 2.5 years as they progressed through school. Their performance across time was then linked to survey information and subject-matter test scores of their teachers. The data are representative of 40 percent of the public primary and public junior secondary schools in Indonesia

[9] This observation is not new however, as early work by Jiyono (1985-86) (cited in World Bank (1989)) already observed similar trends.

and have full geographic coverage—from districts in Sumatra in the west to the southern Maluku islands close to Papua in the east.[10] Figure 1 highlights the 20 Indonesian districts that were randomly selected to take part in this unique survey. The distance from the most western school to the most eastern school in the data spans roughly the full width of the United States of America.

Figure 1: A representative sample of 20 districts was selected to take part in this longitudinal survey



This report is structured into four chapters, including this summary introduction in Chapter 1. Chapter 2 introduces the teacher certification program in Indonesia in detail and discusses its broad impact on teachers' well-being, teachers' behavior and student-learning outcomes. We use some of the technical results presented in more detail in the annexes to understand why the teacher certification program failed to produce some of the desired outcomes. Chapter 2 concludes that without careful systemic amendments the certification program cannot be expected to lead to improvements in teacher quality, even in the medium to longer term. Chapter 3 analyzes education in Indonesia beyond certification, and investigates what *does* matter for student learning in Indonesia. It takes a closer look at how much school-attending children learn as they age, and investigates the role and importance of schools and teachers. One of the primary results is that direct proxies for teacher competencies (e.g., subject-matter test scores) are far better markers for teacher quality than indirect proxies (e.g., bachelor's degrees, or years-of-teaching experience). This has important consequences for the design of an in-service teacher professional management system, which is discussed in Chapter 4. Chapter 4 presents a set of policy options aimed at reorganizing the pre-service (training and hiring), as well as supporting and incentivizing in-service quality upgrading.

[10] About 80,000 students were tested in each of the three rounds, November 2009, April 2011 and April 2012.

CHAPTER 2

An empirical review of the teacher certification program in Indonesia

- The experience of Indonesia's teacher certification program confirms what many skeptics already assumed: paying teachers double does not make them teach better.
- The impact of teacher upgrading through obtaining a bachelor's degree on student-learning outcomes is modest at best.
- In view of these findings, it is imperative that the process of certification and the certification allowance are tied to demonstrated competencies, rather than to formal degrees alone.
- By controlling the intake into teacher's training institutions—selecting the cream of the crop—Indonesia can ensure that the quality of new teachers improves, gradually raising the quality of education over the long term.

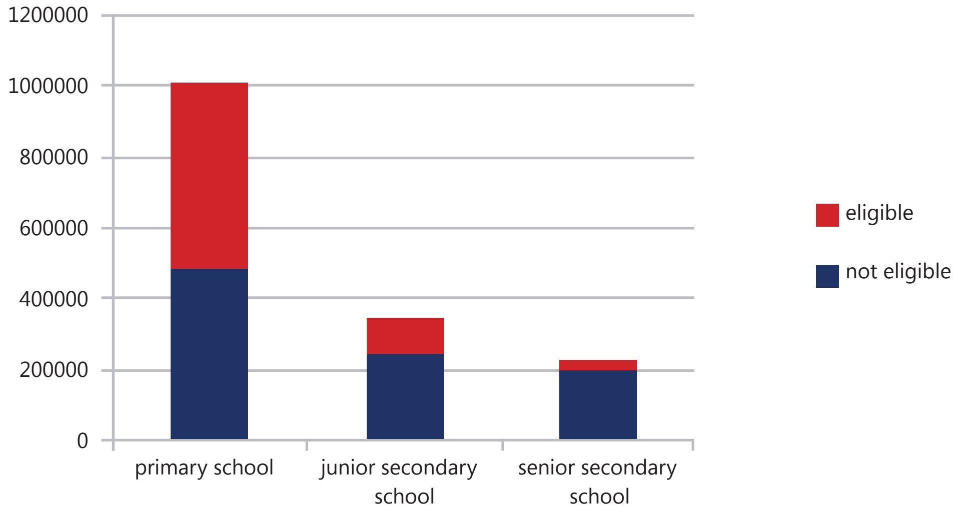
Would acquiring a bachelor's degree from one of the teacher training institutions in the country be sufficient to ensure a minimum level of teacher quality? And, similarly, would large scale academic upgrading to the bachelor's degree level lead to higher quality education in Indonesia in the medium term? Many observers worried that guaranteeing the quality of teacher training programs was going to be challenging in the Indonesian context. This is also why the initial design of the certification program mentioned the importance of explicitly verifying the skills teachers had obtained in the upgrading process. At the outset, therefore, it was unclear whether or not the certification program would lead to improvements in the quality of education. A large body of empirical research has raised skepticism about the benefits of formal academic qualifications of teachers, but such research typically concentrates on more advanced economies where teachers with a bachelor's degree are compared with those who have a master's degree, for example (Hanushek & Rivkin, 2006). The situation in Indonesia is strikingly different, as a quarter of all 2.7 million teachers in 2005 (and a third of all primary teachers) had no more than a high-school diploma. In this context, it is not unreasonable to expect that if this group obtained a university bachelor's degree, the quality of the learning environment would improve.

So, how many teachers needed the upgrade to the bachelor's degree level? The eligibility rules for certification are not completely straightforward, as some exceptions to the standard requirements were introduced. But, generally, in the public school system, civil servants with a university bachelor's degree automatically qualified for certification. In private schools, full-time teachers *(guru tetap yayasan)* with a university bachelor's degree were also automatically eligible. The exceptions were applied, first to senior teachers, who were exempt from obtaining a bachelor's degree. The seniority of teachers was determined by their rank in the civil service (Rank IV), or if a teacher was aged 50 or older and had more than 20 years of teaching experience.[11] Second, district-appointed contract teachers could also be eligible, provided they had a bachelor's degree or were sufficiently senior. These district-appointed contract teachers are a minority and make up less than 1 percent of the overall teaching force. For clarity we ignore these in some of the figures below.

[11] The latter condition is generally non-binding. Teachers aged 50 and older, with 20 years of teaching experience, usually also have a higher rank in the civil service.

Just over half of all teachers were qualified for certification in 2005/06, according to any of the above criteria. Most of the "ineligibility" was in primary schools at that time, where the Teacher Law set the university bachelor's degree for teachers as a new standard.[12] Figure 2 below shows the number of civil service teachers (in public schools) and the number of permanent private school teachers (in private schools) and breaks it down into those who were eligible for certification in 2005/06 and those who were not. About half of all teachers were initially ineligible for certification, and the certification allowance provided a huge financial incentive for them to upgrade to the bachelor's degree level.

Figure 2: Eligibility for certification in 2005/06, by school type



Source: NUPTK Teacher census 2005/06.

Note: This figure looks at civil servant teachers (public schools) and *guru tetap yayasan* (permanent private school teachers) from the NUPTK teacher census and counts how many of them are eligible for certification. About half of the primary teachers in 2005/06 were not yet eligible for certification, effectively needing a bachelor's degree to become eligible.

With such great numbers of teachers needing to upgrade their academic qualifications, the impact of academic upgrading on student outcomes could have been substantial. But academic upgrading is not the only mechanism through which the certification program could have impacted student outcomes. The certification allowance, disbursed periodically upon successful completion of the certification program, meant an effective doubling of take-home pay. This would release teachers from the financial stress they allegedly had felt in the old system, which in turn should have helped them to focus better on their teaching duties. Lastly, it was hoped that higher pay would attract brighter high-school graduates into the teaching profession. While prior to the Teacher Law the best high-school graduates might have opted for careers in engineering or business, the hope was that higher salaries would persuade them to become teachers instead.

The "effect" of certification on teacher quality overall is a combination of three forces working together. The three distinct mechanisms through which the teacher certification program might lead to improvements in the overall quality of teaching are as follows:

The behavioral mechanism. Certification and the associated certification allowance would not only provide recognition but also a significant pay raise. The hope was that teachers who were certified would be more likely to show up for class on time, prepare better for lessons, and forgo outside employment. These changes were expected to improve the quality of teaching and, ultimately, student-learning outcomes.

The academic upgrading mechanism. In order to receive access to the certification program, unqualified teachers had to upgrade their academic qualifications to the bachelor's degree level or equivalent. In the process of upgrading, teachers were expected to acquire new skills that would make them better teachers.

[12] Prior to 2005/06 only a two-year post-secondary-school diploma was required, which, in effect, was not even very actively enforced.

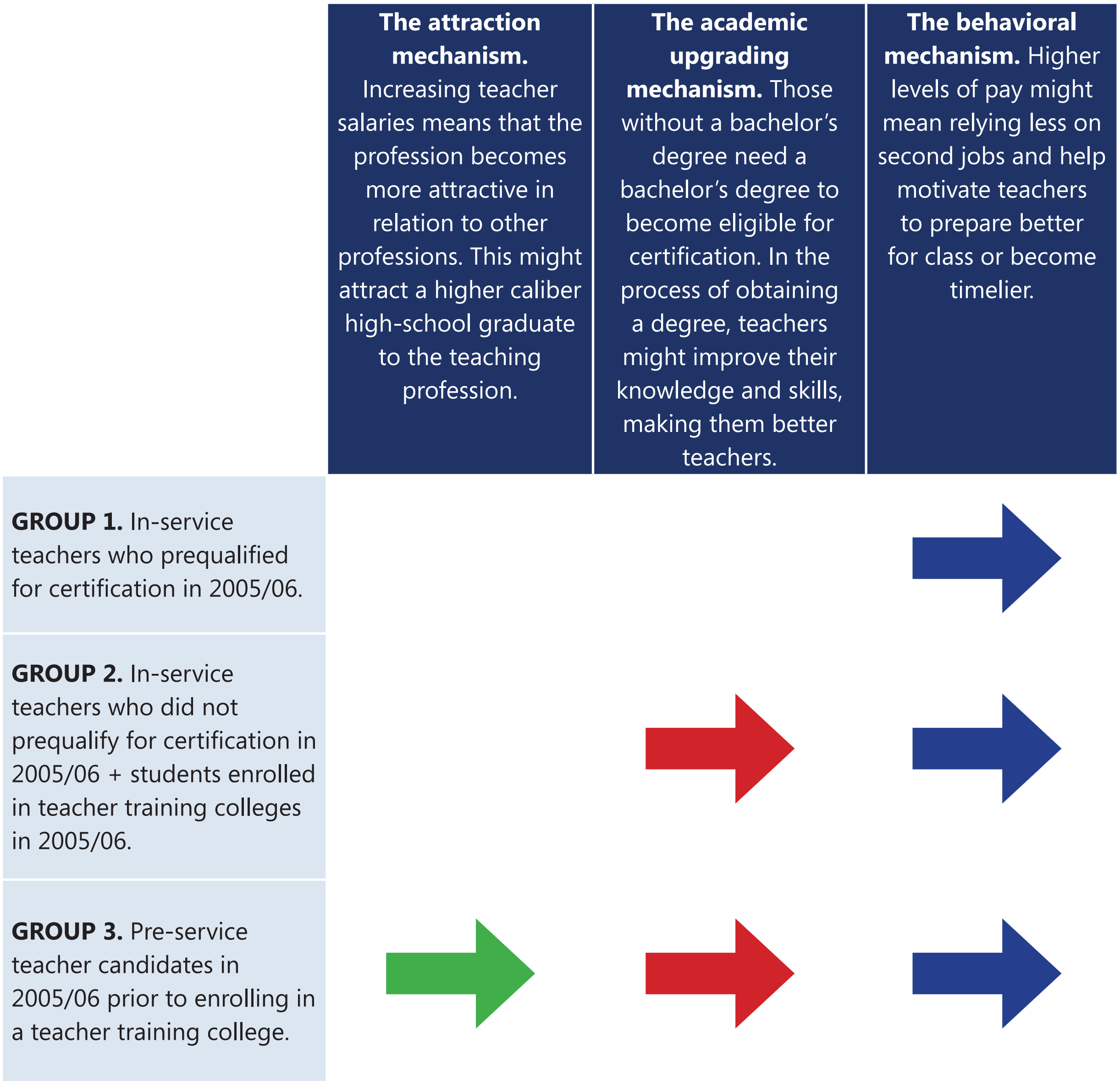
The attraction mechanism. The salary increase, tied to certification, was hoped to make the teaching profession more competitive in relation to other professions. The aim was that it would attract brighter high-school graduates to the teaching profession. For example, the base salary of a junior teacher with a university bachelor's degree roughly translates to the median of the income distribution of all workers with a bachelor's degree in Indonesia, while the base salary and the certification allowance combined translates into roughly the 90th percentile. Clearly, whereas prior to certification teachers were in the middle of the salary spectrum, today they are near the top. Furthermore, Indonesian teachers (at least those in the majority public schools) are civil servants and have a "job for life", inclusive of a pension plan. Teaching in Indonesia is an attractive option for many and certification was expected to motivate high-profile high-school graduates to become teachers, rather than choosing other careers.

The overall effect of certification depends on the combined potency of each of these mechanisms. Not all teachers are affected alike and this matters for how teacher certification may impact student-learning outcomes, today and in the future. Teachers who already prequalified for certification by having a bachelor's degree, for example, are obviously not affected by the "academic upgrading mechanism", as their competency levels were already deemed sufficient. For these teachers the teacher certification program did not mean much more than a doubling of take-home pay, so that they are only, potentially, affected by the behavioral mechanism.

Figure 3 shows schematically how different types of teachers are affected by the three different mechanisms discussed above. We categorize three groups or types of teachers each affected by certification in a different way. **GROUP 1** teachers prequalified for certification when the certification program was introduced in 2005/06. Because their competency levels were deemed sufficient, no additional coursework or training was required. The only way they could be affected, therefore, was by increased professional recognition and a doubling of pay. As such, only the behavioral mechanism applied to them. **GROUP 2** teachers did not prequalify for certification in 2005/06 when the program was introduced. These teachers first had to obtain their bachelor's degrees in order to prequalify. **GROUP 2** teachers are affected twice, in a sense. First, they acquire skills in the academic upgrading process, and second, they become recognized as professionals and receive a higher level of pay. **GROUP 3** teachers are those who were not in the system at all (neither teaching, not enrolled in a teacher training college) in 2005/06. These are high-school graduates who might or might not choose to pursue a career in teaching. The higher pay levels could increase competition for vacancies in teacher training colleges, which could in turn increase the quality of intake into the teacher training colleges. The latter mechanism has the potential to change the entire make-up of the teacher force with time, hopefully leading to better quality teachers.

The overall effect of certification is the sum of the behavioral effect, the academic upgrading effect and the attraction effect. The overall effect of certification therefore will likely change over time, as the influence of the attraction effect becomes stronger as more newly attracted candidates finish their university education and find jobs as teachers. In Sections 2.1, 2.2 and 2.3 of this chapter we evaluate the empirical relevance of each of these three theoretical mechanisms in sequence.

Figure 3: Matrix of the different types of teachers and the mechanisms by which overall performance of the teacher workforce might improve



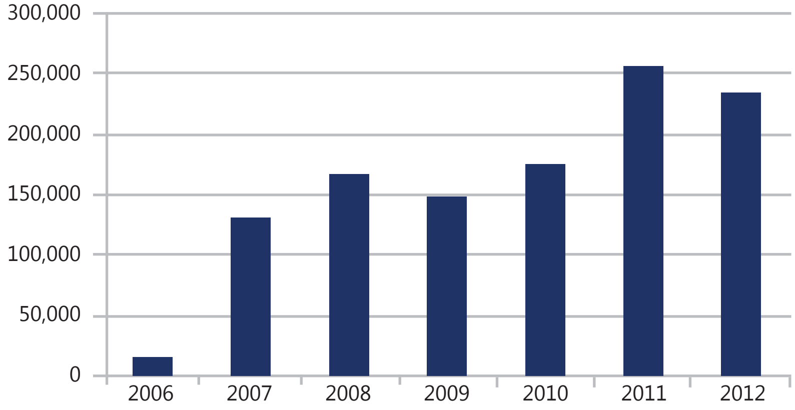
	The attraction mechanism. Increasing teacher salaries means that the profession becomes more attractive in relation to other professions. This might attract a higher caliber high-school graduate to the teaching profession.	**The academic upgrading mechanism.** Those without a bachelor's degree need a bachelor's degree to become eligible for certification. In the process of obtaining a degree, teachers might improve their knowledge and skills, making them better teachers.	**The behavioral mechanism.** Higher levels of pay might mean relying less on second jobs and help motivate teachers to prepare better for class or become timelier.
GROUP 1. In-service teachers who prequalified for certification in 2005/06.			→
GROUP 2. In-service teachers who did not prequalify for certification in 2005/06 + students enrolled in teacher training colleges in 2005/06.		→	→
GROUP 3. Pre-service teacher candidates in 2005/06 prior to enrolling in a teacher training college.	→	→	→

Note: In-service teachers who prequalified for certification in 2005/06, based on having a bachelor's degree or otherwise, are only affected by the behavioral mechanism. Those already in the system (in-service teachers + students enrolled in a teacher training college) are affected by the academic upgrading mechanism first, and second, based on a doubling of pay, also by the behavioral mechanism. Finally, pre-service teacher candidates who have yet to decide whether to pursue a career in teaching or not are also affected by a third mechanism, the attraction mechanism.

2.1. The behavioral channel: Does paying teachers double make them teach better?

Around 1.15 million certified teachers were in the system in 2012. Starting in 2006, the teacher certification program was rolled out across the country at a massive rate of around 200,000 teachers a year. Some objective criteria were developed to determine which teachers were to be certified first. It was observed, however, that these rules were not always followed. Although some problems were reported, huge numbers of teachers are certified today and, in most cases, the certification allowances have been disbursed. Figure 4 below shows the number of teachers in the 2012 NUPTK teacher census who were certified, by year of certification.

Figure 4: The number of certified teachers, by year of certification



Source: Author's calculations based on the NUPTK 2012 teacher census.

Should salaries matter for teacher performance and student-learning outcomes? In the debate leading up to Indonesia's 2005 teacher reforms, some stakeholders argued that higher salaries would improve the performance of incumbent teachers and, hence, student-learning outcomes. For example, after a study revealed in 2004 that 19 percent of teachers in Indonesia were absent from school on any given day (Chaudhury, Hammer, Kremer, Muralidharan, & Rogers, 2006), the chairman of the teachers' union blamed low salaries and poor working conditions for the low observed levels of teacher effort (Santoso, 2004). The union leader argued that the Government could not expect to lower absenteeism until teachers were paid adequately. Other research shows that teacher absenteeism reduces student-learning outcomes (Duflo, Hanna, & Ryan, 2012; Miller, Murnane, & Willett, 2008). If the union leader was right, then higher salaries could indeed lead to improved teacher effort and improved student outcomes.

The argument used by the union leader goes against basic economic theory as it relies on a concept of reciprocity or fairness, where teachers provide additional effort only when salary levels are sufficient. Ideas of fairness have been introduced into economic thinking by Akerlof & Yellen (1990) among others and are sometimes put forward as explanations for why some employers offer salaries above the going market rate. But more standard economic logic suggests that teachers have no explicit economic incentive to do anything differently once they are certified and paid more, for example become timelier, or exert more effort in lesson preparation. This is because the allowance is not tied directly to any of these changes in behavior. Early empirical literature seems to confirm the latter argument and suggests that investments in teacher salaries are not particularly cost effective (see Pritchett & Filmer (1999) for a review).

Whether concepts of fairness are empirically relevant depend also, in part, on whether the professional allowance constitutes a "large-enough amount". To evaluate this we link teachers' salaries to pay levels of other workers via Sakernas, Indonesia's labor force survey. For example, a junior teacher with a bachelor's degree entering the civil service system would have a base salary of about IDR 2.5 million a month, or about US$210 at the current IDR 12,000/US$ exchange rate. The labor force survey indicates that this salary is about average, as 53 percent of all workers with a bachelor's degree earn more. The certification allowance doubles the pay level to about IDR 5 million a month, or US$420 at the IDR 12,000/US$ exchange rate. This is a respectable pay level by Indonesian standards, as now only 10 percent of workers with a bachelor's degree earn more [author's calculations based on Sakernas, 2012]. This basically means that certified civil servant teachers are among the top income-earners in the country.[13] This is especially true in remote rural areas, where salaries tend to be lower on average.

[13] It is important to realize however that there is a sizable group of people in Indonesia that earns amounts far greater than US$420 per month. This group is perhaps hard to reach by field teams such as those who collect data for the labor force survey Sakernas. Potential under-coverage of the rich, or general underreporting of income levels, should be taken into account when interpreting these numbers.

So, did the doubling of teachers' pay lead to better student performance, via improved motivation and effort on the part of the teachers involved? The results of a randomized controlled trial (RCT) show that the doubling of teachers' salaries did not make teachers perform better, as measured by improved student-learning outcomes (De Ree, Muralidharan, Pradhan, & Rogers, forthcoming), confirming early, yet much less rigorous research (Pritchett & Filmer, 1999). But, not surprisingly perhaps, doubling teachers' salaries did have clear welfare effects by reducing reliance on outside employment (second jobs), as well as reducing financial stress among teachers. The complete results of the RCT are presented in De Ree, Muralidharan, Rogers, and Pradhan (forthcoming), but some of their results are discussed below.

Randomized controlled trials are the gold standard for program evaluation by resolving the issue of "selection", which usually emerges when outcomes are compared across recipients and non-recipients of a program. For example, we cannot credibly attribute quality differences between certified and uncertified teachers to the certification itself. The reason for this is that certified teachers are different for a variety of reasons. They are, for example, of higher rank in the civil service on average, and more likely to have bachelor's degrees, etc. Hence, if one observes that certified teachers are better than uncertified teachers it may only be due to the fact that more of them have a bachelor's degree. In the field of program evaluation this problem is called the "selection" problem. The group of certified teachers is a specific *non-random* selection of the population of teachers at large, with entirely different characteristics to the population of uncertified teachers.[14]

The RCT aims to make a more credible comparison than the simple comparison between teachers who are certified and those who are not. How was this executed in Indonesia? The research project started with a representative sample of 240 public primary schools and 120 public junior secondary schools, evenly spread across 20 districts (Figure 1). Within this sample of schools, one-third was randomly selected as the treatment group, while the rest, two-thirds, automatically became part of the control or comparison group. Details of the data are presented in Annex A.

The random division into treatment and control groups is **Step 1** in the execution of a RCT. **The random division ensures that both groups are quantitatively similar, prior to the project's intervention**. This is the key element underpinning the validity of any RCT as a tool for evaluation. **Step 2** in the design is the *intervention*, which should only benefit the schools, teachers and students of the treatment group. In the control group there was no intervention and everything, including the rate at which teachers were certified, remained as it would have without this study, so that their outcomes can be used as a "business-as-usual" benchmark. By comparing outcomes between the treatment and control groups on average, one measures the average effect of the intervention.

The Government, in collaboration with the World Bank, introduced the following intervention to the treatment schools:

> **The Project Intervention:** All the teachers in treatment schools who were eligible for certification in 2009 but not yet certified were allowed immediate access to the certification process. Teachers were eligible for certification if they had their bachelor's degree, or if they were sufficiently senior. The eligible teachers in the treatment group received an official letter from the relevant government body, granting them direct access to the teacher certification program in October 2009, one month before the field teams collected baseline data. The intention of the intervention was that these teachers would be certified in the 2010/11 school year, and start receiving the certification allowance.

[14] A more pressing problem with evaluating the effect of certification is that there are factors that are unobserved. Observable factors like "having a bachelor's degree" can be statistically controlled for. But one reason why certified teachers are different from uncertified teachers could be that those who are certified *first* are "better" according to some subjective criteria than those teachers who are certified *later*.

The general trend was that once teachers were admitted to the certification program, no one, or only a negligible minority of teachers, failed the certification process. Most teachers passed directly, through a successful assessment of a portfolio of past training and experiences, or after completing a 90-hours training course.[15] The certification process therefore does not have a strong skills-upgrading component and is hardly selective in weeding out lower quality candidates. Nevertheless, formally, the experiment evaluates the combined impact of the certification allowance and the certification process.

Figure 5: Differences between treatment and control at baseline (November 2009), one month after the intervention

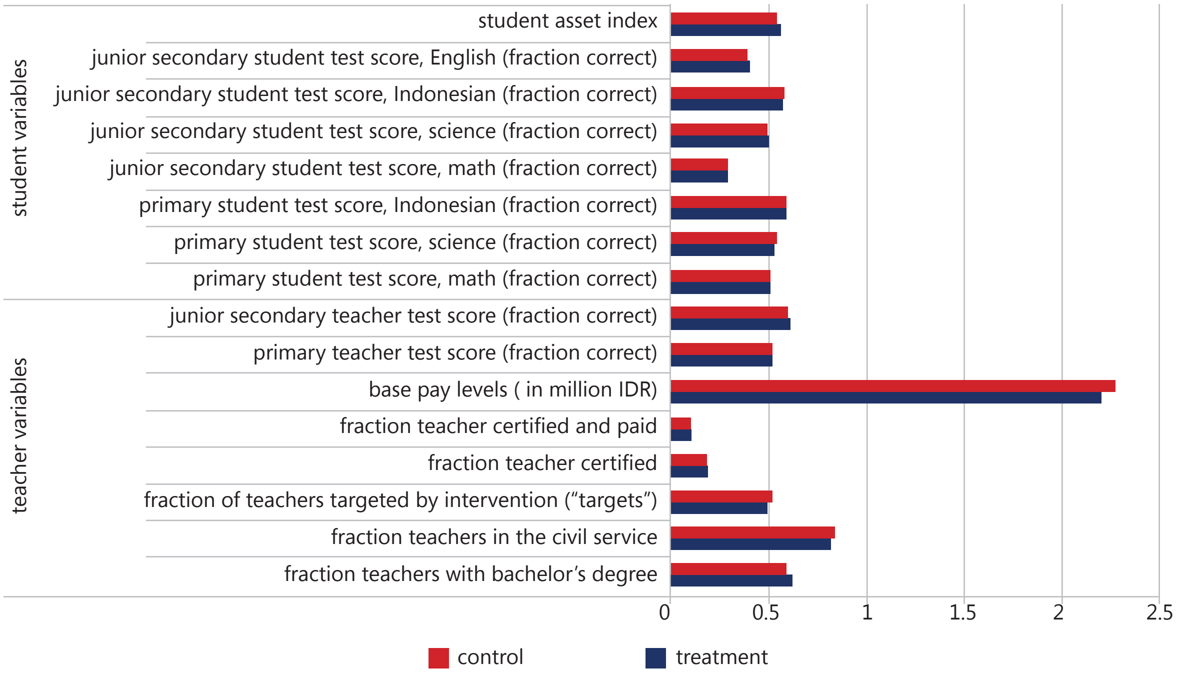


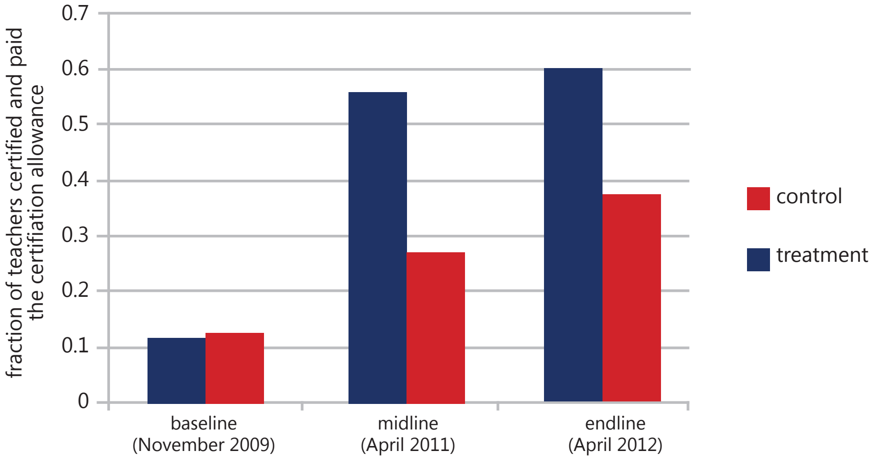
Figure 5 shows that **Step 1** of the RCT was successfully executed: the treatment and control groups were quantitatively similar on average, prior to the project's intervention. Of course, there are some differences due to sampling errors. Teachers from schools in the treatment group, for example, are somewhat more likely to have a bachelor's degree, i.e., 60 vs. 63 percent. By contrast, average base pay levels are slightly higher in the control schools than in the treatment schools: IDR 2.20 million vs. IDR 2.25 million, a difference of about US$ 4 per teacher per month on average. Also for the indicators of quality, such as the subject-matter test scores of students and teachers, minor differences were observed. These differences however are quantitatively small and not statistically significant.

Prior to the intervention, the percentage of certified teachers was similar between treatment and control groups, and slightly under 20 percent. These teachers were certified between the introduction of the certification program in 2006, when the first cohort of teachers entered the certification process, and November 2009, when the field teams visited the sample schools for the first time. Another 50 percent of teachers was eligible but not yet certified in November 2009. In treatment schools, these teachers were directly targeted by the project's intervention. We refer to this group of teachers therefore as "targets". With 20 percent already certified and 50 percent targeted, we would have expected to see 70 percent of teachers certified in treatment schools in April 2011, the time of the second field visit. In the control schools we would have expected to see a substantially lower number of certified teachers, as the project did not intervene there.

[15] The regulations have changed a number of times over the years. Today, all (or most) teachers need to follow the 90-hour training prior to becoming certified. In addition, and prior to entering this training, teachers have to pass a competency assessment. It was observed however that most teachers were passing these assessments even with low raw scores, which has raised concerns about its selectivity.

Figure 5 already showed that there is a difference between being certified and having received the certification allowance. Whereas roughly 20 percent of teachers was certified at baseline, only 12 percent had received the certification allowance. There tends to be a delay between the moment teachers are certified and when they are paid the allowance for the first time.[16] Figure 6 shows the fraction of teachers who were certified and were paid the certification allowance, in the first, second and third round of data collection, and, once again, comparing the teachers in the treatment group with those in the control.

Figure 6: Fraction of teachers who are certified and paid, at baseline, midline and endline



Note: Author's calculations based on three rounds of survey data. The number of teachers interviewed at each round is roughly 3,000.

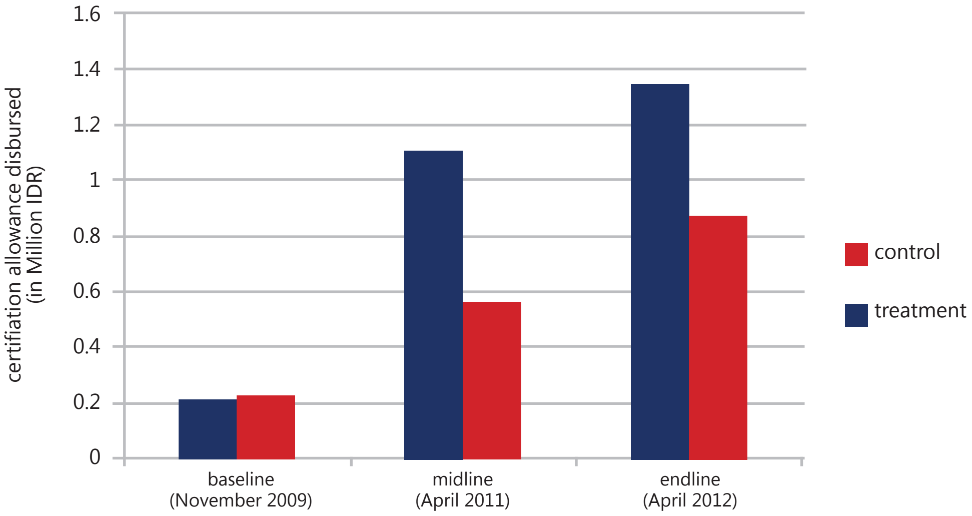
Clearly, as a result of the intervention, many more teachers were certified and paid the certification allowance, both at the second and third visits. In treatment schools around 55 percent of all teachers were certified and paid the allowance in April 2011, versus around 25 percent in control schools. The gap between treatment and control schools extends towards the third and final visit to the sample schools, although it decreases in magnitude. The difference between treatment and control schools narrows over time because control schools steadily catch up with treatment schools. The reason for this is that many eligible teachers in treatment schools are already certified, while those in the control schools are still targeted by the general roll-out of the program across the country. Eventually, therefore, the difference between treatment and control schools will entirely disappear, when all eligible teachers become certified.

How much more money was being disbursed on a per month basis? Figure 7 shows that at the midline, the treatment schools received roughly IDR 550,000 per month *more*, on a per teacher basis.[17] As the intervention only certified and increased pay levels of 30 percent of the teachers in the treatment schools, we can establish that IDR 550,000 per month on a per teacher basis, means roughly IDR 1.85 million per month per *targeted* teacher.

[16] See De Ree, Muralidharan, Pradhan, & Rogers (forthcoming) for more on the timing of payment, and on how this impacts on the overall interpretation of the results.

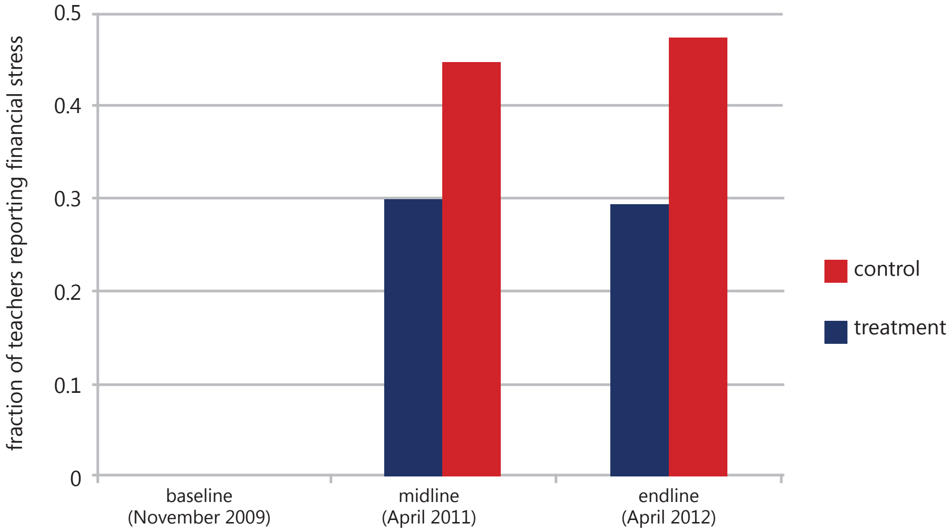
[17] At the endline, the difference decreased somewhat to IDR 500,000.

Figure 7: Certification allowance disbursed per teacher, on a per month basis



The increase in take-home pay had some of the intended welfare effects. Figure 8 shows that financial stress experienced by teachers decreased. This is an important first result, as self-reported financial stress among uncertified teachers was high: around 65 percent of the uncertified teachers reported problems supporting their households financially.

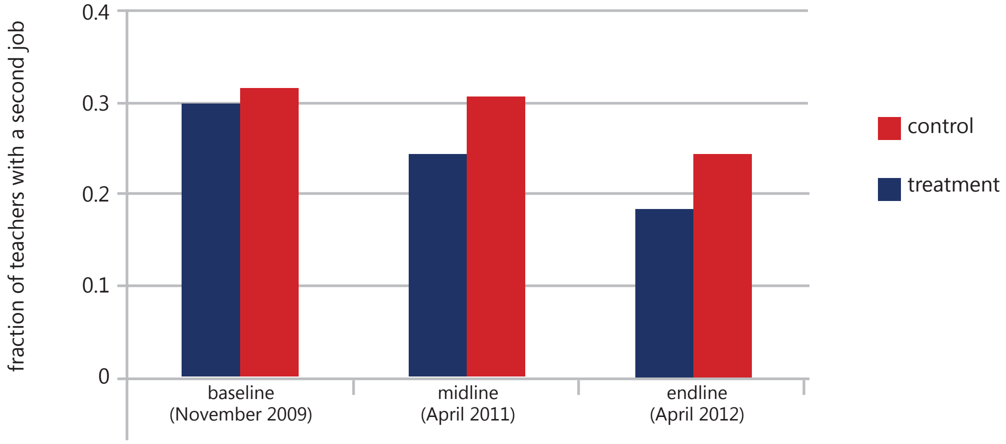
Figure 8: Fraction of teachers reporting financial stress (selected on "targets" only)



Note: This question was not asked at the baseline. The analysis was done for "targets" only, i.e., teachers who were eligible but not certified prior to the intervention.

The boost in take-home pay also led to a decrease in the reliance on outside employment. Despite this, a proportion of teachers still worked in second jobs to complement their income, but the incidence decreased over time, and declined faster in the treatment schools. Most of the decline was due to a decline in second jobs in agriculture, which was the most common type of outside employment in the sample.

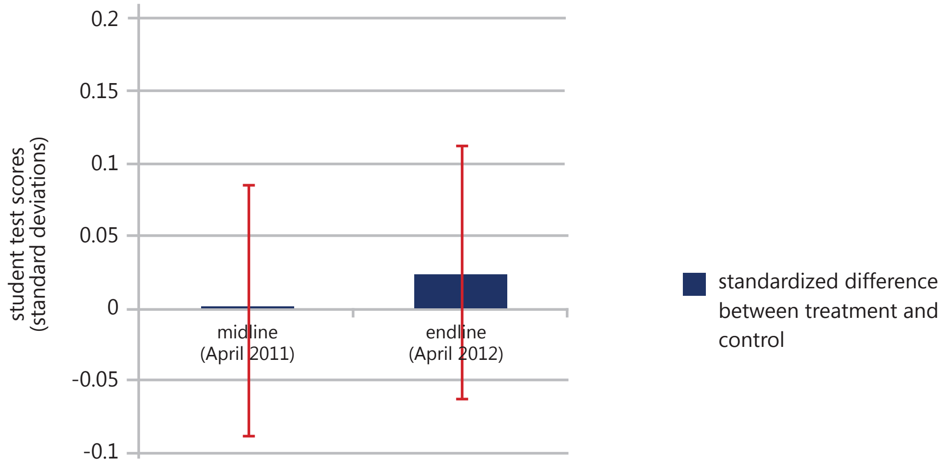
Figure 9: Fraction of teachers with a second job (selected on "targets" only)



Note: Reliance on second jobs decreased, and it decreased faster in the treatment schools than in the control schools.

Teachers left their second jobs and felt more financially secure as a direct result of the increase in take-home pay. Not surprisingly, certification has substantially improved the livelihoods of teachers in Indonesia. It seems therefore that the union leader mentioned earlier in this chapter had a valid point. Prior to certification, teachers had financial problems, and were potentially performing poorly because of this. The certification program removed some of these concerns. The main question, however, is whether teachers have reciprocated by exerting more effort? We evaluate this question by looking at the change in academic achievement of the students they taught. If teachers invested more in successful teaching strategies because they had more time for preparation, background studying, etc., and were less preoccupied with financial problems at home, then the test scores of their students should have increased. From the field teams' first visit to the sample schools in November 2009 until the last visit in April 2012, five semesters (2.5 school years) passed. Because certified teachers were much more prevalent in the treatment schools than in the control schools, students in the treatment schools were taught more frequently by certified teachers.

Figure 10: Student learning outcomes across treatment and control



Note: The reported effects are obtained from a regression of midline and endline standardized student test scores on a dummy variable indicating whether a school is in the treatment group, standardized baseline student test scores (set to zero when baseline scores are not available), a dummy variable indicating whether a student's baseline test score is not available, and a full set of 20 district dummy variables. The estimated effect at the midline is practically zero, while the estimated effect at the endline is 2 percent of a standard deviation. Standard errors allow for arbitrary clustering at the school level and are estimated at around 0.04. This means that these estimates are not statistically significantly different from zero and reasonably precisely estimated. 95-percent confidence bands around the effect estimates are presented. See De Ree, Muralidharan, Pradhan, & Rogers (forthcoming) for the complete analysis, more estimates and details.

Revealingly, the differences in student test scores between the treatment and control schools at the midline and the endline are small and far from statistically significant (Figure 10). Recall that prior to the project's intervention, schools, teachers and students were statistically similar on average. Students emerged from similar socioeconomic backgrounds on average, their schools had similar teachers on average and, consequently, their test scores were also almost the same when the field teams visited the schools for the first time in November 2009. After the intervention, more teachers in the treatment schools became certified, were paid the professional allowance, left their second jobs, and were altogether more financially secure. But the findings of the analysis show that this failed to lead to any measureable difference in student-learning outcomes. The conclusion therefore is that simply paying teachers higher wages does not lead to better teacher performance or student-learning outcomes.

The results of this study have also found their way into the public debate. In a recent article in *The Jakarta Post*, the country's leading English-language newspaper, Professor Yatim Riyanto, an education expert who helped design the certification program early on, mentioned the following: "We initially assumed that a salary increase would encourage teachers to perform better in schools. However, it turned out that most certified teachers have done almost nothing to improve their [teaching] skills or competency, making them no different than uncertified ones" (Widhiarto, 2014).

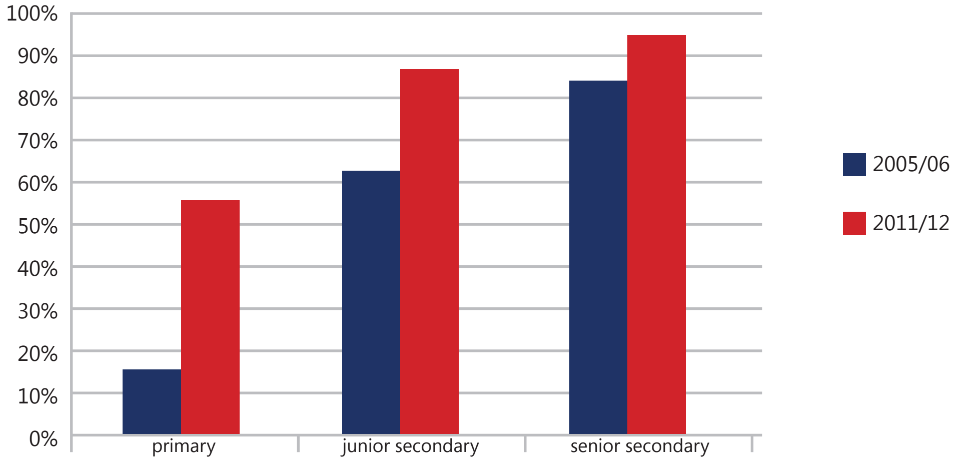
The question that remains is whether teachers did not do better because they did not feel they had to, or whether they tried but were unsuccessful. We cannot really answer this question based on our data, but it is reasonable to assume that the Indonesian teachers surveyed in this study did not perform to the full extent of their potential. Other experimental evidence shows that teachers in developing countries (India in the cited study) can do better if they are sufficiently financially motivated to do so. Maralidharan & Sundararaman (2011) show that if financial rewards are tied to performance (as measured by increased student-learning outcomes), teachers in rural primary schools in Andhra Pradesh (India) tap into some unused potential. It is hard to tell whether Indonesian teachers, as with their peers in India, also have a reservoir of unused potential and whether, if pushed hard enough, they could tap into it. It is clear, however, that the unconditional transfer of the certification allowance was insufficient to activate this latent potential.

2.2. Academic upgrading of in-service teachers: Does it help, and does it help enough?

As a marker for quality, the Government introduced a university bachelor's degree as the single most important criterion for entering the certification program. Behind all this, of course, lay the assumption that once teachers obtained their degrees, they would be better equipped to teach, increasing the quality of education in Indonesia. This is a logical assumption and the results of the empirical analysis, presented in Annex C, suggest that it is true. However, as we explain in this section, bachelor's degrees are not nearly enough to provide the improvement necessary to catch up with more advanced countries in the region, for example, Thailand and Malaysia.

Teachers have responded *en masse* to the new standards laid out in the Teacher Law, and many of them have since obtained a university bachelor's degree. Whereas the majority of teachers in junior secondary schools and above already had bachelor's degrees in 2005/06, the percentage of primary teachers with bachelor's degrees has increased almost fourfold. The latest government statistics from 2012 suggests that 55 percent of primary school teachers have bachelor's degrees, up from 17 percent in 2005/06. Clearly, the prospect of becoming certified and receiving a doubling of pay urged many teachers to take action by investing personal time (and money) in obtaining a bachelor's degree.

Figure 11: Percentage of teachers with bachelor's degrees, by school type, by year



Note: Author's calculations based on the NUPTK 2005/06 and NUPTK 2011/12 data sets. The NUPTK database is an administrative database of all teachers in Indonesia. The data show that the percentage of teachers with a university bachelor's degree has increased markedly.

Annexes B and C present estimates of empirical value-added regressions based on data from the matched student-to-teacher database. The estimates suggest that teachers with bachelor's degrees are better than those without bachelor's degrees. We find this consistently across primary and junior secondary schools. For primary schools the difference is estimated at 0.1 of a standard deviation in student achievement and for junior secondary schools at 0.2 of a standard deviation. While it is not clear whether these findings can be interpreted as being purely causal, we argue that these effects are probably upper bounds on the real causal effects. There are two main reasons for this, which are discussed in Footnote 18.[18]

Regardless of some of the issues around causality, let's accept for now the 0.1 (for primary schools) and 0.2 (for junior secondary schools) of a standard deviation effects as our best estimate of what really happens to student learning outcomes on average when teachers obtain degrees. Proceeding to discuss the implications of this result in a broader sense, it is helpful to think about the overall effect of academic upgrading by considering the following formula:

$$I = P \times V$$

where I is the quantity of interest, the overall impact of an intervention on the population at large. V is the size of the effect of the intervention on an individual student, and P is the fraction of students the intervention applies to. To establish the impact of the recent surge of academic upgrading on the population of Indonesian students we need to know V , how much does a single student benefit from a teacher with a bachelor's degree, and P , how many more students have teachers with bachelor's degrees now, compared to before 2005 when the certification program was introduced.

[18] Most teachers in Indonesia who did not qualify for certification when the certification program was first introduced were given the opportunity to upgrade qualifications through distance-learning courses offered by the Open University (Universitas Terbuka). Most of these teachers, however, did have some post-secondary education, often a two-year diploma (D2), which meant that they needed another two years to complete their degree. But bachelor's degrees acquired through distance-learning courses are arguably of lower intensity than bachelor's degrees obtained in the traditional way, i.e., prior to becoming a teacher in a more residential setting. It is conceivable, therefore, that teachers who obtained bachelor's degrees in a traditional way received better training (or at least more intense training) than those who obtained their degrees through the two-year skills upgrade. The value-added estimates we present in Annex C of this report pool these effects. More generally, it is intuitive that the true causal effect of obtaining a bachelor's degree in Indonesia is smaller than the 0.1 effect presented in Annex C. The argument is that teachers with bachelor's degrees are entirely different people than those without bachelor's degrees. Those with bachelor's degrees are potentially more motivated and/or more able in general, as they had their degrees already when it was not even officially required. This means that part of the reason why teachers with bachelor's degrees are better than those without could be that they are more motivated or more able in general. If this argument is correct, teachers who obtain a degree not out of intrinsic motivation but simply because new regulations require them to do so, should improve less than the current difference in performance between teachers with and without bachelor's degrees.

Figure 11 indicates that about 40 percent of primary teachers and about 20 percent of junior secondary teachers upgraded between 2005 and 2012. So we should not be far off by assuming that 40 percent of primary students and 20 percent of junior secondary students are currently affected by the academic upgrading of their teachers. But because the estimated effect size *V* for primary teachers is smaller, the overall effect on student-learning outcomes in primary and junior secondary schools is similar. Both are estimated at 0.04 of a standard deviation on a year-to-year basis.[19] As these effects accumulate over time, i.e., students benefit each year from a better educated teacher force, we estimate that the effect of academic upgrading is about 0.08 of a standard deviation, about twice as large as the year-to-year effect. In other words, the upgrading of 40 percent of the primary teachers and 20 percent of the junior secondary teachers leads to an increase in student-learning outcomes by 0.08 of a standard deviation on average in the population. These impacts are modest at best, and not nearly enough to catch up with some of the more advanced countries in the region. In fact, an overall increase of 0.08 of a standard deviation would be too small to be even measureable by international trend studies such as PISA. For example, 0.08 of a standard deviation would translate into a mere 6 points on the scale used by PISA.

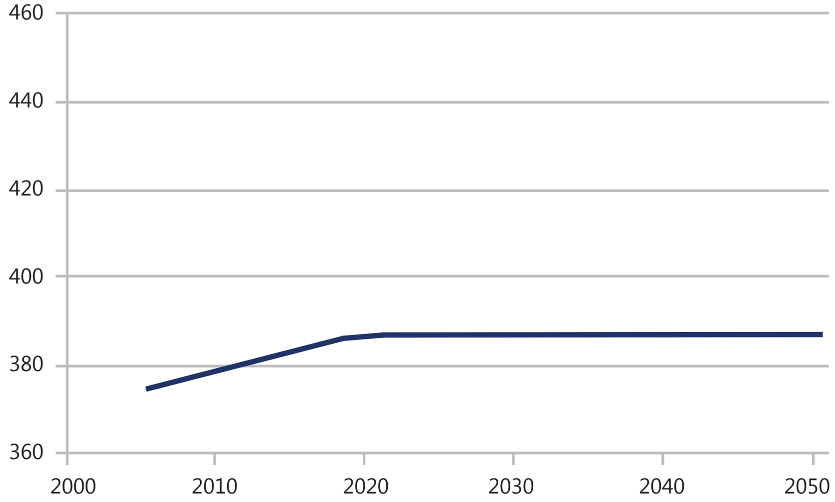
The total effect of certification via the academic upgrading mechanism, however, has not yet fully materialized. Another 45 percent of primary teachers and another 15 percent of junior secondary teachers still do not have a bachelor's degree. This means that school children have yet to experience the full benefits of the upgrading process. Based on our estimates, we can project what *would happen* if all primary and junior secondary teachers have obtained their bachelor's degree. Again the differences between primary and junior secondary schools in terms of the overall effect *I* , are negligible. By using the formula above, we project that the year-to-year impact is 0.08 of a standard deviation.[20] And, as this year-to-year effect is compounded across the full 9-year cycle in basic education, the total effect is roughly double, or about 0.16 of a standard deviation.

Figure 12 projects, based on these empirical findings, what might happen to PISA mathematics scores if all teachers obtained their bachelor's degrees in the next five years. Indonesia's performance on the PISA mathematics component has not changed significantly since 2000. The average score across years is 375, with a standard deviation in the population of about 70. In terms of "PISA points" a 0.16 of a standard deviation increase would be $0.16 \times 70 \approx 11$. Therefore, if all Indonesian teachers obtained a bachelor's degree, we project that PISA scores would increase with 11 points, from the current 375 to 386. We should emphasize that while we use the PISA scale we are not saying that the test score data used mainly in the empirical analysis measure the exact same psychological function as the instruments used by PISA. The scaling, however, provides a ballpark idea of the empirical magnitudes of the effects.

[19] $I = P \times V \approx 0.40 \times 0.1 = 0.04$ for primary schools, and $I = P \times V \approx 0.20 \times 0.2 = 0.04$ for junior secondary schools.

[20] $I = P \times V \approx 0.08 \times 0.1 = 0.08$ for primary schools, and $I = P \times V \approx 0.40 \times 0.2 = 0.08$ for junior secondary schools.

Figure 12: What might happen to aggregate student-learning outcomes when all teachers obtain a bachelor's degree in the next 5 years?



Note: The projections rely on the empirical findings presented in Annex C. It shows what might happen to performance on the mathematics component of PISA, if all teachers obtained a university bachelor's degree. The analysis shows that this would lead to an increase of 0.16 of a standard deviation in the population in the medium term, equivalent to 11 PISA points.

Figure 12 shows that a 0.16 of a standard deviation gain in the long run is not nearly sufficient to catch up with some of Indonesia's more advanced neighbors. Thailand and Malaysia, for example, scored 427 and 421, respectively, on the most recent mathematics component of PISA 2012. Even if all Indonesia's teachers had bachelor's degrees, Indonesia's math scores would still not break through the 400 threshold.[21]

Even in a more favorable scenario, the predicted effects of academic upgrading remain modest. Let's suppose we are wrong in our estimate of the true causal effect of a bachelor's degree. Suppose that the true causal effect is in fact 0.2 of a standard deviation in student learning for primary and 0.4 for junior secondary, and that all teachers without a bachelor's degree in 2005/06 obtained one. The overall effect on the whole population of students, then, would be 0.16 of a standard deviation on a year-to-year basis, and 0.32 of a standard deviation in the long run, when affected students have completed the 9-year cycle in basic education. On the PISA scale, 0.32 of a standard deviation translates into about 22 PISA points. This is a measurable change, and an important one. But despite that, the effect would still not be enough to close half of the learning gap between Indonesia, and its peers Thailand and Malaysia, for example. The results therefore clearly indicate the size of the challenge in generating real changes in student-learning outcomes.

To be able to draw the conclusions above, we have relied on some statistical assumptions and in doing so we may still have under-estimated the long-term effects of academic upgrading. Nonetheless, it is important to emphasize that the evidence presented in this section—based on the matched student-to-teacher database—supports the fact that Indonesia's performance on PISA has hardly improved over the past 12 years since it started participating in 2000, while hundreds of thousands of teachers have since obtained their bachelor's degree. Both the general "macro level" results of PISA and the "micro level" results based on the matched student-to-teacher database arrive at the same conclusion from different angles and based on different data. The evidence suggests that for Indonesia to make these real improvements, teachers must upgrade in their *skills*, first and foremost.

[21] It is important to keep in mind that the projection in Figure 12 manipulates a single parameter: the fraction of teachers with a university bachelor's degree. The projections therefore assume that nothing else is changing in the meantime.

2.3. Attracting better-quality aspiring teachers into the teaching profession

The results presented in Sections 2.1 and 2.2 are sobering. Paying higher salaries does not make teachers teach better and, while the surge of academic upgrading to the bachelor's degree level seems to help somewhat, it does not help anywhere near enough. The findings, however, are in line with findings from academic research more generally. It is commonly found that the degrees or experience of teachers do not correlate strongly with student learning (Rivkin, Hanushek, & Kain (2005); Hanushek & Rivkin (2006); Glewwe, Hanushek, Humpage, & Ravina (2011)). Whether the certification program in its current form leads to strong learning effects in the future, therefore, depends on the strength of a third mechanism: namely, does certification, and the associated professional allowance, attract brighter and more motivated high-school graduates into the teacher training colleges? And if so, is the education system then able to hire these high-performing graduates into schools?

We have the least amount of conclusive empirical evidence for evaluating the effectiveness of this last of the three mechanisms, as most of the effects of certification through the attraction mechanism lie ahead of us. But, despite this, there are some clear indications that the teaching profession has indeed become more popular among high-school graduates. However, it is questionable whether this has led to a general increase in the average quality of students enrolling in teacher training colleges. A survey of students in a sample of 15 teacher training colleges has found that cohorts who enrolled more recently have higher national exam scores than earlier cohorts (relative to the national average of the respective cohort) (Indonesia Ministry of Education and Culture (2009) and Chang, Shaeffer, Al-Samarrai, Ragatz, De Ree, & Stevenson (2013)). This suggests a trend that could benefit Indonesia's school children in the longer term. But it remains unclear how much better these students really are, or will be, in terms of their future performance in the classroom.

It is important to bear in mind that the research cited above compares the quality of intake over time, while keeping constant the number of teacher training colleges. Other data sources show that the number of teacher training colleges itself has been growing, basically to match the increase in demand for vacancies. While the quality of intake in some of the more established teacher training colleges might have gone up, this positive effect may have been offset by the establishment of new centers. As competition for vacancies in some of the more established colleges led to an increase in the quality of intake, it is conceivable that lower-caliber high-school graduates (the dropouts of the initial selection process) now populate these newly established, typically privately operated, teacher training colleges. It is unclear how the rapid expansion of the number of teacher training colleges has impacted the average quality of those enrolled. But if the number of teacher training colleges generally adjusts to match the demand for vacancies, then the system will fail to weed out lower-caliber high-school graduates. This is a worrying development. Other countries in which there is excess demand for vacancies in teacher training colleges use this opportunity to select heavily at the gates. Finland—one of the top-performers on PISA—is an example of this. Currently, Indonesia is not taking advantage of the popularity of the teaching profession to select the most promising high-school graduates out of the total pool of applicants.

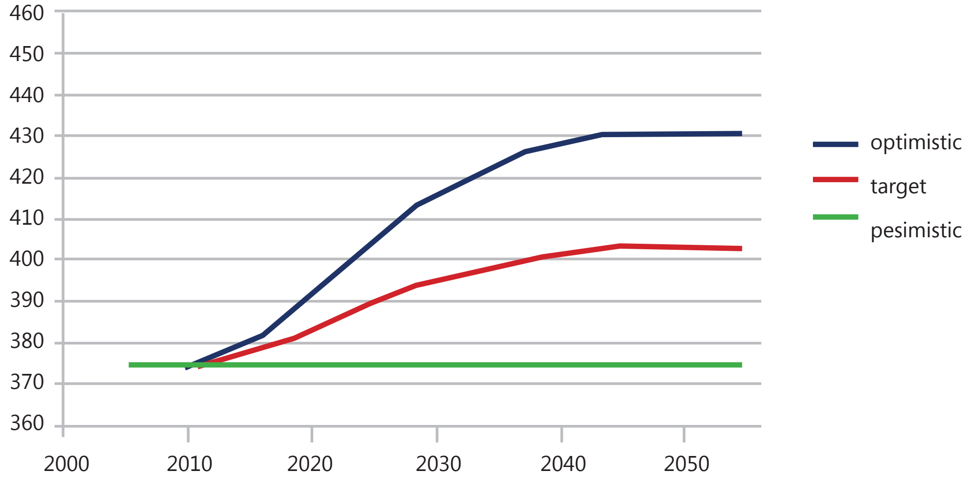
Even if the education system is able to attract a higher caliber of high-school graduates, it is uncertain whether once they graduate they will be provided with jobs as teachers. First, administrative data on enrollment in teacher training programs suggest that there are currently over 1 million students enrolled (excluding in-service teachers upgrading via the Open University). Based on a four-year program cycle, this means that the labor market for teachers needs to absorb about 250,000 university trained teachers every year. As currently only about 50,000 teachers retire each year—a number that is expected to increase to 100,000 each year over the next decade—the market is clearly saturated. Secondly, it is unclear at best whether the current system has the checks and balances in place to hire the best, or even the best trained candidates and provide them with teaching jobs in schools. A worrying statistic in relation to this issue is that only half of the junior primary teachers in the civil service (below age 30) have a university bachelor's degree. In the concluding Chapter 4 we discuss the seemingly inefficient teacher training and teacher hiring procedures in Indonesia in more detail.

The attraction mechanism is potentially very powerful, but it requires well-functioning system of rules and regulations if it is to achieve its full potential. It is particularly unclear how this mechanism will develop over the coming few years. What we can do at this point is make projections based on what could happen in different scenarios. As we do not really know whether the new teachers hired into the system are better than the old cohorts they replace, we consider three scenarios in the projections of Figure 13.

The first projection is unusually optimistic, and assumes that new cohorts of teachers produce 0.4 of a standard deviation of student learning more on a year-to-year basis, than the retiring cohort they replace. Such an effect size is roughly equivalent to the difference in quality between the top 15 percent of teachers in the country (among the very best) and average teachers (see Annex B and Section 3.2 for estimates of the variation in teacher and school quality in Indonesia). The second projection assumes an effect size of half that of the first projection. The incoming cohorts produce 0.2 of a standard deviation of student learning more than the retiring cohort they replace. Such an effect size is more moderate and perhaps a realistic target, being roughly equivalent to the difference in quality between top 30 percent and average teachers. A final projection assumes that new teachers are no better than the retiring cohorts, a particularly plausible scenario in a system that practically does not restrict entry into teacher training colleges, and has a weak system of checks and balances in hiring the best trained teachers from the 250,000 graduates (to replace the 50,000 to 100,000 teachers retiring each year).

The process of replacing the existing teacher force with potentially better quality new intake is terribly slow. When it became clear that the teacher certification program was being introduced in 2005/06, stiffer competition for vacancies in teacher training colleges may have started, and the first cohort of these potentially higher ability candidates could have pushed out the lower ability candidates. Four years later, in 2009, this cohort would have graduated and could have started working as teachers. Each year subsequent to 2009, when a new cohort of these higher-caliber aspiring teachers enters the market, they would replace the respective retirees of that particular year. After 10 years about a quarter of all teachers would be replaced, and it would still take another 25-30 years before the 2009 teacher force is completely replaced by this new breed of potentially better teachers.

Figure 13: Long-term projected effects of better quality teacher intake on mathematics achievement



Note: The projections loosely rely on empirical findings presented in Annex B. These project what might happen to performance on the mathematics component of PISA in the long term, when new cohorts of teachers entering the system are much better (optimistic), just better (target), or not better at all (pessimistic) than the cohorts that retire. In the optimistic scenario, a retiring teacher of average quality is replaced with a teacher performing at a level of the current top 15 percent. The curve in the middle – target—shows what happens when each retiring teacher of average quality is replaced with a teacher performing at the level of the current top 30 percent. The bottom curve – pessimistic— shows what happens when new cohorts of teachers are no better than the retiring cohorts they replace.

Figure 13 presents what could happen to learning outcomes in the long term as retiring teachers are replaced with better quality new teachers. Which one of the three scenarios is most realistic in the current circumstances is by no means certain. What is certain, however, is that the replacement process is slow. The curve is more upward-sloping around 2020, for example, than after 2030. This is (in part) because many more teachers retire in 2020 than in 2030. Figure 13 shows that, in the most optimistic scenario, Indonesian children would score about 0.8[22] of a standard deviation higher in 2045 than they do today, or about 56 points on the scale used by PISA. This would happen when the complete teacher force had been replaced by newly attracted cohorts of (much better) teachers. If this scenario is realistic, Indonesia's student outcomes will reach Thailand and Malaysia's current levels in math by around 2040. In the long run, these effects will close about half of the learning gap between Indonesia and the current average in OECD countries about midway into the century. In the much shorter run, however, for example five years from today in 2020, the effect is much closer to 0.25 of a standard deviation, or 15 PISA points. Under a potentially more realistic scenario, attraction yields roughly 28 points in the long term.

The most optimistic projection only makes sense if new cohorts of teachers are indeed much better than the old ones they replace. But is this likely if the system of teacher training and teacher hiring remains organized in the way it is today? The short answer is: probably not. For the optimistic scenario to be realistic, drastic policy changes are required. The increased interest of high-school graduates in becoming teachers has been met by an increase in the supply in teacher training colleges. This led to an oversupply of university trained teachers competing for jobs. Most urgent therefore is to better regulate the teacher hiring process. With such great numbers of aspiring teachers competing for jobs, it is imperative that only the very best candidates are hired as teachers. For the longer term, the intake into teacher training colleges needs to be controlled, while at the same time a system of licensing and accreditation should only allow the best teacher training institutions to offer courses. If all this happens, then the optimistic scenario could well transpire. But without any of these improvements to the pre-service system, there is no reason to suspect that new cohorts will be better than the retiring cohorts they replace. The systemic changes are discussed in more detail in Chapter 4 where we discuss options for policy.

[22] Two times the 0.4 effect, as the 0.4 year-to-year effect is compounded as students benefit each year as they progress through the 9-year cycle in basic education.

CHAPTER 3

Beyond certification: what matters and what doesn't matter for student learning?

- Because improvements in the pre-service supply of teachers take such a long time to have a measurable impact, there is no alternative but for Indonesia to also upgrade the quality of the existing teacher force.
- Our findings show that Indonesia has some excellent teachers, but also many that are underperforming.
- The research shows that a realistic and achievable way of improving teacher quality and therefore student-learning outcomes would be to improve teacher's subject-matter knowledge. Current subject-matter knowledge of primary teachers is poor on average, but realistic improvements could greatly boost learning outcomes in the medium term.
- Rigorous subject-matter assessments can be built into teachers' in-service career progression.

The previous chapter has shown that, since its implementation in 2005/06, certification has not led to great gains in learning outcomes and it is not at all obvious that it will lead to changes in the future. And, even if Indonesia is able to fix the pre-service system of teacher training and hiring, it would still take decades before improvements in intake lead to substantial changes in student achievement scores. Can Indonesia afford to wait so long? It seems not. High economic growth rates are bound to level off if the education system is not able to meet the demands of an ever-changing labor market (World Bank, 2014). Indonesia, therefore, cannot avoid striving to increase the performance of its in-service teachers. To better understand what Indonesia's Government might be able to do to help transform the education system, this chapter digs deeper into what is happening inside Indonesian schools and classrooms. The analysis is based on the matched student-to-teacher database that we have used extensively throughout this report.

3.1. How much do Indonesian students learn in a year?

Our first observation from the data is that the academic performance of Indonesia's school children improves somewhat substantially as they age.[23] This should not come as a real surprise given that children mature physically and mentally and spend much of their time in school. This finding is also a constructive starting point in trying to understand where these learning gains come from. Indeed, students may not always learn much in school, but learning, in its broadest sense, still takes place.

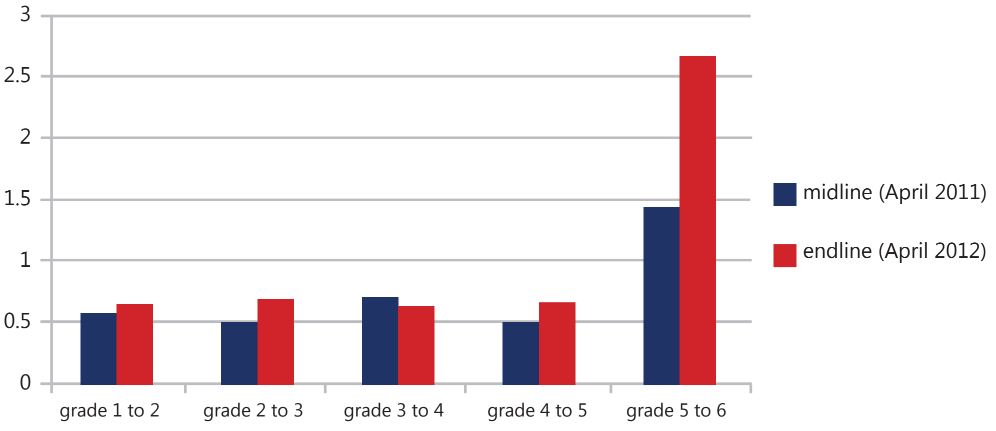
How do we know that children learn? With the design of achievement tests it is often implicitly assumed that children learn. Tests for grade 6 are *designed* to be more difficult than tests for grade 5. In our sample we also test different grade levels with different tests. Although this is standard practice, it is problematic to make quantitative judgments

[23] The Indonesian Government sees a role for the schools in the academic as well as the moral development of children. In this report we restrict ourselves solely to the development of the academic achievement of children.

about learning levels of 5th graders and 6th graders when they are tested with different instruments. Raw scores obtained from different tests are not directly comparable. The instruments used to collect testing data for our matched student-to-teacher database resolve this problem by including some of the same questions in two "adjacent test forms". With adjacent test forms we mean tests that are a grade level apart, or that are for the same grade but collected in a different year. For example, some questions, typically a handful, are the same between a grade 5 and a grade 6 test at the midline or the endline. In this way, the test forms are effectively linked. These questions, or items as they are commonly referred to in the testing literature, are called linked items, or anchor items. While it is usually not permitted to compare raw scores from different tests, we can in fact directly compare the raw scores on the anchor items. Annex D describes in detail how this can be done.

Figure 14 zooms in on a sample of primary students. We use two rounds of the survey, the midline (collected in April 2011), and the endline (collected in April 2012), and compare the raw scores on the anchor items from one grade level against the next. For example, the first blue bar in the figure compares the raw scores on the anchor items between grade 1 and grade 2 at the midline. The first red bar compares the raw scores on the anchor items between grade 1 and grade 2 at the endline. Students in second grade score on average about 0.6 to 0.7 of a standard deviation higher on the exact same questions than students in first grade (who are one year younger on average). **The "0.7 of a standard deviation" means that the average student in second grade performs at roughly the level of the top 25 percent of first graders.** This clearly indicates that children learn in a year. The differences we find between adjacent grades are stable, until we compare grade 5 to grade 6 scores, where we find larger gains.

Figure 14: Learning gains from grade to grade



Source: Author's estimates based on the matched student-to-teacher database. See Annex D for more results.

The learning gains we observe in the data are achieved as the young mind matures and is fed by inputs received at school, at home, and in the world between the two. Figure 14 suggests some variation across grade levels, as gains in the last grade of primary school seem far greater than in earlier grades. One should take this as a first indication that children somehow gain more as they go from grade 5 to grade 6, than in earlier grades. On the other hand, large effects might also emerge if the linked items used are too specific to the curriculum. For example, if a topic of the grade 6 curriculum is used as a linked item between grade 5 and 6 we would expect particularly poor performance in grade 5. We also find that test reliability rates for these scores are particularly low, which is another cause for concern. Consequently, we would encourage further research into this issue, potentially also with these data. However, the analysis presented here is the first of its kind in Indonesia and it documents learning gains that could be somewhat substantial. The take-home message therefore is that aside from the findings in the last year of primary school, the findings are reasonably stable and consistently show important year-to-year improvements in learning outcomes on average.

At first glance perhaps these findings seem to somewhat contradict reports by others, who argue that learning profiles in developing countries are often "shockingly low" (Pritchett & Beatty, 2012). But both findings may indeed be consistent with one another. Annex D shows that raw (percentage correct) scores on the anchor test between grade 3 and grade 4 increase from about 45 percent in grade 3 to 55 percent in grade 4, intuitively a somewhat small gain of 10 percentage points. But because we estimate that the spread of true achievement levels in grade 3 is not particularly wide either, a 10-percentage-point change still translates into 0.6 to 0.7 of a standard deviation.

A potentially different way of looking at a learning gain of 0.7 of a standard deviation is by referring once more to the international comparisons of PISA. The performance of Indonesian 15-year-olds in Indonesia, for example, is roughly 100 PISA points below the OECD average, and differs somewhat depending on whether one looks at math, science or literacy scores.[24] The standard deviation of PISA scores in the Indonesian population is roughly 70 PISA points. In other words, Indonesian 15-year-olds are about 100/70=1.4 Indonesian standard deviations behind the OECD average. With 0.7 of a standard deviation of *learning in a year*, the results from Figure 14 suggest that Indonesian 15-year-olds are about two years behind the OECD average in terms of academic achievement. Notice that the psychological function that is measured by the PISA test may deviate from what is measured by the testing instruments used in our matched student-to-teacher database. The analogy between PISA and the findings based on the survey data should therefore be used with some level of caution.

3.2. Differences in quality across schools and teachers in Indonesia

The previous section documents how much children improve *on average* as they progress through primary school. In this section, we use the matched student-to-teacher database again to look at the variation in learning gains *between* Indonesia's schools and teachers. We measure substantial differences between teachers and schools across Indonesia. Good schools with good teachers produce a good deal more learning than lower-quality schools with lower-quality teachers. In addition, we estimate that anywhere between one-third and two-thirds of the variation in student-learning outcomes can be attributed to differences in the quality of schools and teachers. The remainder, also anywhere between one-third and two-thirds therefore, is due to parental inputs, societal inputs, and individual abilities and motivations, i.e., elements outside the control of schools.[25] The results show that teachers and schools can, and do, make a difference in the lives of many Indonesian children, for some in a good way and for others in a bad way. But the estimates also indicate that other inputs, for example, influences of parents and the broader community also contribute for an important part to the learning outcomes of Indonesian children. The technical details of the analysis and some more results are presented in Annex B.

Another conclusion from the analysis is that there are high quality teachers and schools in Indonesia. It is important to realize this, as it shows that the system is able to produce high-quality teachers and retain them in schools. Successful schools and teachers therefore, are not something alien to the Indonesian education system: success is already there. Indonesia has good schools and teachers; it just does not have enough of them, which is part of the reason why average student performance is low.

Whether a teacher is good or bad is determined by the amount of *value* they are able to *add* to the learning of a child in a given year. When theorists in the field of human capital development talk about good teachers and bad teachers, they therefore use the term "value-added", where high value-added teachers are those that add significantly to a student's academic achievement. Statistical models used to quantify the amount of learning schools and teachers add each year to a student's academic achievement are therefore called "value-added models". Value-added models follow logically from theories of skill formation, where academic achievement is perceived as an aggregate of all past inputs from birth to the most recent (Todd & Wolpin, 2003). Many of the technical results presented in the annexes rely on this methodology.

[24] The latest PISA 2012 results are 375 on math, 396 on reading, and 386 on science.

[25] We cannot be precise on the values mentioned here because of a statistical technicality (see Annex B for a more in-depth discussion), but the findings seem broadly in line with results of a meta-study (Hattie, 2013).

Annex B describes in detail the statistical analysis performed on the matched student-to-teacher database to measure the heterogeneity in quality across Indonesian classrooms, or "classroom value-added". The analysis follows recent advancements in academic research. In particular, the analysis performed shows similarities with Aaronson, Barrow, & Sander (2007) and Jacob & Lefgren (2008). It is however the first of its kind within the Indonesian context. The variation in year-to-year learning across classrooms of the same grade is substantial, indicating that some teachers and some schools are doing much better than others. However, we should be clear about what the empirical analysis is able to establish and what it is not. For example, we cannot statistically distinguish teacher level value-added and school-level value-added. Part of the reason for this is that there is a problem of definition. What if, for example, a principal tends to hire better teachers? It is of course due to the better teachers that the classes are doing better academically but without the principal (a school-level variable) these better quality teachers would not have been hired.

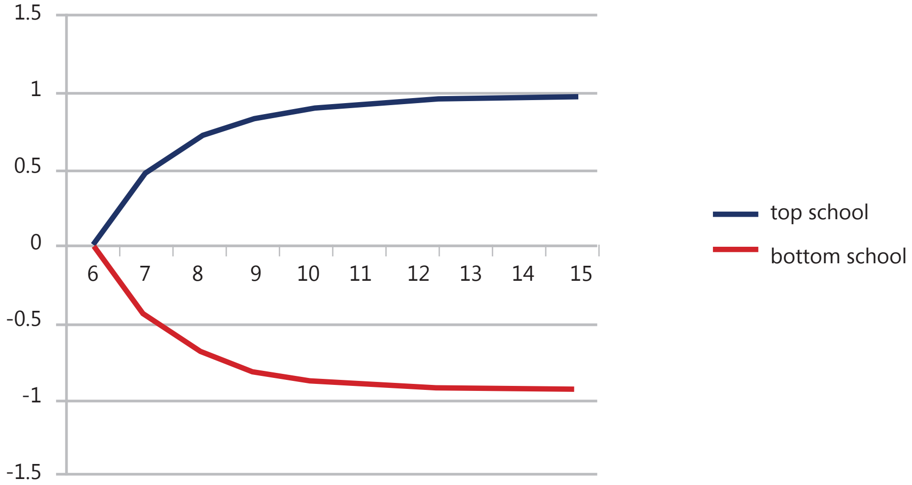
Instead, what we *can* establish empirically is that the differences in learning outcomes observed across Indonesian classrooms are *not* due to "selection effects". It is conceivable for example that more able students, or those who are better supported at home, tend to populate the best schools with the best teachers. If such classes do well, it might not only be because teachers are better, but also because the students have more potential or receive more support at home. The results indicate, however, that whereas students with wealthier parents do better each year, this does not affect estimates of the variation in classroom value-added effects.

The results indicate that some very good things are happening in some classrooms and some worse things in others. For example, the difference between a good classroom (say, with a teacher in the top 10 percent) and an average classroom constitutes roughly 0.47 of a standard deviation in student learning in a given year. To put this into perspective, this is more than half of the difference between the academic achievement of 4th graders and 5th graders. This difference materializes after only a single year of schooling. But what happens if students have two years of these good teachers in a row, or three? The value-added effects are compounded when students benefit from better teachers year after year. This year-to-year compounding, however, happens imperfectly. Having two years with two good teachers in a row does not mean $0.47 + 0.47 = 0.94$ of a standard deviation of additional learning. Instead, the gain (in standard deviation terms) is greatest in the first year with a good teacher. But because the student has gained so much in the first year, an equally good teacher in the second year has more difficulties increasing learning levels (again, in terms of standard deviations in the population) even further.[26]

Regardless of these technical considerations, students with *good* teachers for a couple of years in a row will quickly outpace students with *average* teachers over those same years. The difference after three years, for example, is more than the difference in achievement between an average 4th grader and an average 5th grader. These numbers clearly emphasize the important role teachers and schools can play. Figure 15 shows graphically how this works, and how important schools can be for the academic achievements of students. In the diagram we present what could have happened to two otherwise identical children (with assumed average levels of innate ability) after they enrolled in primary school around nine years ago, at the age of six. Let's suppose that one of the two attended a bottom 10 percent primary and secondary school with associated lower-quality teachers, whereas the second child attended a top 10 percent primary and secondary school with associated higher-quality teachers.

[26] In fact, last year's value added effect is only "worth" half of that today. Our estimates of the time persistence parameter are similar to findings from other researches, for example, Andrabi, Das, Khwaja, & Zajonc (2011), who rely on a similar empirical strategy. Perhaps the most compelling evidence for the idea that knowledge gains appear to fade out are experimental results of Kane & Staiger (2008) for example. See Annex C for more results on the imperfect compounding of value added effects.

Figure 15: Projected differences in learning outcomes across primary and junior secondary school of two, otherwise identical students



Note: The figure projects what would happen to two otherwise identical students, where one of them enters a top 10 percent school, and the other a bottom 10 percent school. Differences in achievement would emerge rapidly in the early years of primary school and converge towards a high and a low level respectively, by the end of junior secondary school. After 9 years of basic education, these otherwise identical students, are 2.0 standard deviations apart in the student achievement distribution.

Two otherwise identical 6-year-olds would reach entirely different levels of achievement at age 15, depending on whether they attend a top-quality school or a bottom-quality school. That is, independent of their respective backgrounds and the ability levels they started out with, and solely due to the quality of the schools they attended. The difference would be roughly "2.0 standard deviations" in the population of achievement scores at age 15. We refer once more to the distribution of performance of Indonesian 15-year-olds on the PISA tests, to make clear what "2.0 standard deviations" really means. If the two students in the example above are of average ability, home support, etc., the unlucky child would normally end up at the bottom end of Indonesia's achievement distribution. More specifically, his or her performance level would be rated as "below level 1" by the PISA experts, where level 1 is its most basic proficiency level. This student most likely would have difficulty applying even the most basic mathematical concepts in real world situations. On the other hand, the other student would reach "level 2" in the PISA qualifications, the level of proficiency that PISA deems is "needed to participate effectively and productively in society", with a predicted PISA score of 435 in mathematics.

This goes to show that Indonesia has good schools and good teachers—it just does not have enough of them. Schools and teachers can make or break a child's future prospects on the labor market and in life more generally. Another important conclusion of the assessment here is that good schools and teachers do exist in Indonesia, and these are not only the top schools targeting Indonesia's small elite. This is important contextual information, as it is certainly not the case that all schools perform at a low level. This means that real change does lie within reach, somehow.

3.3. Which observable characteristics of teachers matter?

The previous two sections established that Indonesian children advance while in school and that a substantial portion of the variation of the learning outcomes can be attributed to the schools they attend. This section aims to go further in trying to understand what makes for a good teacher. In the analysis, we have a particular focus on the relative importance of *indirect indicators of quality*, such as bachelor's degrees and teacher experience, versus *direct indicators of quality*, such as *demonstrated* levels of subject-matter knowledge. Previously, we have discussed that the Government considered a bachelor's degree as marker for quality and the key requirement for its certification program. Section 2.2 provided the empirical evidence that even if all teachers obtained a bachelor's degree (at least in the current system of teacher training) no great gains in learning can be expected. This is a policy sensitive result and suggests that there are indeed problems with the quality of teacher training in Indonesia. But it also suggests that the bachelor's degree as the single most important condition for professional certification is dubious. The analysis presented here shows that more useful measures of teacher quality can be used as requirements for career progression and certification. It provides a basis on which more effective policies can be designed. Chapter 4 discusses these policy options in more detail.

Students of teachers with a bachelor's degree learn more than students of teachers with lower level qualifications, in primary schools about 0.1 of a standard deviation on a year-to-year basis (see Annex C for a discussion of the empirical analysis and detailed results).[27] This finding should be familiar by now as we have used it in Section 2.2 for making projections about aggregate student achievement in the population, in response to the surge of academic upgrading currently taking place in Indonesia. Although the size of these effects is not particularly large, it does indicate that teachers with a bachelor's degree are better teachers. Further analysis into the robustness of this finding did not raise any doubts about the validity of this conclusion (see Annex C).

But, as already mentioned in Section 2.2, the difference in terms of value-added between the teachers with and without a bachelor's degree is only small to moderate. By combining this result with the results from Section 3.2, we conclude that the differences in the academic status of teachers cannot explain the differences in teacher and school quality observed across Indonesian classrooms.[28] This means that many classrooms operate very well, even when teachers do not have a bachelor's degree and, likewise, that there are many underperforming classrooms where teachers do have a bachelor's degree.[29]

To the informed reader, this result may not come as a real surprise. A large amount of academic literature already suggests that indirect proxies of teacher quality, such as academic qualifications or seniority, are only weakly related with student-learning outcomes (Rivkin, Hanushek, & Kain (2005), Hanushek & Rivkin (2006) and Glewwe, Hanushek, Humpage, & Ravina (2011)). Direct proxies have been found to work better in discerning high-quality teachers from lower-quality ones (Hanushek & Rivkin, 2006). In the following paragraphs, we discuss this issue empirically for Indonesia by organizing a "statistical horse race" between the bachelor's degree, an indirect measure of competency, and the subject-matter knowledge of teachers, a direct measure of competency. We are able to evaluate which attribute of teachers—a bachelor's degree or the level of subject-matter knowledge—is a better predictor of student-learning outcomes. The technical background behind the approach is discussed in Annexes B and C.[30]

The results are clear-cut, even more so for primary schools. Observed levels of subject-matter knowledge are much better at predicting student-learning outcomes than academic qualifications. In fact, in ideal circumstances[31], differences in teachers' levels of subject-matter knowledge can account for half of the variation in student-learning outcomes observed across Indonesian primary classrooms, much more than the mere 13 percent that can be attributed to teachers having

[27] About two-thirds of the primary teachers in our sample without a bachelor's degree have a 2-year post-secondary diploma. Another one-third has only a (senior) high school diploma.

[28] Because about 60 percent of primary teachers in our sample have a bachelor's degree, the variation explained is :
V (0.1 × bachelor's degree) = 0.01 × V (bachelor's degree) = 0.01 × (0.6 × (1-0.6))= 0.0024 The standard deviation of the portion explained out of a teacher's academic qualifications is the square root of $\sqrt{0.0024} \approx 0.05$, or about 13 percent of the standard deviation of the estimates classroom effects (0.36).

[29] Whereas the data seem to justify the conclusion that teachers with bachelor's degrees are better teachers than those without, it does not necessarily follow that teachers with bachelor's degrees are better because they have bachelor's degrees. It is quite plausible, for example, that intrinsically more motivated, or more able, teachers are more likely to obtain degrees in the first place, and that it is this underlying characteristic of teachers with bachelor's degrees that is the real reason why they are better teachers overall. This matters for a discussion on academic upgrading, one of the main mechanisms through which certification was hoped to have an impact on learning outcomes. In fact, we might conclude that the real, causal impact of obtaining a bachelor's degree is smaller than the effect sizes reported in this section.

[30] Column (2)-(3) of Table 3 presents the results for primary schools and column (9)-(10) for junior secondary schools.

[31] With ideal circumstances we mean that when teacher's subject matter knowledge was perfectly observed, that is with 100 percent reliability. At lower reliability levels observed test scores account for less of the variation in learning outcomes across classrooms.

bachelor's degrees.[32] In other words, subject-matter knowledge, provided that it is measured precisely (using high-quality tests), is a powerful indicator of teacher quality overall, much more powerful than a bachelor's degree. It should be noted however that noisy test scores (with imperfect reliability) would not do as well in singling out good teachers.

Another finding is that, by taking a teacher's subject-matter knowledge into account, having a bachelor's degree is no longer statistically important. In other words, results from good subject-matter tests are excellent at singling out good teachers, regardless of whether a teacher has a bachelor's degree or not. Direct measures of a teacher's competency, i.e., test scores, are far more accurate than indirect proxies such as bachelor's degrees or experience.

Especially when we look at primary school data, these results are generally insensitive to changing the empirical specifications. The results discussed above, for example, are insensitive to the inclusion of measures of the socioeconomic background characteristics of children (which alleviates some of the concerns of endogenous matching of wealthy students to better-trained teachers). For primary schools the results are also robust to the inclusion of district and school level fixed effects. This means that even within schools, these large differences in learning outcomes can be observed between teachers with high levels of subject-matter knowledge and those with lower levels. For junior secondary schools however, we do find that it matters whether we look within districts or within schools. For junior secondary schools, therefore, cannot draw firm conclusions on the importance of subject-matter knowledge.

This analysis provides powerful evidence, especially for primary schools, that teachers with high levels of subject-matter knowledge are better teachers on average. This is an important result. It is likely that these correlations can also, at least to some extent, be treated as causal effects. For example, based on the results of videotaping Indonesian classrooms, World Bank research has shown that Indonesian teachers with higher levels of subject-matter knowledge use their knowledge to apply a wider range of teaching practices, and also use similar practices more effectively (Ragatz, forthcoming). Interestingly, many of the ways in which high subject-knowledge teachers teach differently appear to be directly related to their levels of subject-matter knowledge. High subject-knowledge teachers, for example, tend to use more open-ended questioning. Using open-ended questioning effectively requires an understanding of different and more elaborate ways of approaching a mathematical problem. It is quite possible, therefore, that if their subject-matter proficiency were to improve, teachers would start making use of a wider array of techniques, and become better teachers.

However, high levels of subject-matter knowledge may also signal higher levels of ability or motivation more generally. If this is the case, it may not be that the subject-matter knowledge itself makes the difference but rather the underlying difference in ability or motivation. To address this concern, the matched student-to-teacher database allows digging one level deeper into the test score data of students and teachers. Both teachers and students in the database were tested in three subjects, namely math, science and the Indonesian language. We can therefore investigate a student's *relative* performance in these subjects and compare them to a teacher's *relative* performance in the same three subjects. Indeed, if teachers' subject-matter knowledge were truly important, we would observe that students progress faster in math than in other subjects when their teachers are relatively better at math. This idea has been used before in a different setting by Metzler & Woessman (2012) and separates out the effects of subject-matter knowledge from other teacher competencies. See De Ree (forthcoming) for more results.

In doing so, we still find support for the idea that a teacher's level of subject-matter knowledge is important. This finding strengthens our confidence in drawing conclusions about the causal relationship between teachers' subject-matter knowledge and student-learning in Indonesia. Efforts to increase levels of subject-matter knowledge of teachers therefore, would make teachers more effective. Column (8) of Table 3 presents the results of the approach. We estimate

[32] The variance of the teacher's subject-matter knowledge score is 1 by construction. The estimated parameter on the subject matter effect is about 0.2. The variance of the prediction out the variation in teacher's subject-matter knowledge is :
V (0.2 × subject matter knowledge) = 0.04 × V (subject matter knowledge) = 0.04 × 1 The square root of that, naturally, is 0.2, which is about half of 0.36, the estimate of the standard deviation of the classroom effect.

that a 1.0 standard deviation increase in the subject-matter knowledge of a teacher yields 0.15 of a standard deviation in additional student learning on a year-to-year basis. [33] If all teachers improve by 1.0 standard deviation, the effects of higher knowledge are compounded as students benefit year after year. The overall effect, in the medium term, would therefore be about 0.30 of a standard deviation.

Before further discussing the reality of improving levels of teachers' subject-matter knowledge in Indonesia, we briefly mention some of the additional empirical results from Annex C. There are positive correlations between student learning and a teacher's age (older teachers appear to do better), and students background characteristics (those from wealthier backgrounds do better). On the other hand, we do not find relationships between learning and class size, or whether or not teachers are civil servants.

3.3.1. The reality of improving the subject-matter knowledge of teachers

The improvements in teachers' subject-matter knowledge that would be needed for real, measurable changes in aggregate student performance are not particularly extreme or unrealistic. We have found that a 1.0 standard deviation increase in teachers' subject-matter knowledge, established across the board, among all teachers in Indonesia, would yield roughly 0.3 of a standard deviation in additional student achievement in the medium term.[34] This is equal to an increase of about 20 points on the PISA scale, which is substantial. But "standard deviation increases in teachers' subject-matter knowledge" are not necessarily intuitive quantities, as they only refer to how teachers do in relation to their peers. In this section we provide a more intuitive account on what these findings mean in everyday life.

What is the level of teachers' subject-matter knowledge in Indonesia today, and what does a standard deviation increase actually mean? We answer these questions by looking at two questions from the teacher subject-matter test used in our matched student-to-teacher database—a sample of roughly 1,700 primary school teachers. The questions were selected to show a pattern. The first question is relatively straightforward while the second is more difficult.

Question 1: *Abi has twice as many marbles as Budi, while Doni has five times as many marbles Abi. If Abi has 10 marbles, how many marbles do Abi, Budi, and Doni have combined?*

a. *55 marbles*

b. *65 marbles*

c. *75 marbles*

d. *80 marbles*

e. *90 marbles*

One could approach the question as follows: Abi has twice as many marbles as Budi. If Abi has 10 marbles, Budi must have only 5. Doni has five times as many marbles as Abi, which means he has 50 marbles. The three combined would therefore have 5+10+50=65marbles, so the correct answer is answer "*b*". This question is straightforward and we should expect primary school teachers to be able to answer this question correctly. In reality, only 52 percent of teachers chose the correct answer.[35]

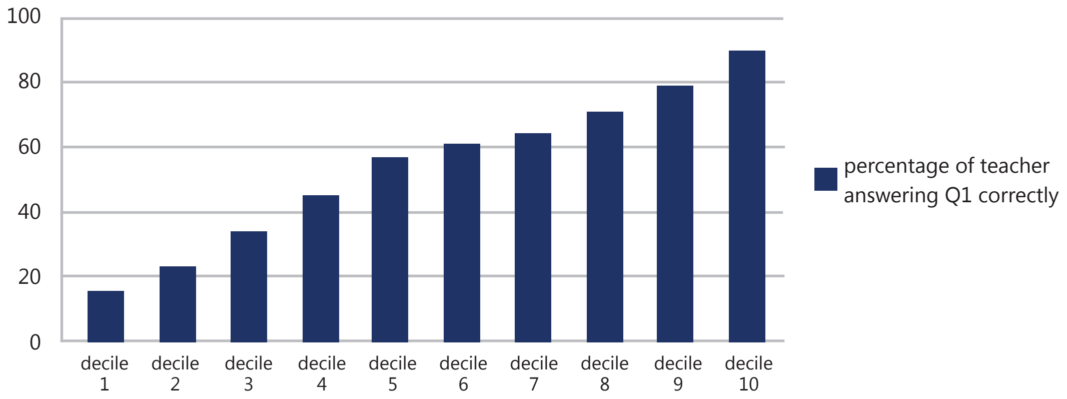
[33] The causal effect we find is somewhat smaller and less precisely estimated than the earlier findings (columns (2)-(7) of Table 3), which is as expected.

[34] The 0.15 of a standard deviation effect on a year-to-year basis, through compounding across years, becomes 0.3 of a standard deviation after the full 6-year primary school cycle.

[35] Another 24 percent said the right answer is answer d. Much smaller percentages (less than 10 percent each) chose options b., c., and e.

Figure 16 investigates how the performance on **Question 1** relates to the overall score on the teacher test. We divided the group of 1,700 test takers into ten equal-sized groups based on their overall raw scores, i.e., deciles. Teachers in the 10th decile are the top 10 percent performers in the sample, reflective of the 10 percent most knowledgeable primary teachers in the country at large. These teachers had no problems with this question and around 90 percent of this group answered **Question 1** correctly. At the lower end of the spectrum, the bottom 30 percent or the bottom three deciles, we observe that teachers had great difficulty answering this particularly easy question correctly. Note that if a group answered **Question 1** by guessing randomly across answer a., b., c., d., and e., about 20 percent would guess correctly.

Figure 16: Scores on test Question 1, by overall score decile



This analysis shows that there is a large group of primary teachers that cannot answer some of the most elementary mathematical questions. Of the bottom 30 percent of teachers, the first three deciles, 80 percent chose the wrong answer to this question, equivalent to random guessing. This bottom 30 percent represents a group of 400,000 Indonesian primary teachers in the population at large.

The second question we discuss is considerably more difficult.

Question 2: *How many multiples of 8 and 12 are there between 1,000 and 2,000?*

a. *40*

b. *41*

c. *42*

d. *43*

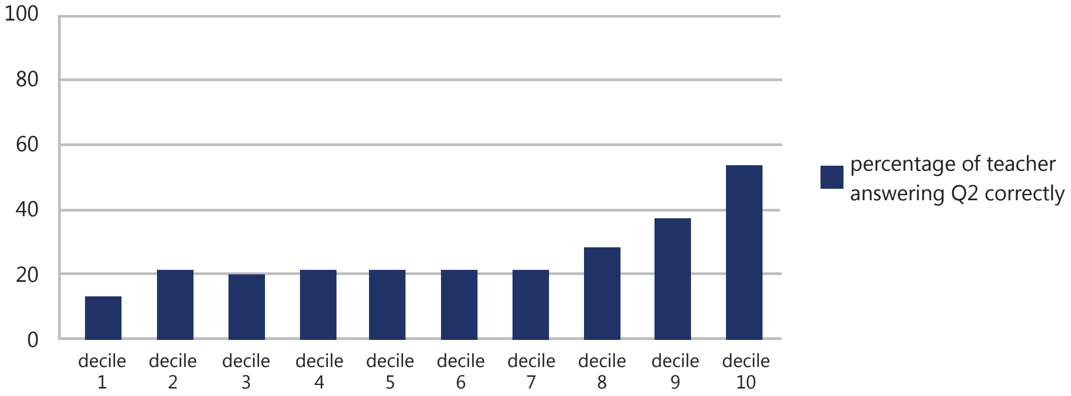
e. *44*

There is no doubt that this question demands a much better command of the mathematical tool box and we would not expect all teachers to answer this question correctly. The correct answer is "*c*". One way to reach the solution is provided in a footnote.[36] Figure 17 shows how our teachers fared on this question. Not surprisingly perhaps most teachers did not know how to approach this question. Of the 7 bottom deciles, we find that 20 percent answered the question correctly, the expected score if teachers do not know the right answer and guess randomly.

[36] One may reach the right answer to this question as follows. The smallest multiple of 8 and 12 is 24 (8 × 3 = 24 and 12 × 2 = 24). We can therefore rephrase the question by saying how many multiples of 24 are there between 1,000 and 2,000? Because $\frac{2000}{24} = 83.333$, there are 83 multiples of 24 between 0 and 2000. Because $\frac{1000}{24} = 41.666$ there are 41 multiples of 24 between 0 and 1,000. This means that there are 83-41=42 multiples of 24 between 1,000 and 2,000. The correct answer therefore is "c".

What is interesting here is that there is a significant group of primary school teachers in Indonesia that did know how to approach the question and reach the right conclusion. Among the top 10 percent of most knowledgeable primary teachers in the country, 50 percent chose the right answer. Even in the lower deciles—the 8th and the 9th deciles—we observe teachers who know what to do when faced with this difficult mathematical problem. In absolute numbers, we estimate that 1 in every 20 primary teachers (or about 75,000 primary teachers in total) knows how to approach **Question 2** and provide the right answer. This is an encouraging result. It shows that Indonesia is able to produce quality teachers and retain them in schools.

Figure 17: Scores on test Question 2, by overall score decile



In conclusion then, this analysis provides insight into the level of subject-matter knowledge among Indonesian primary teachers. We find that many teachers have major difficulties with even the most basic mathematical problems. We observe this more generally for other questions used in the test. It is doubtful whether some of these underperforming teachers can ever be effective teachers, at least in the subject of mathematics. Their lack of knowledge is probably confusing for students and probably discouraging in the longer term. Bright students who see their teachers making mistakes may not be confident enough to correct them. In Indonesia, cultural barriers may also prevent children from doing so.

At the same time, we find that there are exceptionally knowledgeable teachers in Indonesia. This is not a large group in percentage terms, but this group is sizeable in absolute numbers. On average, every two or three primary schools would have one of these high-performing teachers. These top performers might be used as catalysts to provide assistance to some of the lower-performing teachers to help them in their process of professional advancement.

So, how much improvement in the teacher force do we need? We have argued that an achievable first goal might be a standard deviation gain in teachers' subject-matter knowledge. Recall that we predict that this gain would lead to an increase in student-learning outcomes equivalent to roughly 20 points on PISA, a substantial first step. But what does a standard deviation increase in teachers' subject-matter knowledge actually mean in practical terms. Would it mean, for example, that all teachers perform well on **Question 1**?

Figure 18 and Figure 19 present the predicted performance on **Questions 1** and **2**, following a 1.0 standard deviation increase in teachers' subject-matter knowledge overall. We expect all teachers to do better—the poor performers, as well as those at the higher end of the spectrum. A 1.0 standard deviation increase in teachers' subject-matter knowledge means that 72 percent of teachers would answer **Question 1** correctly, versus 52 percent today. Note that for real change to occur it is not at all necessary for all teachers to answer **Question 1** correctly, as the projection still allows for some teachers at the very bottom of the scale to have problems with this question. For example, we would expect about half of the bottom three deciles to know the right answer to **Question 1**, as opposed to only 20 percent today. We still do not need most teachers to improve markedly on **Question 2**, the much harder question.

Figure 18: Predicted performance on test Question 1 after a one standard deviation increase in subject-matter knowledge, by overall score decile

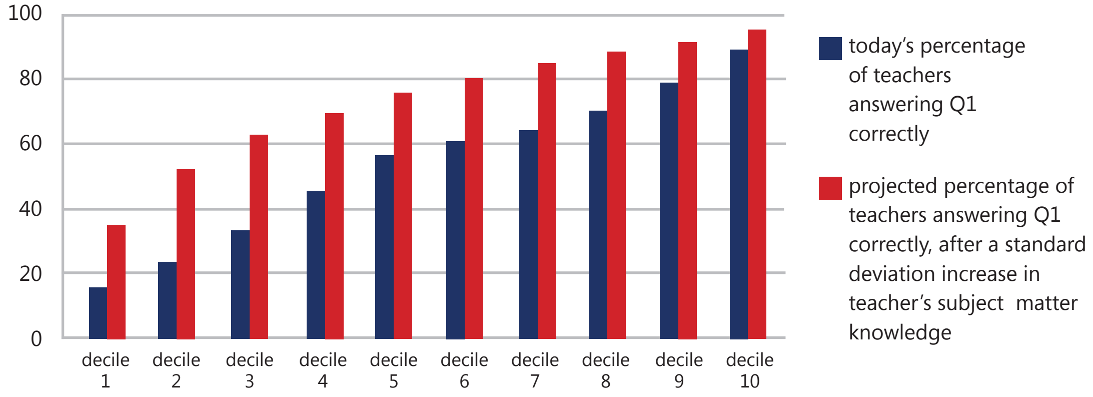
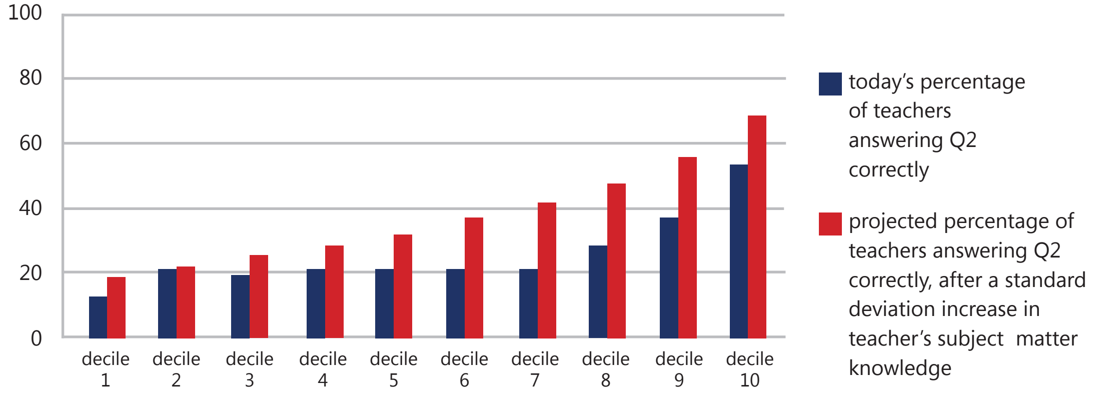


Figure 19: Predicted performance on test Question 2 after a standard deviation increase in subject-matter knowledge, by overall score decile

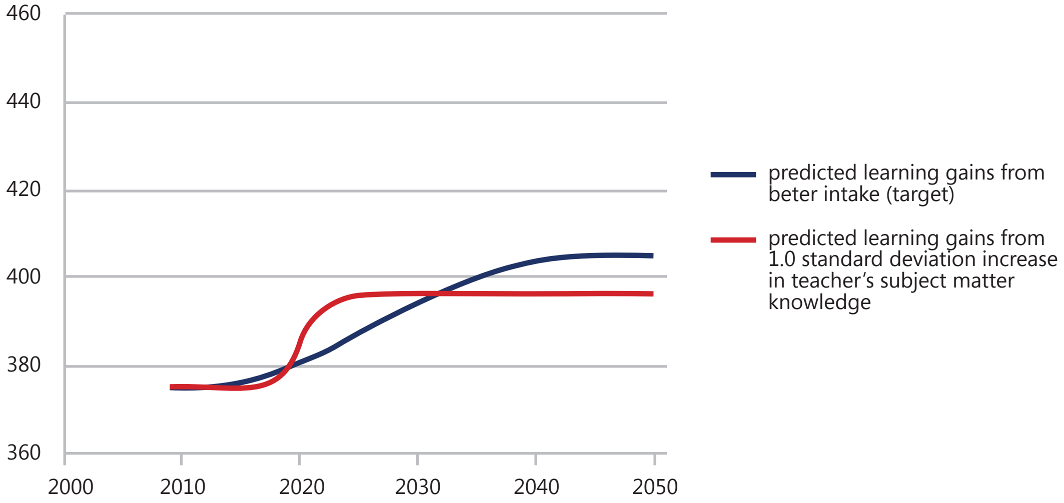


There is no doubt that a boost in subject-matter knowledge of this kind would require hard work from all teachers, and targeted support from local and national governments. But it is reasonable to expect that such work would quickly reap measurable returns in average student achievement. A 1.0 standard deviation increase in the subject-matter knowledge of teachers is an entirely realistic target and, if it is implemented universally, it is also likely to translate far more rapidly into improvements in student-learning outcomes than the slow changes brought about by the replacement of retiring cohorts that we have seen in Section 2.3.

Figure 20 forecasts what would happen to student-learning outcomes if teachers' subject-matter knowledge were to increase by a 1.0 standard deviation between 2015 and 2019. As mentioned before, a 1.0 standard deviation increase in teachers' subject-matter knowledge means that student achievement scores would increase by the equivalent of roughly 20 PISA points within the next eight years. As a reference, Figure 20 also includes the *target for replacement* as presented in Figure 13. The initial target for better intake—if average teachers retire, they are replaced by teachers from the top 30 percent—reaches a higher level, but only in the much longer term.

The gains in student learning through a 1.0 standard deviation increase in teachers' subject-matter knowledge are not enough to catch up with countries such as Thailand and Malaysia, but they are the kind of gains in mathematics achievement that are unparalleled in Indonesia since it first participated in PISA in 2000.[37] Also, these projected gains are about double what we can expect if all teachers upgraded to the bachelor's degree level (Section 2.2). And even after a 1.0 standard deviation increase in teachers' subject-matter knowledge there seems room for more improvement still. A subsequent goal, after making the successful first step, could be to target a 2.0 standard deviation increase in teacher subject-matter knowledge. This is projected to lead to a 0.6 of a standard deviation increase in aggregate student learning over time, equivalent to about 40 points on PISA. Indonesia's new administration could choose to declare the achievement of such a gain in teachers' subject-matter knowledge one of its top priorities. Rigorous subject-matter assessments could be built in at any stage of a teacher's in-service career progression. If the new administration were able to implement this carefully, and ultimately successfully, it is entirely plausible that Indonesia could start making its first strides into systematic and universal improvements in quality education.

Figure 20: Efforts to increase subject-matter knowledge of Indonesian teachers could rapidly translate into improved student outcomes



[37] Between 2003 and 2006, Indonesia gained 30 points on PISA mathematics. The difference is not statistically significant however and in 2009, the scores dropped back to lower levels. These kinds of fluctuations are probably due to general sampling variation in the data. Indeed, Indonesia has not gained statistically significantly in mathematics on PISA since 2000. PISA reading scores, however, have improved statistically significantly.

CHAPTER 4

Policy options for sizable and lasting changes in education quality

For the past 10 years, Indonesia's economic growth rates have been consistently moderate to high. But the country's quality of education, as measured by international assessment studies such as PISA and TIMSS, has generally failed to follow. This is an early warning sign of a looming middle-income trap. If Indonesia wants to keep growing economically, it needs to ramp up labor productivity through improvements in education, among other things. In a growth scenario, companies require new and better skills to operate profitably in an increasingly globalizing and competitive world, especially given deeper ASEAN integration with the implementation of the ASEAN Economic Community this year. If the education system is not capable of providing more productive workers, growth rates will eventually level off, or Indonesia will have to import skilled labor to fill the gap. In improving the quality of its education system, as with many other structural changes, Indonesia is now at a point where it needs to start addressing these issues of change directly.

The empirical evaluation presented in this report is sobering. Despite its success in increasing school enrollment figures in Indonesia, the Government has found the challenge of improving the quality of education far more difficult. We recognize, however, the formidable political and operational challenges that Indonesia faces in making real and effective changes to the system. This chapter offers some policy options based on the results of this empirical study to help facilitate Indonesia's journey along the road towards better education. At the same time, the results of the analysis may also help governments and policymakers elsewhere, who face similar scenarios and challenges, in their own journeys to a successful transition in education system revival.

Besides some of the sobering conclusions of this report, there is also some important good news. The analysis generally indicates that past reforms have contributed to a situation in which progress towards higher-quality education is now a realistic goal. This assertion does not mean that real change will come automatically, however—in fact, far from it. Recent reforms brought millions more children into schools and increased teacher salaries via the certification program. Both have helped to lay the groundwork—an indispensable one—for a successful future. Teaching is once again a popular profession. The position is fully secure, inclusive of a pension plan, and with minimal chances of being dismissed. Salaries, now inclusive of certification allowances, are generally sufficient or even generous (depending on where a teacher lives), and certainly higher than most other jobs available to those with a university bachelor's degree.

The remainder of this report will discuss policy options that we believe could lead to sizable and lasting improvements in the quality of education in Indonesia. The policy options are broken down in two sets: those referring to *pre-service* teacher training and teacher hiring; and those referring to *in-service* teacher management and competency upgrading.

4.1. Pre-service teacher training and teacher hiring

The career prospects of recent graduates from teacher training colleges are bleak (at least for those who want to become teachers). There are two reasons for this. First, recent government statistics show that each year around 250,000 university-trained teacher candidates enter the labor market. Given that only 50,000 teachers retire each year (a number that will increase to roughly 100,000 each year over the next decade) the labor market for teachers is saturated. Clearly, in a saturated market it is hard to find jobs. Second, the NUPTK teacher census suggests that over 60 percent of teachers below the age of 30 in the system are contract teachers without much job security and on low salaries. Those graduates who do eventually find a job, therefore, often spend years of uncertainty over whether their positions will become more permanent.

The number of students entering teacher training colleges can be controlled. Today, there are about 1 million students enrolled in teacher training colleges across Indonesia, studying for jobs that probably most of them will never do. This is an inefficient use of resources: it creates a mismatch between skills that are produced and skills that are demanded by the labor market. But something else might be happening as well. One of the reasons for certification and the associated pay increase was to reestablish the esteem and the attractiveness of the teaching profession. It was thought that higher salaries would draw a higher-caliber high-school graduate into the profession. Although this logic makes sense, a clogged-up labor market could lead to the opposite effect. High-caliber high-school graduates are different from ordinary, or average, high-school graduates in that they have more options in pursuing a career. As high-school candidates internalize the poor career prospects in teaching, high-caliber candidates are the ones that opt out first. This is not because they do not want to teach, but because they are uncertain about whether they will find secure and reasonably well-paid jobs.[38] The renewed attractiveness of the teaching profession, therefore, could lead to the opposite outcome of what was intended, namely, discouraging rather than encouraging high-caliber high-school graduates from pursuing a career in teaching. Restricting the number of applicants entering teacher training colleges seems critical if this trend is to be curbed.

In limiting the number of students entering teacher training colleges, only the very best candidates should be selected out of the total pool of applicants. In some of the high-performing education systems, such as those in Finland and Singapore for example, a stringent selection mechanism is applied to applicants of teacher training colleges (McKinsey & Company, 2007). These countries rely on the attractiveness of the teaching profession—the demand for vacancies in teacher training colleges exceeds the supply—to select the very best among the pool of applicants, and only just enough to replace retiring workers. In a way, Indonesia enjoys the same luxury as Finland and Singapore, because the popularity of the teaching profession is high among high-school graduates. Indonesia, however, makes little use of this situation and does not rigorously control entry into teacher training colleges.

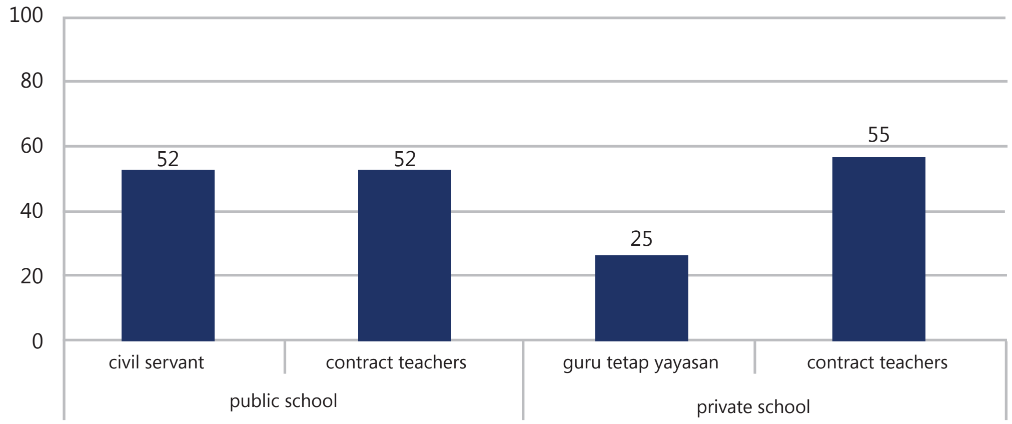
The process of selecting the best candidates out of the total pool of applicants is complex. Ultimately, individual teacher training colleges should play an important role in a system of selection. However, part of the implementation can be organized at a higher level. The results presented in Section 3.3 suggest that it makes sense to make a first selection on the basis of literacy and numeracy skills. McKinsey & Company (2007) mentions that besides high overall literacy and numeracy skills, "strong interpersonal and communication skills, a willingness to learn, and a motivation to teach" are important selection criteria. These latter competencies are generally less precisely defined, which makes their use in selection more subjective and complex. Individual, accredited teacher training colleges should be entrusted and supported in executing this second and final selection. The selection based on literacy and numeracy skills can be organized at a higher level, a consortium of teacher training colleges or universities for example, or even at the level of local or national governments.

[38] Part of the issue here is that teacher hiring systems do not ensure that those with the best qualifications are hired.

The quality of teacher training institutions should be strictly ensured. The local and national governments are responsible for ensuring the quality of the teacher training colleges. In this case, quality does not only mean the quality of the "hardware", in the availability of buildings, laboratories, libraries, etc., but also includes the "software": do teachers in teacher training colleges have sufficient competencies to train new teachers. Today, a common practice is that teachers in teacher training colleges are directly recruited from the best graduates of the very same teacher training colleges. In such situations, the teacher trainers themselves have no real firsthand experience of teaching in real schools. The system of licensing and accreditation currently in place should be further developed and improved. It should consider the need for new teachers from schools (by region or catchment area) and should ensure a minimum level of quality. The development of a more efficient system of licensing is a continuous process of fine-tuning, and more analytical work is warranted to better understand the ways in which teachers' productivity in Indonesian classrooms can be improved.

If the hiring process of teachers into schools were to become more transparent and merit-based, quality will increase. McKinsey & Company (2007) argue that the best school systems regulate entry into teacher training, as opposed to regulating the hiring of graduates from teacher training colleges by schools. Regulating entry into teacher training colleges is economically more efficient. But, given the looming oversupply in the labor market for teachers, Indonesia needs to regulate both. Hiring procedures in Indonesia are not always efficient and/or based on merit. Broadly speaking, it is not clear that systems are in place to guarantee that the best candidates are hired as teachers. Figure 21 highlights some of the problems that occur in practice. Fifty-one percent of the recently hired primary school teachers, that is to say those aged between 24 and 30 years old, do not have a university bachelor's degree. This empirical fact is hard to reconcile with a system that requires a bachelor's degree by law, and with a system of teacher training that produces more than enough qualified candidates. This mismatch suggests that the teacher hiring process is not functioning well. This trend, in which many young contract teachers without appropriate training or qualifications are landing jobs in schools, can be curbed by better regulating the hiring process.

Figure 21: Percentage of primary school teachers (aged between 24-30 years of age) without a university bachelor's degree



Note: Author's calculations based on the 2012 NUPTK teacher census. In public schools teachers the figure discriminates civil servants, with a permanent contract, and contract teachers with a temporary contract and a low salary. In private schools, we categorize separately between the *guru tetap yayasan*, the counterpart of the civil service teachers in private schools, as well as the contract teachers. Overall, across all categories, 51% of primary teachers in this age range have no bachelor's degrees.

4.2. In-service teacher continuous professional management and teacher quality upgrading

To achieve sizable gains in students' academic achievement *in the short term*, competency upgrading of in-service teachers seems to be the only available option available to the Government. The research presented in this report shows that it is far from easy to implement policies that generate large and sustained increases in student-learning outcomes. The research also shows that, to achieve such gains in the short term, the only available option within the sphere of influence of the Government is upgrading competency levels of in-service teachers. Pre-service teacher training and hiring is important, even crucial for the long-term success of the education system, but due to the slow process of retirement and replacement, the impact of improvements in the pre-service are only measurable in the medium to longer term. Indonesia does not appear have the luxury of being able to wait this long.

A system of teacher management and continuous professional development should value demonstrated competencies first place, and bachelor's degrees and seniority second. The research presented in this report emphasizes the need for continuous professional development (CPD) and in-service competency upgrading. A Teacher Professional Management System (TPMS) was established by the Ministry of Administrative and Bureaucratic Reform (MenPAN) in 2009, and was further developed by the Government's Board for Education and Human Resources Development (BADAN) in 2013. Due to the recent implementation of the 2013 curriculum, the development and implementation of the TPMS has been delayed. However, Indonesia's new administration is encouraged to continue the development and implementation of this framework.

The TPMS emphasizes the interplay between competency assessment, performance appraisal and continuous professional development. Competency assessments and performance appraisals identify gaps in a teacher's knowledge and skills. The information obtained from assessments and appraisals should subsequently feed into a system of planning, training and continuous professional development. The sequence is subsequently repeated. By having new assessments and appraisals after a period of in-service training, teachers and other stakeholders are held accountable for the progress they make. An important change with respect to earlier systems is that the TPMS more explicitly emphasizes "learning" and "improving" knowledge and teaching skills.

The TPMS may function best when teachers and other stakeholders are financially rewarded for meeting higher competency standards. As people tend to respond to incentives, the TPMS may work best if steps in teachers' career progressions, or professional certification, are made contingent on meeting well-defined targets and goals. In this way, teachers bear some responsibility for the system's quality upgrade, as their monetary interests are aligned with the broader interest of society, i.e., better teachers in Indonesian classrooms. The challenge is to set targets and goals based on demonstrated competencies, rather than based on more easily observed proxies, such as seniority, bachelor's degrees, or participation in working groups. Targets in demonstrated competencies should be based on **valid** indicators that can be **reliably** obtained. An assessment is **valid** if it measures a competency that is important (ideally scientifically proven to be so) for better teaching. An assessment is **reliable** if it measures that competency with reasonable levels of precision. Developing valid and reliable tools of assessment and appraisal, and implementing them effectively in a TPMS, is a formidable challenge that cannot be resolved overnight. The new Indonesian administration is encouraged to create a task-force to address this issue.[39]

[39] In today's system of career progression, teachers accumulate credit points, which are based on seniority, participation in teacher working groups, peer-assessments, etc. When sufficient credits are accumulated, teachers advance to the next level, which also means an increase in salary. In theory, this system rewards quality at each step, by assuming that teachers become better teachers as they participate in working groups, and by assuming that peer-assessments prevent poor performers from advancing. The current system, however, does not check how much teachers learn by participation in working groups, or whether peer-assessors really rate underperformers with a low score. It is unclear, therefore, how well today's system really incentivizes high performance.

High-stakes assessments and appraisals do not necessarily have to be holistic to be useful. Different tools of assessment and modes of performance appraisal can be used for different purposes. In theory, classroom assessments are holistic, capturing the whole of the teacher-quality concept. However, the reliability of such assessments is difficult to guarantee. This is true in general, but particularly in a country as large and diverse as Indonesia. Appraisals based on classroom observations are inherently subjective, as opinions about what constitutes "good teaching" may differ from one assessor to the next. This inherent subjectivity decreases the reliability of assessments based on classroom observations. Another, perhaps more prominent problem, is that assessments based on classroom observations are potentially subject to manipulation. This issue plays an important role once appraisals based on classroom observations become high-stakes, for example by being decisive for career progression. Subject-matter tests or pencil-and-paper based tests on the theoretical principles of pedagogy, for example, are far less holistic. They are therefore less valid as a way for measuring dimensions of the broader concept of teacher-quality,[40] but the upside is that such pencil-and-paper-based tests are highly reliable, at least when they are well designed. For the purpose of high-stakes assessments, therefore, the Government might consider making trade-offs between validity and reliability, where less valid (but highly reliable) instruments are preferred over more valid instruments, for which reliability cannot be guaranteed. At the current stage of the implementation of the TPMS, the Government may need to exercise caution when appraisals, based on classroom observations, are used as the deciding factor in a teacher's career advancement or professional certification. It is important, however, that performance appraisals based on classroom observations do not lose their function in the TPMS as a low-stakes tool to identify gaps in a teacher's competencies.

The teacher competency assessment tool, the Ujian Kompetensi Guru (UKG), could be further developed into a deciding factor for teachers' career advancement and for professional certification. In 2012, all certified teachers had to participate in a competency test. The assessment became known by its acronym, UKG. The overall performance on the UKG was said to be poor in the sense that average raw scores were low. Low raw scores are not everything, however, as the test may have been too difficult or partially invalid. The UKG nonetheless is a step in the right direction and a start in setting up a system in which competencies of teachers are measured and assessed. The UKG can be strengthened and expanded to capture the entire population of teachers and used as a high-stakes component in the process of selecting teachers for career advancement and for professional certification. Teachers with insufficient passing grades would not be allowed to advance. This will provide them and other stakeholders with incentives to take the TPMS seriously and embark on a process of professional development to improve competencies. For this to work, however, the validity and reliability of the UKG has to be guaranteed, and broadly supported by all stakeholders, including teachers. This would include that passing grades are set at reasonable values, and that the results on different episodes of the UKG are comparable.

[40] De Ree (forthcoming) finds support for the hypothesis that subject-matter knowledge is causally related to better student learning outcomes. This result makes that a reliable subject matter test, is a valid assessment tool.

Annex A: The survey data

To analyze the effects of teacher certification in Indonesia, The Government of Indonesia in partnership with the World Bank, and financially supported by the Government of The Netherlands (via the Dutch Education Support Program), has collected a large and extensive database of student and teacher test score data, supplemented with teacher interview data. This section describes briefly the sampling and data collection.

Sampling of districts and schools

Impact evaluation studies are often implemented on a pilot scale. Therefore, they are often challenged on their external validity. Do the findings from the pilot maintain when they are applied to the population at large? The impacts found in a randomized trial are typically context dependent and what is found in one situation, country, or district, cannot be always extrapolated to other situations, countries or districts. With regard to the external validity of our findings we aimed to achieve some degree of representativeness in the data we collected (De Ree, Muralidharan, Pradhan, & Rogers, forthcoming).

The sampling of 360 schools happened in two stages. In the first stage, 20 districts were sampled. In the second stage, 12 public primary and 6 public junior secondary schools were sampled from each district. The sampling frame was constructed from the 2005/06 NUPTK teacher census. The census consisted of roughly 1,600,000 teachers at the time, in 130,000 public primary and 17,000 junior secondary schools in 454 districts.

A number of selections have been made on this population, so that our sample is not representative of all teachers and schools in Indonesia. The first selection was the exclusion of 5 districts that were considered too dangerous to visit, 20 other districts were excluded because they were sampled for a study into the effects of teacher working groups, and 46 districts were excluded because they were considered "small" (less than ten junior secondary schools or less than forty primary schools). 383 districts ultimately were part of the sampling frame. These 383 districts represent 91 percent of the total number of teachers and schools in Indonesia.

The 20 districts were sampled from 10 sampling blocks or strata, proportional to the size of the district in terms of the number of teachers. The number of districts we sample from each stratum depends (positively) on the number of teachers employed in the area.[41] [42] After sampling, two of the selected districts, Bengkulu Utara and Maluku Tenggara Barat, split up. These divisions do not affect the analysis. But because of this, 22 districts ultimately were part of the analysis. In the text we refer to the original 20 districts, as they were defined at the time the sample was taken.

[41] The 10 strata were the eastern part of Indonesia 1 (1 district sampled), the eastern part of Indonesia 2 (1 districtsampled), Java 1 (3 districts sampled), Java 2 (3), Java 3 (4), Kalimantan (1), Sulawesi (2), Sumatra 1 (2), Sumatra 2 (1) and Sumatra 3 (2)

[42] The districts that were sampled are. Eastern part of Indonesia 1: Maluku Tenggara Barat. Eastern part of Indonesia 2: Lombok Timur. Java 1: Ciamis, Jakarta Timur, Purwakarta. Java 2: Bantul, Kudus, Semarang. Java 3: Lamongan, Lumajang, Probolinggo, Tuban. Kalimantan: Hulu Sungai Selatan. Sulawesi: Gowa. Sumatra 1: Deli Serdang, Tapanuli Tengah. Sumatra 2: Tebo. Sumatra 3: Bengkulu Utara, Ogan Ilir.

In each of the original 20 districts, 12 public primary and 6 public junior secondary schools were sampled. Within districts a number of additional selections on the sampling frame were applied. Because the intervention associated with the randomized controlled trial (see De Ree, Muralidharan, Pradhan, & Rogers (forthcoming) and Section 2.1 of this report) grants preferential access to the certification process for all eligible (but not yet certified) teachers in treatment schools we gain statistical power by oversampling schools that have more teachers that are eligible for certification, but are not yet certified. In fact, schools with relatively small numbers of eligible but not certified teachers were excluded from the sampling frame. The NUPTK census has information on age and education status of all teachers. This information is used to make this selection. Teachers with bachelor's degrees, or who were older than 50 years of age, but were not certified in 2006 were coded as eligible but not yet certified.[43] Primary schools with fewer than three teachers who are eligible, but not certified, and junior secondary schools with fewer than four teachers who are eligible but not certified, were excluded. Very large schools were also excluded: primary schools with more than 20 and junior secondary schools with more than 40 eligible but not certified teachers were also excluded from the sampling frame.

The remaining primary schools (junior secondary schools) were evenly divided into seven (three) school-level strata. The strata's are determined on the basis of the number of eligible but not certified teachers within each of the 20 sampled districts. Stratum 1 would have the smallest number of eligible but uncertified teachers within each district, whereas stratum 7 (or 3 for junior secondary schools) would have the highest number of eligible but not certified teachers. Sampling from these strata ensures that we have a reasonable representation of small and of large schools.

For budgetary reasons we further excluded strata 1, 6 and 7 for primary schools and strata 3 for junior secondary schools from the sampling frame. From each of the remaining 6 strata, we randomly select 3 schools. In each district, we thereby sample 12 primary and 6 junior secondary schools.

Randomization into treatment and control

Within each of the respective school-level strata, one school was randomly selected into the treatment group. The other two schools automatically became control. All eligible but not yet certified teachers in the 120 treatment schools (80 primary and 40 junior secondary) were granted preferential access to the certification process. In the 240 control schools this did not happen. The selections described above imply that the sample is not representative for all public primary and public junior secondary schools in Indonesia. In fact, within the selected twenty districts, the sampling frame represents approximately 39 percent of the number of public primary schools teachers, 45 percent of the public junior secondary schools teachers, 35 percent of the public primary schools and 46 percent of the junior secondary schools. Because of the earlier exclusion of a number of districts, the sample would be representative for a smaller number of schools or teachers in the country at large. We could say that the data we collect are roughly representative for 40 percent of the public schools and teachers in the country. Do notice, however, that the geographic representation is over 90 percent.

Instruments: tests and interviews

In the 360 sample schools we test all core subject teachers and all students from grade 1 until grade 9.[44] The test components of this study were developed by the Indonesian Government's center for educational assessment (Puspendik). This Government unit is for example also responsible for developing and implementing the national school exams. In addition, all core subject teachers and headmasters were interviewed.

[43] This coding is not entirely accurate, but good enough for restricting the sampling frame. In fact, a teacher is eligible with a university bachelor's degree, or with civil service rank IV or seniority (over 50 years of age and over 20 years of teaching experience).

[44] The first round of data collection, the baseline, happened in November 2009. November is five months after the start of the school year and it was decided that each grade g would be tested with the instruments of grade $g - 1$. This implied that grade 1 (primary) and grade 7 (junior secondary) students were not tested at baseline. This has implications for the possibility of tracking students when they progress through school.

Primary schools

1. Teacher tests (*N* ≈ 1,700). All class teachers in the sample schools were given a multiple choice subject matter test.[45] Primary school teachers usually teach all subjects that are part of the curriculum. The test had a mathematics, science, and Indonesian language component. In addition to these subject matter components, also their pedagogical and social competencies were broadly assessed based on a set of multiple choice questions.

2. Teacher interview (*N* ≈ 1,700). All class teachers were interviewed. Teachers were questioned about their demographics, opinions, information on schooling, certification status and income. An important element of the teacher survey asks about the specific classes each teacher is teaching. This information is used to link students to teachers in the analysis.

3. Student test (*N* ≈ 40,000). All students in primary school, from grade 1 to grade 6, were given a grade-specific multiple choice subject matter test. The subject-matter tests had mathematics, science and Indonesian language components. At the back of the answer sheet the students had to fill out a set of additional questions on household assets and parent education levels. This information can be used for example to construct student specific asset indices.

Headmaster interview + facility questionnaire (*N*= 240). The headmaster (or a replacement) of each primary school was interviewed about the state of the school, the number of students and teachers, finances, etc.

Junior secondary schools

1. Teacher test (*N* ≈ 1,400). All core subject teachers in the sample schools were given a multiple choice test. Core subject teachers are mathematics, physics, biology, Indonesian language, and English teachers. These teachers were given a subject specific multiple choice subject matter test, i.e., mathematics teachers did a mathematics test. If a teacher would teach multiple core subjects, he or she would be tested on all these subjects. Similarly to primary school teachers, pedagogical and social competencies were also broadly assessed, based on a set of multiple choice questions.

2. Teacher interview (*N* ≈ 1,400). All core subject teachers were interviewed using the same instrument as used for the primary school teachers.

3. Student test (*N* ≈ 40,000). All students in junior secondary school, from grade 7 to grade 9, were given a grade-specific multiple choice subject-matter test. The subject-matter tests had mathematics, science, Indonesian language and English language components. The set of questions at the back of the answer sheet is the same as for primary school students.

4. Headmaster interview + facility questionnaire (*N*= 120). The headmaster (or a replacement) of each junior secondary school was interviewed at each round. These questionnaires are the same as for primary schools.

The same data was collected three times. The first round of data was collected in November 2009. We refer to this round of data collection as the baseline. Tests and survey data were collected at different points in time for the baseline. This meant that the overlap between tests and surveys was not perfect: some teachers who were tested were not interviewed and vice versa. The second round of data was collected in April/May 2011 right before the end of the 2010/11 school year. Each grade was tested with a grade appropriate subject matter test. At midline we administered the interview and the test at the same visit, such that we have almost complete overlap between teachers who were tested and who were surveyed. A final round of data was collected in April/May 2012 right before the end of the 2011/12 school year.

[45] In a small number of schools it happened that mathematics for example was taught by a specially assigned mathematics teacher, rather than a class teacher. This teacher, then, would have been tested as well. Even when such teachers teach only a single subject, they were still tested with the general primary school teacher test.

Annex B: Teacher value-added: estimating the distribution of school and teacher quality across Indonesian primary classrooms

Theorists and empiricists in the field of education often evaluate the importance of teachers in terms of how much they contribute to a child's (scholastic) achievement. The term that is widely used in this context is a teacher's "value added", where teachers add (more or less) to the achievement of students. The statistical models used to confront this theoretical idea to real-world data on student test scores are called *value added models*. Value added models are derived from explicit theories of skill formation, where each input from birth until today contributes to current achievement (Todd & Wolpin, 2003) . Value added models come in various forms. The version that we rely on heavily in this report is the following:[46]

$$y^*_{it} = \alpha + inputs_{it} + \gamma\, y^*_{it\text{-}1} \qquad (1)$$

Where $y^*_{it}\,\forall\tau = t,t\text{-}1$ are normalized measures of true achievement, with mean 0 and standard deviation 1. The value added model links last year's true achievement scores $y^*_{it\text{-}1}$ to current true achievement scores $y_{it}{}^*$and $inputs_{it}$ happening in-between. $inputs_{it}$ in equation (1) may be all factors that contribute to learning in a given year, from school inputs (e.g., teachers), to parental support, to individual innate abilities, interests, motivations and talents. The model describes that the reasons for observing differences in achievement scores between students at a given point in time, is due to either differences in inputs between period t - 1 and t (some may receive more or better inputs than others) or to differences in prior achievement (which is due to the accumulation of inputs from birth until period t - 1).

This Annex discusses the estimation of the variation in test score gains that is explained at the level of the classroom. In other words we are interested in the parameters of the regression of $inputs_{it}$ on a full set of classroom dummy variables:

$$inputs_{it} = \sum \delta_c 1\ (i \in c) + u_{it} \qquad (2)$$

Where $1(i \in c) = 1$ if child i is in classroom c, and 0 otherwise. The δ_c's are the parameters on the classroom dummy variables $1(i \in c)$. The regression (2) defines the quantities of interest. We cannot run the regression in practice, simply because the totality of $inputs_{it}$ are not observed. The parameters of interest, the δ_c's, therefore are estimated from the value added specification (1). Incorporating equation (2) into the value added specification (1) yields:

$$y^*_{it} = \alpha + \sum \delta_c 1\ (i \in c) + \gamma\, y^*_{it\text{-}1} + u_{it} \qquad (3)$$

[46] See for example Guarino, Reckase, & Wooldridge (2012), for a discussion of some different versions of the value added model.

The parameters of interest in this Annex are the δ_c's, the classroom effects, and γ, the time persistence parameter. Below we describe how we estimate these quantities, based on test score information we obtained from our data base.

Two major obstacles need to be overcome before we can estimate the parameters of (3). Both issues are well established, and the literature has yielded a number of ways of dealing with them (see for example, Andrabi, Das, Khwaja, & Zajonc (2011)). The first obstacle is the likely persistence in unobserved inputs. The second obstacle is the measurement error in test scores.

Persistence in unobserved inputs

It is likely that the u_{it} term in the model is persistent over time. For example, children differ in terms of their levels of innate ability or in terms of the support they receive at home. Smarter or better supported children "learn more" in a given year, but also have higher prior test scores. This phenomenon makes that (even if we had access to true achievement scores y^*_{it} an ordinary least squares regression on (3) would not yield consistent estimates of the parameters of interest. We assume that the persistence in the error term uit is appropriately represented by an individual student fixed effect α_i and an idiosyncratic term ε_{it} which is assumed uncorrelated over time:

$$y^*_{it} = \alpha + \sum \delta_c 1\,(i \in c) + \gamma y^*_{it-1} + [\alpha_i + \varepsilon_{it}] \tag{4}$$

In what follows we use a Blundell and Bond type estimator to estimate the parameters of (4) (Blundell & Bond, 1998). Blundell and Bond (1998) argue that in a stationary dynamic process, like one presented by (4), the first differenced lagged score $\Delta y^*_{it-1} = y^*_{it-1} - y^*_{it-2}$ is uncorrelated with the composite error term $\alpha_i + \varepsilon_{it}$ and correlated with the lagged score y^*_{it-1}. In the analysis we therefore use Δy^*_{it-1} (or a proxy thereof, see below) as an instrumental variable for y^*_{it-1} in a two-stage least squares (2SLS) setting.

Measurement error in test scores

The second obstacle is that of measurement error in test scores. The y^*_{it} 's in the model, are so-called *true achievement scores*. "True" refers to the idea that they are measured precisely, i.e., without error. In real life, observed test scores, for example those obtained from our survey, measure "true" achievement levels with some level of imprecision, or noise. Differences between observed scores and the theoretical true scores are due to good/bad luck with guessing when the test-taker is not fully sure about an answer, being more or less tired on the day of the test, etc. The fact that test scores tend to not completely reflect a student's level of true achievement is important in empirical value added modeling.

In our analysis we assume that measurement error in test scores is "classical", in the sense that it is additive, zero in expectation, and independent of the true score. Moreover, to maintain the interpretation of the parameters in terms of standardized true score gains, we need to maintain the normalization we have introduced before $V(y^*_{i\tau}) = 1$. Consequently, we can write down the relationship between the observed scores $y_{i\tau}$ and the true scores $y^*_{i\tau}$:

$$\frac{y_{i\tau}}{\sqrt{\rho_{y_\tau}}} = y^*_{i\tau} + e_{i\tau}\ \forall \tau = t, t\text{-}1 \tag{5}$$

Where $y_{i\tau}$ is a standardized observed test score with $E\,[\,y_{i\tau}\,]=0$ and $V(\,y_{i\tau})=1$, and ρ_{y_τ} is the coefficient of reliability, defined as $\frac{V(\,y^*_{i\tau})}{V(\,y^*_{i\tau}+e_{i\tau})}$. The premultiplication with $\frac{1}{\sqrt{\rho_{y_\tau}}}$ ensures that $V\,(\,y^*_{i\tau}) = 1$ is maintained.[47]

[47] $V\left(\frac{y_{i\tau}}{\sqrt{\rho_{y_\tau}}}\right) = \frac{1}{\rho_{y_\tau}} = \frac{1}{\frac{V(\,y^*_{i\tau})}{V(\,y^*_{i\tau} + e_{i\tau})}} = \frac{V(\,y^*_{i\tau} + e_{i\tau})}{V(\,y^*_{i\tau})} = 1 + V(\,e_{i\tau})$

Incorporating the noisy observed test scores in the original specification yields:

$$\frac{y_{it}}{\sqrt{\rho_{y_t}}} = \alpha + \sum \delta_c 1\ (i \in c) + \gamma \left(\frac{y_{it\text{-}1}}{\sqrt{\rho_{y_{t\text{-}1}}}} \right) + [\alpha_i + \varepsilon_{it} + e_{it} - \gamma e_{it\text{-}1}] \qquad (6)$$

In the empirical operationalization the standardized observed test scores are scaled with (the square root of) estimates of the reliability coefficients. The literature on value added generally ignores this transformation of the data.

The main issue however with measurement error is that $e_{it\text{-}1}$ and the period t - 1 observed score $y_{it\text{-}1}$ is correlated by definition, unless $V(e_{it\text{-}1}) = 0$. In the analysis we address this problem in two steps. First, the scaled full lagged achievement score $\frac{y_{it\text{-}1}}{\sqrt{\rho_{y_{t\text{-}1}}}}$ is replaced with a scaled lagged achievement score, which is based only on the odd numbered items $\frac{y^o_{it\text{-}1}}{\sqrt{\rho_{y^o_{t\text{-}1}}}}$.

$$\frac{y_{it}}{\sqrt{\rho_{y_t}}} = \alpha + \sum \delta_c 1\ (i \in c) + \gamma \left(\frac{y^o_{it\text{-}1}}{\sqrt{\rho_{y^o_{t\text{-}1}}}} \right) + [\alpha_i + \varepsilon_{it} + e_{it} - \gamma e^o_{it\text{-}1}] \qquad (7)$$

Subsequently, following Blundell & Bond (1998) we use the lagged change in the observed even item score

$\Delta \frac{y^E_{it\text{-}1}}{\sqrt{\rho_{y^E_{t\text{-}1}}}} = \frac{y^E_{it\text{-}1}}{\sqrt{\rho_{y^E_{t\text{-}1}}}} - \frac{y^E_{it\text{-}2}}{\sqrt{\rho_{y^E_{t\text{-}2}}}}$ as an instrumental variable for the lagged odd item score $\frac{y^o_{it\text{-}1}}{\sqrt{\rho_{y^o_{t\text{-}1}}}}$.

Note that in many ways this approach is close to an approach in which biased parameters are corrected, post-estimation, using estimates of reliability coefficients like Cronbach's alpha, or the even-odd split-half reliability ratio. The approach presented here however is easier to implement as it does not involve a post estimation adjustments.

In the end our approach heavily relies on the assumption that $\Delta \frac{y^E_{it\text{-}1}}{\sqrt{\rho_{y^E_{t\text{-}1}}}}$ is uncorrelated with the individual effect α_i , following Blundell and Bond, and is uncorrelated with the noise term in the lagged odd item test score $e^o_{it\text{-}1}$ equation (7). The approach is restrictive, but there seems no easy solution to further relaxing these assumptions with only three rounds of student test score data. See Andrabi, Das, Khwaja, & Zajonc (2011) who face a similar issue. They provide a discussion on why having four consecutive rounds of testing data is useful in the proper estimation of value added models.

Model (7) is estimated with two-stage least squares, with $\Delta \frac{y^E_{it\text{-}1}}{\sqrt{\rho_{y^E_{t\text{-}1}}}}$ as the excluded instrument, and a full set of classroom dummy variables. The main quantity of interest in this section is $SD(\delta_c)$, the standard deviation of the classroom effects. The regression obtains a full set of estimates $\hat{\delta}_c$. The literature acknowledges the fact that $V(\delta_c) < V(\hat{\delta}_c)$, where $V(\cdot)$ is the variance across classrooms. We correct our estimate of the standard error of the classroom effects like Aaronson, Barrow, & Sander (2007) and Jacob & Lefgren (2008), by subtracting the average squared standard error associated with the $\hat{\delta}_c$'s, from V($\hat{\delta}_c$).

The first two columns of Table 1 presents estimates of the standard deviation of the classroom effects, corrected (as discussed in the previous paragraph) and uncorrected, and separately by subject and pooled across subjects.

Table 1: Estimates of the standard deviation of classroom effects across Indonesian public primary schools

	Variation in value added across Indonesian primary school classrooms			
	standard deviation of estimated classroom effects (corrected)	standard deviation of estimated classroom effects (uncorrected)	standard deviation of estimated classroom effects (corrected). Controlled with student assets indices	standard deviation of estimated classroom effects (uncorrected). Controlled with student assets indices
Pooled across subjects	0.36	0.43	0.35	0.43
Mathematics	**0.53**	**0.66**	**0.52**	**0.65**
Science	0.45	0.59	0.45	0.59
Indonesian	**0.13**	**0.46**	**0.12**	**0.45**

We find sizable differences in learning across classrooms. Some classrooms do much better on a year-to-year basis than others. We also find that these differences are biggest in mathematics and science, and less pronounced for Indonesian language. The substantial differences in the speed of learning across Indonesian primary school classrooms suggest that some teachers are much better than others. We cannot however draw such a firm conclusion based on the analysis so far. It is conceivable for example that more able or students who are better supported at home (for example, the wealthier students) tend to end up in the better schools with better teachers. If such classes do well, it might not only because teachers are better, but also because the students have more potential or more support at home. In column 3-4 of Table 1 we present the result of a replication of the analysis, where we have included a set of 5 dummy variables of a student's socio economic background in the regression. The inclusion of these asset indices affects the estimates only marginally.[48] Indeed, if sorting of better supported or more able students to better teachers would be important, we would expect that controlling for a student's background would drastically reduce the estimated standard deviation of the estimated classroom effects.

The robustness analysis suggests that endogenous sorting is not a major threat to the approach to estimating the variation in classroom value added, at least not with the Indonesian data. It seems that some classrooms do much better than others because of what happens inside that classroom. One caveat needs to be mentioned however with this otherwise important empirical result. In the approach we cannot distinguish statistically between teacher level and school level inputs. Part of the reason for this is due to the difficulty of definition. How would we classify for example a situation in which if a principal hires better teachers, or is able to guide teachers better into better performance? It is indeed due to the better teachers that the classes are doing better, but on the other hand, without the principal, these better quality teachers would not have even been there to start with.

With these estimates we could gauge how much of the variation in learning outcomes is due to teachers and schools, and how much is due to influences of parents, the broader community and individual talents and interests—the amount that is not due to teachers and schools. We cannot be precise on this percentage because we do not measure (and do not attempt to measure) how persistent the classroom effects are. If the classroom effects are fully persistent, i.e.,

[48] The estimated parameters on the asset dummies are jointly statistically significantly different from zero, and the differences between the poor and the better off are quite sizable. See Annex C for additional results.

students have the same quality teachers and schools throughout their entire career in basic education, the year-to-year classroom effects are compounded. If the standard deviation of the year-to-year classroom value-added effect is 0.36, it is 0.72 in compounded terms—about twice as much. Because the standard deviation in student-learning outcomes is 1.0 by construction, schools and teachers would account for 72 percent of it. The remainder, 28 percent, would be due to a mixture of the influence of parents, individual talents and interests, etc.

But typically, the classroom effects would not be fully persistent. If instead the classroom value-added effects are independent over time, i.e., the quality of this year's teacher is uncorrelated with the quality of last year's teacher, the classroom value-added effects are not compounded across years. The standard deviation of the variation in learning outcomes explained by schools and teachers is then only 36 percent, such that the remainder of that, 64 percent, is due to influences of parents, individual talents and interests. Roughly, we could therefore say that the influence of schools is anywhere between 36 and 72 percent, and the influence of the residual forces is between 64 and 28 percent. In the main text these numbers are rounded: anywhere between 1/3 and 2/3 of the variation (measured in standard deviation terms) in learning outcomes is due to schools and teachers.

Annex C: More empirical results of value-added models

This Annex builds further on the value added regressions introduced in Annex B of this report. In Annex B we estimate the variation in learning across Indonesian classrooms. The Annex concludes that there are substantial differences in learning between Indonesian primary school classrooms, and that these differences are greater for math and science education than for languages. The natural follow-up question is then: which observable characteristics of these classrooms matter? In this Annex we refer to the general value added specification (1) of Annex B. We are interested in the parameters of the following regression:

$$inputs_{it} = X_{it}\beta + u_{it} \tag{8}$$

Again, we cannot estimate this model directly, as the totality of *inputs*$_{it}$ are not observed. Equation (8) defines the parameters of interest however. Incorporating this into the value added specification (1) yields

$$y_{it} = X_{it}\beta + \gamma y_{it\text{-}1} + u_{it} \tag{9}$$

In estimating the parameters of (9) we face some of the same obstacles as discussed in Annex B, that is, the potential correlation between u_{it} and $y^{*}_{it\text{-}1}$ due to time persistence in the unobserved input factors and measurement error in observed test scores. In this Annex we deal with these problems in the same way as in Annex B.

The variables included in X_{it} are described in Table 2. The table shows that most teachers in the sample have a university bachelor's degree, for example. This number is higher than the national average across all grade levels and all schools in Indonesia.

Table 2: Averages of some of the variables used in the analysis

	Primary (grade levels 4, 5, 6)	Junior secondary (grade levels 7, 8)
Teacher has a bachelor's degree (0/1)	0.71	0.93
Standardized teacher test score	-0.00	-0.00
Teacher's age	45.20	43.80
Teacher is certified (0/1)	0.63	0.74
Teacher is civil servant (0/1)	0.91	0.92
Class size	29.93	32.55
Asset dummy 1 (0/1) (poorest)	0.18	0.21
Asset dummy 2 (0/1) (2nd poorest)	0.19	0.22
Asset dummy 3 (0/1)	0.24	0.23
Asset dummy 4 (0/1)	0.22	0.17
Asset dummy 5 (0/1) (wealthiest)	0.18	0.17
Asset information is missing (0/1)	0.00	0.00

Note. Endline data (collected April 2012) are used to describe some of the characteristics of the input variables used in the analysis.

Table 3 presents regression results of versions of regression equation (9). As discussed in Annex B, we need at least three rounds of student test score data to estimate the persistence parameter γ of equation (9). Because for our junior secondary school data we do not have three rounds of consecutive testing data for the same child we set the parameter γ to a fixed value in estimation. We use the estimate obtained from the primary school models (column (5)) in Table 3 as our best estimate.

Table 3: Results from empirical value added modeling

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	primary 1	primary 2	primary 3	primary 4	primary 5	primary 6	primary 7	primary 8	junior sec. 1	junior sec. 2	junior sec. 3	junior sec. 4	junior sec. 5	junior sec. 6
Scaled lagged score (odd numbered items)	0.611***	0.606***	0.558***	0.549***	0.536***	0.498***	0.477***	0.290***	parameter on the lagged score is set at 0.536					
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.08)						
Teacher has a bachelor's degree		0.100**	0.032	0.025	0.015	0.060*	0.035		0.230**	0.074	0.118	0.089	-0.012	-0.047
		(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)		(0.10)	(0.08)	(0.08)	(0.07)	(0.04)	(0.04)
Teacher's subject matter test score (normalized)			0.240***	0.270***	0.268***	0.221***	0.214***	0.152*		0.180***	0.169***	0.148***	0.034*	0.037*
			(0.05)	(0.06)	(0.05)	(0.06)	(0.07)	(0.09)		(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
Teacher's age				0.011***	0.011***	0.011***	0.007**				0.011***	0.008**	-0.004	-0.005**
				(0.00)	(0.00)	(0.00)	(0.00)				(0.00)	(0.00)	(0.00)	(0.00)
Teacher is certified				0.011	0.001	0.018	0.003				0.043	0.033	0.051	0.019
				(0.05)	(0.05)	(0.05)	(0.06)				(0.06)	(0.06)	(0.04)	(0.03)
Teacher is a civil servant				-0.057	-0.068	-0.102	-0.095				-0.005	0.006	0.033	0.011
				(0.07)	(0.07)	(0.07)	(0.08)				(0.07)	(0.06)	(0.06)	(0.06)
Class size				0.000	-0.001	0.002	-0.000				0.002	0.002	0.007**	0.003
				(0.00)	(0.00)	(0.00)	(0.00)				(0.00)	(0.00)	(0.00)	(0.00)
Asset dummy 2 (2nd poorest)					0.081***	0.077***	0.060***					0.092***	0.014	0.012
					(0.03)	(0.03)	(0.02)					(0.02)	(0.02)	(0.01)
Asset dummy 3					0.127***	0.129***	0.096***					0.153***	0.059***	0.040***
					(0.03)	(0.03)	(0.02)					(0.03)	(0.02)	(0.02)
Asset dummy 4					0.180***	0.193***	0.132***					0.206***	0.069***	0.030*
					(0.04)	(0.03)	(0.03)					(0.04)	(0.03)	(0.02)
Asset dummy 5 (wealthiest)					0.245***	0.265***	0.171***					0.377***	0.162***	0.062***
					(0.05)	(0.04)	(0.03)					(0.05)	(0.04)	(0.02)
Asset dummy is missing					0.011	-0.025	-0.014					-0.094	-0.074	-0.084
					(0.18)	(0.16)	(0.15)					(0.15)	(0.14)	(0.14)
Grade level - Subject fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District fixed effects						Yes							Yes	
School fixed effects							Yes							Yes
Student fixed effects								Yes						
Number of observations	38779	38779	38779	38779	38779	38779	38779	28245	85725	85725	85725	85725	85725	85725
Long run effect of a teacher with a bachelor's degree		0.253	0.0720	0.0543	0.0326	0.119	0.0676		0.496	0.160	0.255	0.193	-0.0254	-0.101
Long run effect of a standard deviation increase in teacher's subject matter knowledge			0.542	0.599	0.578	0.440	0.410	0.327		0.388	0.364	0.318	0.0725	0.0799

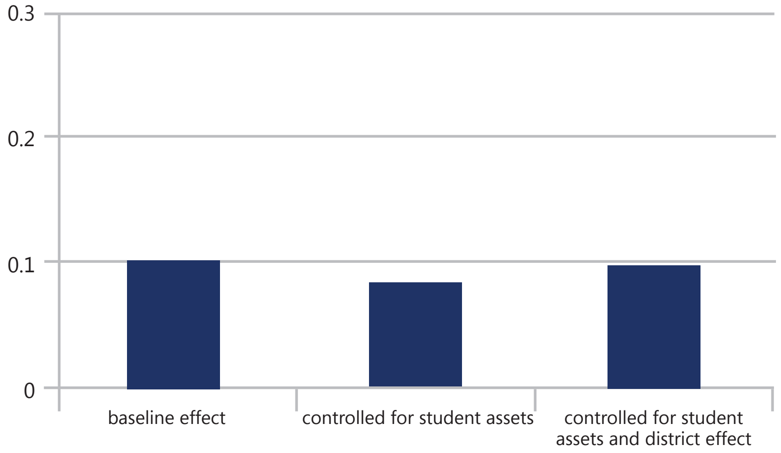
NOTE: significance levels *** 1%, ** 5%, * 1%. Clustered (school-level) standard errors are in parenthesis. Control group for asset categories is the poorest category (asset dummy 1). Long run effects are calculated by dividing the year-to-year effect by 1 minus the estimated parameter on the lagged score

Table 3 shows that estimates are broadly similar between primary and secondary schools. However, there are some differences. The results for primary schools are somewhat more robust to the inclusion of "fixed effects" at various levels.

We find that a university bachelor's degree is statistically significant (column (2) for primary and column (9) for junior secondary). Especially for junior secondary school teachers, this parameter is sizable. Note however, that 93 percent of junior secondary teachers in our sample has a university bachelor's degree. Junior secondary teachers without bachelor's degrees seem to do poorly. For primary schools, these effects are not as pronounced.

While these results suggest that teachers with a bachelor's degree are better teachers, some caution should be kept in mind when drawing this conclusion. Selection issues, where children from better-off households are drawn to the best schools, possessing better trained teachers, might be important. We do indeed find that students with wealthier parents are somewhat more likely to have teachers with bachelor's degrees (results not reported). The inequality of opportunity between rich and poor, however, does not affect our conclusion. If we take the socioeconomic status of students into account, we still find practically the same positive association between the academic qualification of teachers and student learning. Also, if we control for 20 district fixed effects we practically obtain the same results. Figure 22 presents the effect sizes. The first bar presents the difference in learning gains between students with teachers with a bachelor's degree versus those with teachers with lower level qualifications. The second and third bars present the results after controlling for student asset levels, and district effects, respectively. The results regarding the importance of the university bachelor's degree are hardly affected by these control strategies, while the asset index and the district effects themselves are both highly statistically significant.

Figure 22: Effect of having a teacher with a bachelor's degree, on learning outcomes



Note. These are estimates of effect sizes measuring the year-to-year impact of having a teacher with a university bachelor's degree, versus having a teacher with lower level qualifications. The estimates are based on data from primary schools only. The baseline specification controls for a student's prior test scores and a full set of grade level-subject dummy variables (see Annexes B and C). The first bar, the baseline effect, is the effect size reported in column (2) of Table 3 (Annex C). The second and third bar show the estimated effects of having a teacher with a bachelor's degree after controlling for student assets, and for student assets and district effects respectively.

The statistical significance of the bachelor's degree effect however disappears when teacher's subject-matter test scores are taken into account in column (3) and (10) of Table 3. This indicates that direct proxies of ability, such as subject-matter test scores, are better markers of teacher quality than more indirect proxies like a university bachelor's degree. Subject-matter knowledge of teachers is clearly highly correlated with student learning. Teacher's subject-matter test scores are noisy measures of true subject-matter ability, just as student test scores are. In the analysis we therefore have used lagged (midline) teacher's test scores as instrumental variables for current (endline) teacher's test scores.

We also find that a teacher's age is positively correlated with student learning (column (4) for primary and column (11) for junior secondary). It appears that older teachers in the sample are a better than the younger ones. It might be that experience simply makes teachers teach better (an "age" effect). But, it may also be that the newly hired teachers are somehow of lower quality (a "cohort" effect). The data does not easily allow us to discriminate between the two.

Whether teachers are civil servants, whether teachers are certified, or the size of the class, does not correlate with student learning. Student background characteristics on the other hand matter a lot. Students from wealthier backgrounds fare better at school than their classmates from poorer families. Those in the wealthiest group gain 0.25 (for primary) to 0.38 (for junior secondary) of a standard deviation more than those in the poorest group, on a year-to-year basis (see column (5) and column (12) primary and junior secondary, respectively).

Column (6) (primary) and column (13) (junior secondary) presents district-fixed effects models. This means that we are zooming in on the variation in learning outcomes across students, but within the 20 selected districts. For primary schools we find hardly any differences between column (5) and (6) results. For junior secondary schools on the other hand we find that the parameter on teacher's subject-matter knowledge becomes much smaller and the parameter on teacher's age becomes statistically insignificant. Column (7) (primary) and column (14) (junior secondary) presents school fixed effects models. Still the primary school results are unaffected, whereas the junior secondary school results are again less robust.

Overall, we find that teachers with higher levels of subject-matter knowledge do better, and for the primary level we can even observe this phenomenon within schools. The school-fixed effects approach in column (7) and (14) controls for example, for differences in the school budget, the quality of the principal in managing his teachers, etc. The column (7) results, therefore, is a strong indication that knowledge matters. For junior secondary schools we do not find this.

Column (8) presents results of a student-fixed effects model. Whereas column (7), the school-fixed effects model, looked at differences in student learning within schools, across students and subjects, column (8) looks at differences across different subjects for the same students. The approach effectively investigates whether students progress faster in math than in science if their teacher knows more math than science. The results still suggest that teachers' subject-matter knowledge is important. The results however are somewhat weaker, in the sense that the estimate is only statistically significant at the 10 percent level. The approach, including more detailed results, is presented in De Ree (forthcoming).

Finally, Table 3 presents "long run effect" parameters. These are calculated by taking the year-to-year effect parameter β, divided by $1 - \gamma$. In column (2) for example, the long-run effect of a teacher's qualifications is $\frac{\beta}{1-\gamma} = \frac{0.1}{1 - 0.606} = 0.253$. This means that if you have teachers with bachelor's degrees for many years in a row, you would end up at 0.253 of a standard deviation higher in the distribution of test scores than students with teachers without bachelor's degrees for many years. The rationale behind this formula is the following. Primary teachers with a bachelor's degree for example produce 0.1 of a standard deviation of learning value each year relative to teachers without this degree. After two years with a university trained teacher (versus two years with untrained teachers) the difference has increased to $0.1 + 0.606 \times 0.1 \approx 1.606$ standard deviations, a value that depends on the persistence parameter γ here estimated at 0.606. After three years the difference is $0.1 + 0.606 \times 0.1 + 0.606^2 \times 0.1 \approx 0.197$. After many years, say six years, the complete primary cycle the difference has become $0.1 + 0.606 \times 0.1 + 0.606^2 \times 0.1 + 0.606^3 \times 0.1 + 0.606^4 \times 0.1 + 0.606^5 \times 0.1 \approx 0.241$
It can be shown that this sequence is roughly equal to $\frac{0.1}{1 - 0.606}$, which is the value that is reported in Table 3. The long run effect presented in column (2) of Table 3 therefore present roughly the expected difference in scores between students who had teachers with a bachelor's degree for the entire 6-year primary cycle, versus students who had teachers with lower level qualifications for the same period.

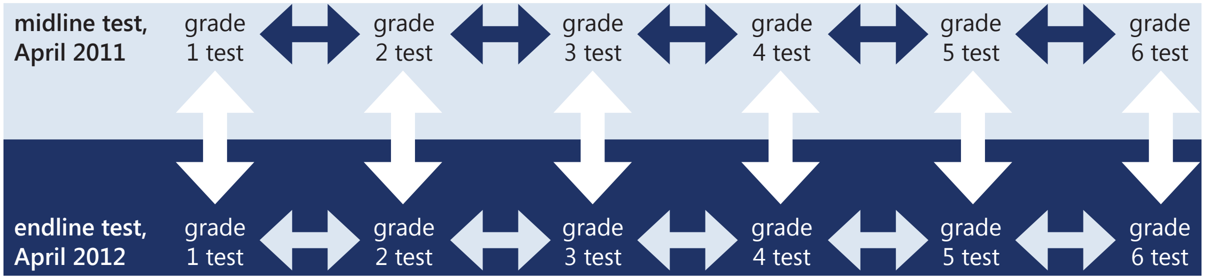
Annex D: Learning profiles, how much children learn in a year

It is of considerable interest to measure what and/or how much students learn each year when they are in school. For example, we would want to know how knowledge levels of 4th graders compare to that of 3rd graders. In our matched student-to-teacher database, we cannot simply compare raw test scores obtained from these two populations and meaningfully compare them. The reason for this is that 3rd graders are given a different test than 4th graders. A grade 4 test is typically more difficult than a grade 3 test. One reason for why tests for 3rd graders are different than tests for 4th graders is that if two populations have different levels of true achievement you might need a very long test (with many questions) to span the entire range of achievement of both populations, while being able to measure achievement at acceptable levels of precision. But for practical purposes, tests should not be too long as students can have trouble maintaining concentration levels.

The matched student to teacher database also uses different tests for each grade. And, also, tests are changed each new round of the data collection.[49] The raw scores on these different tests therefore cannot be meaningfully compared. Instead, the tests used rely on an anchor item design. This means that *some* different test forms have *some* of the same (overlapping) questions. The testing literature refers usually to anchor items or linked items when they mean overlapping questions.

What this means in practice is that some questions, typically a handful, in "adjacent" tests are the same. Using arrows, Figure 23 shows schematically what we mean with adjacent tests. There are overlapping items between test for the same grade in different years (vertical links) and there are overlapping items between tests for different connecting grade levels in the same year (horizontal links). To be clear about this: children in grade 1 at the midline (April 2011) do a different test than children in grade 2 at the midline (April 2011), only some of the questions they do, are the same.

Figure 23: Schematic representation of the anchor test design: all test forms are linked horizontally and vertically, but not diagonally



[49] Students therefore never see the same question (or item) twice.

A variety of statistical approaches have been proposed to use anchor items to make test scores from different tests comparable. Linear approaches were perhaps more popular in the early days (Angoff, 1984), recently designs based on item response theory seem more common. Both have their own advantages and disadvantages. In this report we chose to stay close to the data and simply report raw (standardized) scores on the anchor items, and compare them between relevant populations. The approach is simple so these results are easy to interpret. Future research may look into the specifics of these linked items and perhaps choose another approach. More interesting, even, would be to evaluate the specifics of each question, and see what it is that children learn, and what it is they do not. An example of such important pioneering work can be found in Pritchett (2013).

For the purpose of this report we want to know how much children learn as they progress through school. We do this, by simply comparing cohorts in the same year. Effectively we are comparing average raw scores on the overlapping items between grade 3 and grade 4. It is consistently found that children of higher grade levels scores higher on average. The question is now, how much higher. We are interested in comparing the average raw score on the anchor item of the grade 4 population, in relation to the mean and the spread of the distribution of raw true scores in grade 3. One key thing to worry about when making such assessments is the reliability of the results on the anchor test. (The anchor test is the totality of the anchor items between two test forms.) As the anchor test consists of only a handful of different items, scores are usually not particularly "reliable", in the sense that the variance of the true score on the anchor test may be quite a bit smaller than the variance of the observed score on the anchor test. Low test reliability is of secondary importance when comparing means in a population (in fact, we could even compare means on a single test item). But it is important for assessing the spread of the distribution of true scores.

Suppose measurement error on an anchor test is classical:

$$y = y^{*} + e$$

Where y is the raw score on the anchor test, y^{*} is the unobserved true score on the anchor test, and e is a measurement error term.

Suppose, for example, that we are interested in comparing the scores on the anchor test between grade 3 and 4. How much better are grade 4 children than grade 3 children on average, measured in terms of standard deviations of grade 3 true scores:

$$DIFF_{3,4} = \frac{M_4(y^{*}) - M_3(y^{*})}{SD_3(y^{*})} \qquad (10)$$

Where $M_4(\cdot)$ is the mean score of grade 4 students, $M_3(\cdot)$ is the mean score of grade 3 students and $SD_3(\cdot)$ is the standard deviation of grade 3 scores.Based on assumptions about the nature of the measurement error (measurement error is classical, uncorrelated with the true scores), we can derive that:

$$M_4(y^{*}) = M_4(y)$$
$$M_3(y^{*}) = M_3(y)$$
$$SD_3(y^{*}) = \sqrt{\rho_3}\, SD_3(y)$$

where the latter condition follows directly from the definition of the coefficient of reliability $\rho_3 = \frac{V_3(y^{*})}{V_3(y)}$. Incorporating this in (10) yields:

$$DIFF_{3,4} = \frac{M_4(y) - M_3(y)}{\sqrt{\rho_3}\, SD_3(y)} \qquad (11)$$

The reliability of the grade 3 test (in this example) therefore plays a role in measuring differences between populations. This is because differences in means (in terms of standard deviations of raw observed scores) appear smaller on less reliable tests. In what follows we estimate the reliability coefficient with Cronbach's alpha.

What do we observe for example between grade 3 and 4, both for the midline (April 2011) and the endline (April 2012). (An extra complication in the anchor design is that we have two test, an *a* and *b* version of each test, to prevent children from cheating.)

Table 4: The difference between 3rd graders and 4th graders, in terms of learning level

base form	alternative form	number of anchor items	raw score base form	raw score alternative form	standardized difference	p-value (testing equal mean scores, between base and alternative form)	Cronbach's alpha, base test	corrected standardized change (standardized difference, divided by the square root of Cronbach's alpha)
midline 3a	midline 4a	5	0.43	0.53	0.42	0.00	0.35	0.71
midline 3a	midline 4b	**5**	**0.43**	0.53	0.42	0.00	0.35	0.72
endline 3a	endline 4a	**3**	**0.46**	0.53	0.26	0.00	0.24	0.54
endline 3a	endline 4b	**3**	**0.46**	0.55	0.30	0.00	0.24	0.62
endline 3b	endline 4a	**3**	**0.45**	0.54	0.29	0.00	0.17	0.69
endline 3b	endline 4b	**3**	**0.45**	0.55	0.33	0.00	0.17	0.79

Note. There are no anchor items between midline test 3 b and the midline 4 tests

Table 4 shows that raw scores on the anchor items increased by roughly 10 percentage points. This comes down to about 0.35 of a standard deviation in the raw observed scores distribution. However, because the anchor test consists (as usual) of only a few items, the test have low reliability scores. Correcting for this, in the last column, suggests that the real difference between grade 3 and grade 4 students on average is about 0.6 to 0.7 of a standard deviation.

In the main text we have produced the average differences between grade *G* and grade *G*+1 , where *G* = 1, 2, 3, 4, 5. We find consistently that differences are around 0.6 of a standard deviation, except when we compare grade 5 to grade 6. The results for the grade 5 to 6 comparison are presented in Table 5 below.

Table 5: The difference between 5th graders and 6th graders, in terms of learning levels

base form	alternative form	number of anchor items	raw score base form	raw score alternative form	standardized difference	p-value (testing equal mean scores, between base and alternative form)	Cronbach's alpha, base test	corrected standardized change (standardized difference, divided by the square root of Cronbach's alpha)
midline 5a	midline 6a	4	0.28	0.39	0.50	0.00	0.10	1.59
midline 5a	midline 6b	**4**	0.28	0.39	0.50	0.00	0.10	1.61
midline 5b	midline 6a	**2**	**0.44**	0.61	0.48	0.00	0.14	1.28
midline 5b	midline 6b	**2**	**0.44**	0.61	0.48	0.00	0.14	1.28
endline 5a	endline 6a	**3**	**0.45**	0.65	0.72	0.00	0.14	1.97
endline 5a	endline 6b	**3**	**0.45**	0.66	0.76	0.00	0.14	2.07
endline 5b	endline 6a	**3**	**0.44**	0.65	0.80	0.00	0.06	3.24
endline 5b	endline 6b	**3**	**0.44**	0.66	0.84	0.00	0.06	3.41

We find larger standardized differences between grade 5 and 6. Also, because reliability of these tests is low, they are corrected by multiplying these changes by a factor of about 3, i.e., $\frac{1}{\sqrt{0.1}} \approx 3$ (see equation (11)). At such low levels of reliability one might be worried about whether the differences are measured precisely enough. We encourage, therefore, further research based on these data. But having said that, the analysis presented here is the first of its kind in Indonesia and it documents learning gains that could be substantial, at least in standard-deviation terms. Future research however might want to investigate what "standard deviation increases" mean in terms of what students know and are able to do. Somewhat substantial gains in standard deviation terms, which we find here, might not actually mean that much. That is, students in grade 6 know more on average than those in grade 5, but the difference between them, in terms of what they know and are actually able to reproduce might not be particularly impressive. Such an assessment however would require a value judgment on what "sufficient" or "impressive" is.

Bibliography

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicage public high schools. *Journal of Labor Economics , 25* (1), 95-135.

Akerlof, G.,& Yellen, J. (1990). The fair wage-effort hypothesis and unemployment. *Quarterly Journal of Economics , 105* (2).

Al-Samarrai, S., & Cerdan-Infantes, P. (2013). Where did all the money go? Financing basic education in Indonesia. In D. Suryadarma, & G. W. Jones (Eds.), *Education in Indonesia.* Singapore: Institute of South East Asia studies.

Andrabi, T., Das, J., Khwaja, A., & Zajonc, T. (2011). Do value added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics* , 29-54.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton N.J.: Education Testing Service.

Banerjee, A., & Duflo, E. (2011). *Poor economics: a radical rethinking of the way to fight global poverty.* PublicAffairs.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* .

Chang, M. C., Shaeffer, S., Al-Samarrai, S., Ragatz, A., De Ree, J., & Stevenson, R. (2013). *Teacher reform in Indonesia: the role of politics and evidence in policymaking* (Vol. Directions in Development). Washington, DC: World Bank.

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

De Ree, J. (forthcoming). How much teachers know and how much it matters in class: Analyzing three rounds of subject-specific test score data of Indonesian students and teachers.

De Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. (forthcoming). Double for Nothing? Experimental Evidence on the Impact of an Unconditional Teacher Salary Increase on Student Performance in Indonesia.

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *The American Economic Review* .

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review, 102* (4), 1241-1278.

Fahmi, M., Maulana, A., & Yusuf, A. A. (2011). Teacher Certification in Indonesia: A Confusion of Means and Ends. *Working Paper in Economics and Development Studies* .

Glewwe, P., Hanushek, E. A., Humpage, S., & Ravina, R. (2011). School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. In P. Glewwe, *Education policy in developing countries.* The University of Chicago Press.

Guarino, C., Reckase, M., & Wooldridge, J. M. (2012). Can value-added measures of teacher performance be trusted? *working paper (unpublished)* .

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher Quality. In E. A. Hanushek, & F. Welsch, *Handbook of the economics of education* (Vol. 2).

Hattie, J. (2013). *Visible learning: A synthetsis of over 800 meta-analyses relating to achievement.*

Indonesia Ministry of Education and Culture. (2009). *Dampak Peningkatan Kesejahteraan Guru Terhadap Mutu Input (Quality Enrollment) dan Pemberian Bantuan Dana Kompetitif terhadap Kemampuan Lulusan LPTK*. Jakarta: Research and Development Board (Balitbang).

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics , 26* (1), 101-136.

Jalal, F., Samani, M., Chang, M. C., Stevenson, R., Ragatz, A. B., & Negara, S. D. (2009). *Teacher certification in Indonesia: A strategy for quality improvement.*

Jiyono. (1985-86). *Research on Teachers' Aptitudes and Instructional Materials in Physical Science at the Primary-School Level.* Jakarta: Indonesia Ministry of Education and Culture (Balitbang).

Kane, T., & Staiger, D. (2008). Estimating teacher impacts of student achievement: an experimental evaluation. *NBER working paper No. 14607* .

Maralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy , 119* (1), 39-77.

McKinsey & Company. (2007). *How the world's best-performing education systems come out on top.*

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

Novia, D. R. (2014, December 2). Pendidikan Indonesia Gawat Darurat. *Republika* , p. 5.

OECD. (2010). *PISA 2009 at a Glance.* OECD Publishing.

Pritchett, L. (2013). *The rebirth of education: schooling ain't learning.* Brookings institution press.

Pritchett, L., & Beatty, A. (2012). The negative consequences of overambitious curricula in developing countries. *HKS Faculty Research Working Paper Series* (RWP12-035).

Pritchett, L., & Filmer, D. (1999). What education production functions really show: a positive theory of education expenditures. *Economics of Education Review , 18* (2), 223-239.

Ragatz, A. (forthcoming). *The importance of teacher knowledge in student learning outcomes (TIMSS video study).*

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica , 73* (2), 417-458.

Santoso, D. (2004). Government expects too much from poverty-line teachers. *Jakarta Post* .

Todd, P., & Wolpin, K. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal , 113* (485), 3-33.

Webbink, D., & Gerritsen, S. (2013). How much do children learn in school? International evidence from school entry rules. *CPB discussion paper* .

Widhiarto, H. (2014, October 18). Amid soaring education budget, performance remains low. *The Jakarta Post, Jakarta* .

World Bank. (forthcoming). *Assessing the role of the SChool Operational Grant Program (BOS) in improving education outcomes in Indonesia.*

World Bank. (2014). *Indonesia: Avoiding the trap.* Development Policy Review.

World Bank. (1989). *Indonesia: Basic education study.* World Bank.

World Bank. (2013). *Spending more or spending better: Improving education financing in Indonesia.* Jakarta: The World Bank Office Jakarta.

World Bank. (forthcoming). *Teacher certification and beyond: an empirical review of the teacher certification program and teacher quality improvement in Indonesia (full report).*

World Bank. (2012). *Teacher certification in Indonesia: a doubling of pay or a way to improve learning?*