Public Disclosure Authorized

Public Disclosure Authorized

WPS7556

Policy Research Working Paper 7556

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

Joppe de Ree



WORLD BANK GROUP
Education Global Practice Group
February 2016

Policy Research Working Paper 7556

Abstract

Improving the quality of education is one of today's main challenges for governments in the developing world. Based on a unique matched student-to-teacher panel data set on test scores this paper presents two empirical results for Indonesia. First, through detailed inspection of teacher-level responses to test questions, the paper concludes that subject matter knowledge of primary school teachers in Indonesia is low on average and that a 1.0, but also a 2.0 standard deviation increase in teachers' subject matter knowledge seem to be achievable medium-term goals for education policy making in Indonesia. Second, the paper presents the results of three types of value-added regressions, a (standard) level specification, a school fixed-effects specification, and a flexible student-teacher fixed-effects specification. The student-teacher fixed-effects approach estimates the parameters of a value-added model using test score variation within each student-teacher pair across three different subjects, mathematics, science and Indonesian language. The results suggest that a 1.0 (and 2.0) standard deviation increase in teachers' subject matter knowledge across-the-board can yield increases in student achievement by 0.25 (and 0.50) student-level standard deviations by the time students complete the six-year primary school cycle.

This paper is a product of the Education Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at joppederee@gmail.com.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

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

Joppe de Ree*†

JEL Classification: I21, I25, H42, O15

Keywords: Education, value-added modeling, teachers, subject matter knowledge, Indonesia

*World Bank, email: joppederee@gmail.com

†The research presented in this paper was generously supported by the Dutch Education Support Program (DESP) funded by the Government of the Kingdom of The Netherlands. Special thanks go to Dedy Junaedi (and his teams of field workers, data managers, etc.), Susiana Iskandar, Titie Hadiyati (and team), Amanda Beatty (and team), Halsey Rogers, Menno Pradhan, Karthik Muralidharan, Andy Ragatz, Samer Al-Samarrai, Ai Li Ang, Husnul Rizal, Laura Wijaya, and other colleagues at the World Bank office in Jakarta, as well as to colleagues at the Indonesian Ministry of Education and Culture, such as Yendri Wirda Burhan, Simon Silisabon (and the team at *puslitjak*), Rahmawati, Yani Sumarno (and the team at *puspendik*).

1 Introduction

Improving the quality of education is one of today's main challenges to policy makers in Indonesia and elsewhere in the developing world. Despite the fact that Indonesia doubled real spending on education in the last decade[1], the quality of instruction, as measured through international student comparisons like PISA and TIMSS, has not (or not much) improved. PISA researchers conclude for example that half of Indonesia's 15-year-olds do not have "a basic level of proficiency, at which students begin to demonstrate the reading literacy competencies that will enable them to participate effectively and productively in life [p50. in OECD (2010)]."[2] This suggests that many teachers and schools fail to prepare the next generation for an ever-more demanding global labor market.

This paper presents two empirical results for Indonesia that potentially also apply to the broader developing world. First, section (2) shows that subject-matter knowledge of primary school teachers in Indonesia is arguably low on average, and most likely a binding constraint to learning in many Indonesian primary schools. Second, section (4) presents evidence that increases in teachers' subject-matter knowledge can help increase student-learning outcomes. The inference is based on three types of value-added regressions: a standard model in levels, a school fixed-effects model, and a flexible student-teacher fixed-effects model. The latter estimates the parameters of a value-added model, only using variation within each student-teacher pair across the three different subjects – mathematics, science and Indonesian language.

The student-teacher fixed-effects model attempts to address important threats to conventional value-added models in levels. The fixed-effects results, for example, would be insensitive to situations in which intrinsically more able students students are matched to better schools or teachers. Moreover, models in levels are incapable of separating out a teacher's subject-matter knowledge from other (unobserved) teacher attributes, such as motivation, general intelligence, etc. The idea of looking at within-student variations across different subjects is not new and has been successfully applied before by Metzler and Woessman (2012).[3] Metzler

[1] See World Bank (2013)

[2] Almost 80% of Indonesian 15-year-olds do not have a baseline level of proficiency in Mathematics, that is below level 2 as coded by OECD (2010).

[3] The idea of using within-student variation (across subject scores, or otherwise) has been used before, see for example Dee (2005), Dee (2007), Ammermüeller and Dolton (2006), Clotfelter, Ladd, and Vigdor (2010),

and Woessman (2012) however base their analysis on a cross-section of data only, while this research estimates the parameters of a fully specified dynamic value-added model.

The data we use are unique for the developing world. Subject-matter test score data of all (around 45,000) students of a near representative sample of 240 primary schools are tracked across three rounds of measurement, November 2009, May 2011 and May 2012. Figure (1) presents a map of the 20 districts that were randomly selected to take part in the study.[4] The student level test score data can be linked to survey and subject-matter test score data of their respective teacher(s). A feature of this data is that both students and teachers are tested on the same three subjects – math, science and Indonesian language – in each of the three rounds of measurement. The data therefore allows for estimating value-added models on variation within each student-teacher pair, across the three subjects. At the same time, the data permits to account for biases due to measurement error in test scores (under standard assumptions on the nature of the measurement error).

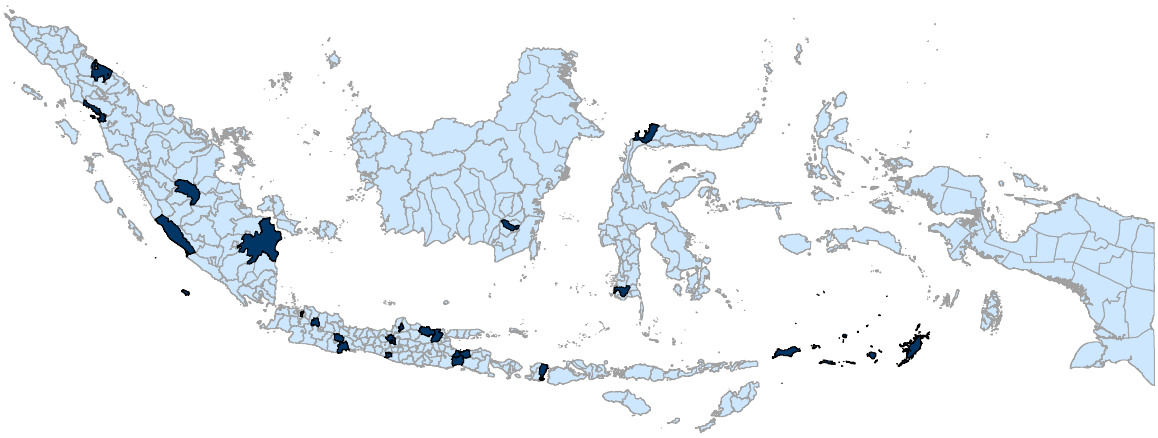


Figure 1: Map of Indonesia, with the 20 selected districts highlighted

Kingdon and Teal (2010), and Chu, Loyalka, Chu, Qu, Shi, and Li (2015).

[4]The data was collected in support of a randomized controlled experiment designed to evaluate the effects of Indonesia's teacher certification program (De Ree, Muralidharan, Pradhan, and Rogers (2015)). The data is nearly geographically representative for the majority public primary schools in Indonesia. (The data on 120 Junior Secondary schools is not used in this paper.) Indonesia spans across 1,700 more or less inhabited islands and is the world's fourth most populous country, and the distance from the most western school in the sample, on mainland Sumatra, to the most eastern, on one of the remote islands of the South Moluccas, roughly spans the distance between San Francisco and New York. See also De Ree, Al-Samarrai, and Iskandar (2012), Chang, Shaeffer, Al-Samarrai, Ragatz, De Ree, and Stevenson (2013) and World Bank (2015), for analysis based on the same data. For more information on the data and the experiment see De Ree, Muralidharan, Pradhan, and Rogers (2015) and World Bank (2015).

The results of the analysis suggest that teachers' subject-matter knowledge is important, and that 1.0 but also 2.0 standard deviation increases in teachers' subject-matter knowledge are realistic medium-term targets for education policy making in Indonesia. A 1.0 and 2.0 standard deviation increase in teachers' subject-matter knowledge across-the-board, is predicted to increase student-learning outcomes by 0.25 and 0.50 student-level standard deviations respectively by the time students complete the six-year primary cycle. These effects are substantial and would amount to 17.5 to 35.0 points respectively on the scale used by PISA if such impacts would persist until age 15.

The paper is organized as follows. Section (2) presents the teacher test score data and provides an intuitive account of the level and spread of subject-matter knowledge among Indonesia's primary school teachers. Section (3) presents the statistical value-added model and presents ways of dealing with the difficulties of estimating the parameters of such models, e.g. these difficulties are due to measurement error in test scores and the time persistence of unobserved inputs. Section (4) presents regression results of three different value-added models: the standard level model, the school fixed-effects model, and a student-teacher fixed-effects model. Section (5) concludes and discusses the findings within the broader context of Indonesia's education system.

2 How much do teachers know?

The level of teachers' subject-matter proficiency in Indonesia is said to be particularly low on average, even though rigorous international comparisons do not exist. Muhammad Nuh, for example, the country's former Minister of Education and Culture reflected on the results of a large scale competency assessment of Indonesian teachers in 2012: ".. on one hand I am happy with the plethora of poor results because it means the data is reliable (The Jakarta Globe 2012)." Low levels of subject-matter proficiency can be a real constraint to effective teaching, even if teachers have a talent for delivering a message to a classroom. This section provides an intuitive account of the level and spread of subject-matter proficiency of Indonesian primary teachers. The data reveals somewhat poor performance overall as the

majority of primary teachers have difficulties with high-school material.

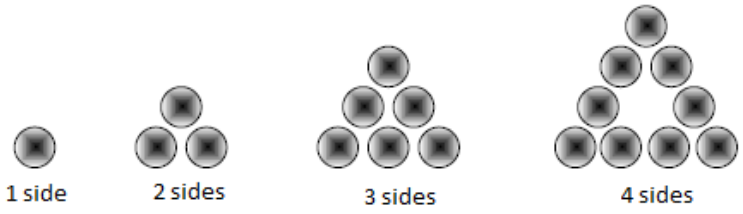
The tests used in this research were developed by the Center for Educational Assessment of the Indonesian Ministry of Education and Culture (*Pusat Penilaian Pendidikan* or *Puspendik*). The tests consist of a 20 item math component, a 20 item science component, and a 20 item Indonesian language component. Another 40 items were used to assess pedagogical knowledge and social and personality traits. Teachers had 2 hours to complete the entire test. This meant that teachers had 1 minute and 12 seconds on average for each question. All questions were 5-option multiple choice questions. Approximately 1,700 primary school teachers in 240 public primary schools were tested.

The test intended to roughly reflect what the Center for Educational Assessment expects teachers to know. Scores of around 60 percent correct or higher, then, are considered "a pass". But what does passing really mean? The Center for Educational Assessment may have been too ambitious for example, by setting targets too high. It is even more difficult to compare such raw scores internationally as standards for *appropriate levels of knowledge* may differ between countries. Comprehensive studies specifically designed to evaluate teachers' subject-matter knowledge across countries, such as PISA or TIMSS for students, do not exist. Hanushek, Piopiunik, and Wiederhold (2014) cleverly measure cross-country variations in teacher knowledge from OECD's Programme for the International Assessment of Adult Competencies (PIAAC). PIAAC does not focus on teachers explicitly, but as the PIAAC samples are large enough, sufficiently many teachers are part of it. The first round of PIAAC however does not survey any of the developing economies.

As cross-country comparisons of teachers' subject knowledge are lacking in the developing world, we rely on introspection of the reader of this paper to make an assessment of whether Indonesian teachers are up to standard. For the analysis we select two items from the 20-item math component of the subject-matter test of the Midline study (fielded in April-May 2011). Midline data of teachers is arguably the most reliable source of information we have, as both baseline and endline had to rely on smaller budgets for the field work. Especially, for midline, we managed in most cases to test teachers in the same classroom as their students, thereby minimizing the risk of collaboration among teachers. The two test items we look at are the

following.[5]

QUESTION 1. Look at the stacks of marbles below!



How many marbles will be needed to build a "triangle" of which one side consists of 8 marbles?

a. 18 marbles

b. 21 marbles

c. 24 marbles

d. 32 marbles

e. 36 marbles

QUESTION 2. An automotive firm uses robots to produce cars. If 3 robots can produce 17 cars in 10 minutes, how many cars can 14 robots produce in 45 minutes assuming that each robot works at the same speed?

a. 325 cars

b. 345 cars

c. 357 cars

d. 353 cars

e. 365 cars

The two questions do not require high levels of mathematical skills and should be doable for high-school graduates. And because a university bachelor's degree is a requirement for primary school teachers in Indonesia they should, by and large, perform well on these questions. The difficulty level of the two questions is roughly representative for the difficulty of the entire test, as the performance (percentage correct) on these two questions on average is about the same as the performance (percentage correct) overall.

[5]World Bank (2015) presents a similar analysis, but looks at two alternative items.

To answer question 1 test-takers must understand how these triangles are constructed. If one side of the triangle has 8 marbles, there are 7 and 6 marbles needed to complete it. The correct answer therefore is $8 + 7 + 6 = 21$, or answer *b*. Item 2 requires more calculation and the numbers used in the problem add to the complexity.[6] There are different ways of solving this problem. One is, by saying that what 14 robots can do in 45 minutes, is $\frac{14\times45}{3\times10} = \frac{14\times15}{1\times10} = \frac{14\times1.5}{1\times1} = 21$ times as much as what 3 robots can do in 10 minutes. Because 3 robots produce 17 cars in 10 minutes, 14 robots can produce $17 \times 21 = 357$ cars in 45 minutes. The correct answer therefore is *c*.

Of our sample, 45% of teachers answered question 1 correctly and 33% answered question 2 correctly, and only 14% of teachers answered both questions correctly.[7] So rather than observing many teachers performing well, only a minority, in the end, did. This finding suggests a low level of subject-matter knowledge among primary school teachers in Indonesia: a sizable share of the teacher population has difficulties with material of the high-school mathematics curriculum. It is quite possible that low levels of teachers' knowledge are a binding constraint for learning in many Indonesian primary classrooms. This is unfortunate for current generations of students of course, but it also presents a direct opportunity for change. At such low levels, improvements in teacher's subject-matter knowledge might translate quickly into improved student-learning outcomes.

For policy makers in Indonesia the question of interest is: how much more knowledge do teachers need to acquire before investments in teacher training start to really pay off? The empirical literature almost universally reports effect sizes in terms of population standard deviations, e.g. Metzler and Woessman (2012) and section (4) of this paper. But population standard deviations are not necessarily intuitive quantities. It is not clear a priori what a standard deviation increase in teachers' subject-matter knowledge means in every-day life. And whether it would be costly or cheap to achieve such changes through government-supported training programs. The remainder of this section provides an intuitive account of whether standard deviations of subject-matter knowledge are *a little* or *a lot*.

[6] One odd feature of the question for example, is that it is unrealistic to expect that robots can produce a car in 10 minutes.

[7] If all teachers would guess randomly on both items and across the five possible answers *a.*, *b.*, *c.*, *d.*, and *e.*, only about $0.20 \times 0.20 \times 100\% = 4\%$ would answer both items correctly.

To make intuitive what a 1.0 or 2.0 standard deviation increase in subject-matter knowledge means in everyday life, we discuss what it would mean with respect to the performance on the two test items introduced before. We model the probability of answering question 1 and 2 correctly at a given true knowledge level x^* as follows:[8]

$$P\left(Q_k = 1|x^*\right) = \Phi\left(\alpha_k + \beta_k x^*\right) \tag{1}$$

for $k = 1, 2$ (question 1 or 2). x^* is a measure of true knowledge, standardized to have $E\left[x^*\right] = 0$ and $V\left(x^*\right) = 1$. Based on this model we can predict how likely an Indonesian primary teacher is to answer question k correctly –or questions of similar content and difficulty– at each knowledge level x^*. And we can also evaluate how much more likely he or she is to answer questions like k correctly, after a 1.0 standard deviation increase in subject-matter knowledge, i.e. by evaluating $D\left(x^*\right) = P\left(Q_k = 1|x^* + SD\left(x^*\right)\right) - P\left(Q_k = 1|x^*\right)$.

Estimating the parameters α_k and β_k of (1) is not straightforward, because true achievement scores x^* are not observed. *True achievement scores* x^* are only proxied with *observed test scores* x with usually a considerable amount of noise, or measurement error. In Appendix A we show how one can estimate the parameters α_k and β_k based on noisy test scores x, assumptions on the nature of the measurement error in x, and estimates of the reliability coefficient ρ. Based on bias-corrected estimates of α_k and β_k and the probit model (1), we present the relationship between the standardized true achievement scores x^* on the horizontal axis, and the predicted probabilities $\hat{P}\left(Q_k = 1|x^*\right)$ of answering question 1 [left panel] and 2 [right panel] correctly.[9]

An average teacher (or more precisely, a teacher who scores the population average $x^* = 0$) is able to solve problems like question 1, 44 percent of the time. Question 2 is considerably more challenging and the typical Indonesian teacher would be able to solve such problems only about 30 percent of the time. The average teacher therefore has some level of mastery

[8]This is a probit model, but a logistic regression model, or something more complex, could also be used for this purpose.

[9]Note that the model does not have a "guessing parameter", such as the 3PL model from item response theory. Even without any knowledge test-takers should guess correctly once every five times (on a five item multiple choice subject-mattertest). However in our data we observe that test-takers not always report answers to each question.

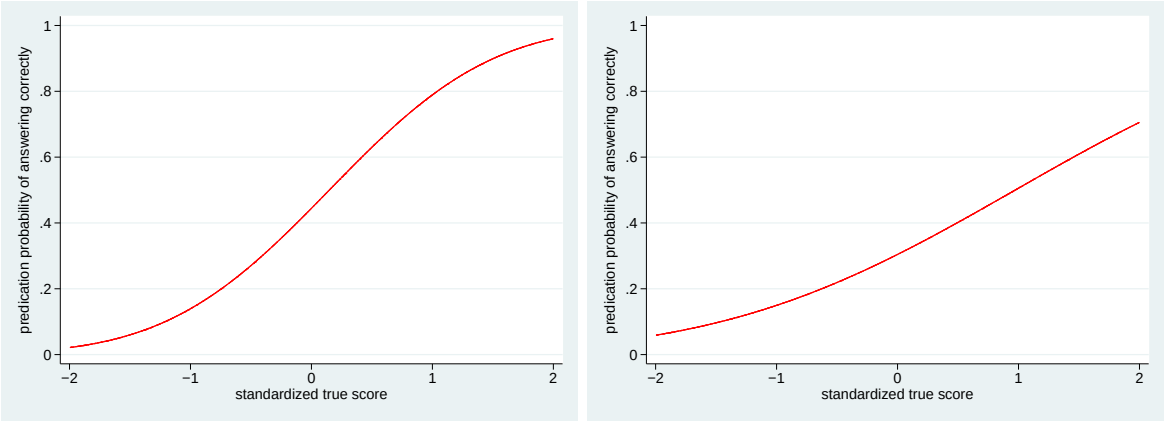


Figure 2: Modeling the relationship between standardized true scores and the probability of answering question 1 (left panel) and 2 (right panel) correctly

over the skills needed to solve question 1, but he or she would not find question 1 easy.

What does a 1.0 standard deviation increase mean? A teacher with a level of subject-matter knowledge of 1.0 standard deviation above average, i.e. $x^* = 1$, has much less problems with the first question and would answer such questions correctly more than 80 percent of the time. He or she generally has the skills needed to solve questions like 1 without difficulty. Question 2 however still remains somewhat of a challenge for a teacher in Indonesia with $x^* = 1$, as he or she would answer questions like these correctly 50 percent of the time.

What does a 2.0 standard deviation increase mean? A teacher with a level of subject-matter knowledge of 2.0 standard deviation above average, i.e. $x^* = 2$, has no problems at all with the first question and would almost always answer questions like question 1 correctly. He or she has all the skills needed to solve problems as straightforward as question 1. Also, question 2 is not really difficult for a primary teacher in Indonesia with $x^* = 2$, although perhaps the numbers used in question 2 makes that the predicted performance even for this class of teachers is not closer to 1.

We can use the results of the model to make predictions for the entire distribution of today's true teachers' test scores. Even though true scores x^* are not observed at the individual level, we can still approximate the distribution of true scores by using standardized observed scores.[10]

[10]True scores x^* have mean 0 and a standard deviation of 1 by construction. Standardized observed scores also have this property. For each individual test taker, the true score is generally not the same as the observed score. The distribution however does look the same, which is what matters here.

Figure (3) presents today's predicted performance (the blue bars), in combination with the predicted performance when all teachers have increased their subject-matter knowledge by 1.0 (red bars) or 2.0 standard deviations (green bars) respectively.

The bars are splits based on quintiles of today's true score distribution. The bottom quintile, i.e. poorest performing 20% of today's population of teachers and representative of more than 200,000 Indonesian primary teachers, currently scores below 20% on item 1. Subsequent to an increase in subject-matter knowledge across the board of 1.0 and 2.0 standard deviations, we would expect these currently poor performing teachers to score 45% and 65% on the first item. For the other quintiles the model predicts more moderate increases.

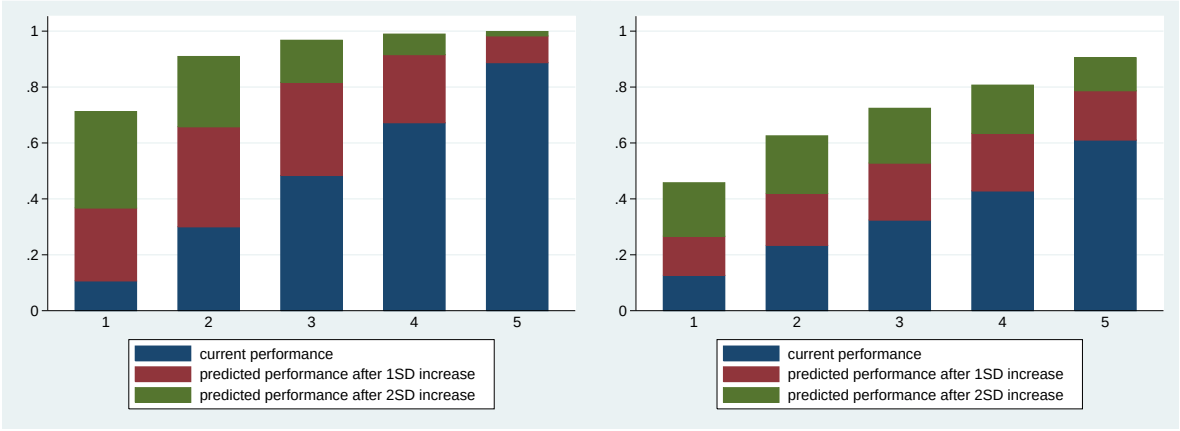


Figure 3: Current and predicted performance on item 1 (left panel) and item 2 (right panel), by overall true score quintile

A 1.0 and 2.0 standard deviation increase in knowledge are reflected in a somewhat substantial improvement in performance on the easier question 1, and somewhat more moderate increases in performance on the more challenging question 2. Designing and implementing policies that ensure lasting increases in subject-matter knowledge among teachers is a formidable challenge for policymakers in Indonesia. But the analysis provides clear evidence that such improvements are not quite out of reach. Much of these improvements can be realized within the existing education support structures in Indonesia. Especially the teacher working groups –known in Indonesia as *kelompok kerja guru* (or *KGG*)– can be used as a vehicle to provide extra training and practice. These teacher working groups exist broadly in Indonesia, and over 90% of the teachers in our sample are more or less active participants. In summary,

this section shows that subject-matter knowledge among Indonesian primary teachers is low, as a majority of the primary school teachers in Indonesia face difficulties with fairly elementary mathematical exercises. But this section also shows that 1.0 and 2.0 standard deviation increases in subject-matter knowledge can be realistic policy targets for the medium term.

3 Statistical value-added models

Theorists and empiricists in the field of education research often refer to 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 a child's scholastic achievement. The statistical models used to confront this theoretical idea to data on test scores, are called value-added models. Value-added models can be derived from education production functions, where each input from birth until today contributes to current achievement (Todd and Wolpin 2003). Value-added models come in various forms (Guarino, Reckase, and Wooldridge 2015). The version I use in this paper is the following:

$$y^*_{sit} = inputs_{sit} + \gamma y^*_{sit-1} \tag{2}$$

Where $y^*_{si\tau}$ are student i's true achievement scores on subject s at period τ. The value-added model links last year's achievement y^*_{sit-1} to current achievement y^*_{sit}, and $inputs_{sit}$ happening in-between. The literature often distinguishes *true* and *observed* achievement scores. True achievement is the actual achievement level of an individual, and observed achievement is a noisy proxy value of this obtained by test scores.[11] $inputs_{sit}$ in equation (2) may be all factors contributing to learning in a given year, from school inputs (e.g. teachers), to parental support, to individual (innate) interests, motivations and talents.

The model describes that the reasons for observing differences in achievement between students at a given point in time are either due to differences in inputs between period $t-1$ and t (e.g. some have better teachers than others) or to differences in prior achievement

[11] The closer the test score is measuring some dimension of true achievement the less noisy test scores are, and the more reliable they are. In real world situations, test scores are always subject to more or less noise.

(which is due to the accumulation of inputs from birth until period $t-1$). The subject-matter knowledge of teachers is potentially one of the important input factors. We decompose the totality of inputs as follows:

$$inputs_{sit} \quad = \quad \beta x^*_{sit} + u_{sit} \tag{3}$$

where x^*_{sit} is a standardized measure of teachers' subject-matter proficiency. β measures the causal effect of a standard deviation increase in a teachers' subject-matter proficiency on student-learning outcomes on a year-to-year basis. The residual term u_{sit} captures the remaining inputs, and is a complicated function of unmeasured teacher abilities, parental factors, and individual student abilities, talents and interests, each influencing learning gains of student i from period $t-1$ to t on subject s. Assumptions about the nature of u_{sit} determine under which conditions we can estimate the causal parameter β consistently.

In this paper, true achievement scores are standardized, i.e. they have a mean of 0 and a variance of 1.

$$E\left[y^*_{sit}\right] = E\left[y^*_{sit-1}\right] = E\left[x^*_{sit}\right] = 0 \tag{4}$$

$$V\left[y^*_{sit}\right] = V\left[y^*_{sit-1}\right] = V\left[x^*_{sit}\right] = 1 \tag{5}$$

Other restrictions on the theoretical model would be possible. One may use test equating techniques in an attempt to place the pre and and post test scores y^*_{sit} and y^*_{sit-1} on the same scale for example [see Andrabi, Das, Khwaja, and Zajonc (2011) and Rothstein (2007) who make such adjustments]. Such approaches are not strictly necessary however for an intuitive interpretation of the results. The conditions (4) and (5) above mean that the value-added specification measures relative gains in the distribution of achievement, rather than absolute gains.[12]

[12] As a result of this, we know also that the average student, say one with a prior achievement score of 0, maintains his position in the distribution when his or her inputs are average as well. With above average inputs, a student will gain in relation to its peers and will score above the mean in the next period. It is straightforward to show that students with above average prior achievement also need above average inputs to prevent a loss in position. Naturally, to maintain a position at the bottom of the distribution, only below average inputs are required.

Combining the theoretical value-added specification (2) and the input split (3) yields something that looks like a dynamic regression model:

$$y^*_{sit} = \beta x^*_{sit} + \gamma y^*_{sit-1} + [u_{sit}] \tag{6}$$

The effect parameter β and the input persistence parameter γ are the parameters of interest of this research. Two major hurdles need to be taken before we can estimate these parameters however. Both issues are well established and the literature has proposed ways of dealing with them (see for example, Andrabi, Das, Khwaja, and Zajonc (2011)). The first problem is the likely *persistence in the unobserved inputs* u_{sit} and omitted variable bias more generally. And the second is the *measurement error in test scores*, which serve as proxies for the unobserved true achievement scores in (6).

3.1 Omitted variable bias: Three different empirical specifications

Suppose, for the sake of exposition, that true achievement scores y^*_{sit}, y^*_{sit-1} and x^*_{sit} are observed in the data. Even in these ideal circumstances, an OLS regression of today's achievement scores y^*_{sit} on the teacher's test score x^*_{sit} and the student's own lagged score y^*_{sit-1}, is probably not yielding consistent estimates of the parameters of interest β and γ. One basic problem is that the residual inputs u_{sit} are likely correlated with teachers' subject-matter proficiency x^*_{sit} and, perhaps more importantly so, with a student's lagged achievement y^*_{sit-1}.

The level model. A text-book issue related to dynamic (panel-data) regression models like (6) is that in most imaginable situations, the term u_{sit} is persistent in time. More able or more intelligent students learn more today, but are also more likely to have higher achievement levels to start out with, because they have also learned more last year. In this scenario there exists a positive correlation between the residual inputs term u_{sit} and the lagged achievement score y^*_{sit-1}. The idea is operationalized by decomposing the residual term u_{sit} into a fixed,

and a time-varying component:

$$y^*_{sit} = \beta x^*_{sit} + \gamma y^*_{sit-1} + [\overline{u}_{si} + (u_{sit} - \overline{u}_{si})] \quad (7)$$

$$= \beta x^*_{sit} + \gamma y^*_{sit-1} + [\eta_{si} + \varepsilon_{sit}] \quad (8)$$

where I assume that ε_{sit} is uncorrelated over time.

In this paper I follow Blundell and Bond (1998) who argue that in stable ($\gamma < 0$) dynamic systems $\Delta y^*_{sit-1} = y^*_{sit-1} - y^*_{sit-2}$ can be used as an instrumental variable (IV) for y^*_{sit-1}, because Δy^*_{sit-1} is uncorrelated with η_{si} and correlated with y^*_{sit-1}. Note that this IV strategy requires *three waves* of panel data on test scores.

But whereas the IV solves some problems it does not solve all. For consistency of both model parameters in the level specification also x^*_{sit}, the teacher's achievement score, should be uncorrelated with $\eta_{si} + \varepsilon_{sit}$. It is easy to think of reasons why x^*_{sit} and $\eta_{si} + \varepsilon_{sit}$ are correlated. One is for example that better performing (high ability or wealthier) students self-select into better, and/or better funded schools. This idea is clearly relevant in the Indonesian context. Schools in urban centers are usually better funded, and therefore better able to attract better teachers. The best schools (or at least those who are perceived as the best) are also able to select the best students, or those willing to pay more to enroll. If more able, or better supported children (who learn more than others regardless of the schools they enroll in) also populate the schools with the best teachers, subject-matter ability of teachers x^*_{sit} and the residual inputs u_{sit} are positively correlated. Such mechanisms pose a threat to the validity of the level specification.

The **school fixed-effects model** estimates the parameters of interest from variation across subjects and classrooms, but *within* the same school. If the matching of better or better supported students to more knowledgeable teachers in better and better funded schools happens only at the level of the school, and not within schools, a school fixed-effects model would account for it. Where ability matching within schools might happen in some cases, it does not seem to be happening at scale in Indonesia (see section (4.1) for an analysis on selection).

The school fixed-effects model relies on the variation in subject-matter knowledge between teachers across different classrooms. And as endogenous matching within schools seems unlikely to be happening at scale, the statistically significant results we find in the results section (4) indicate that teachers with more subject-matter knowledge are indeed better teachers. While this finding is important in its own right, it does not imply that teachers with more subject-matter knowledge are better *because* they have more knowledge. It could be, for example, that teachers who score higher on subject-matter tests tend to be more motivated, and that it is the additional motivation that makes them perform better. For the purpose of policymaking it is important to establish the causal effect of additional subject-matter knowledge, by isolating the subject-matter knowledge effect from other factors such as motivation and other (correlated) skills.[13]

In an attempt to do this, we estimate a **student-teacher fixed-effects model**. The idea of the model is to only rely on variation in test scores for a single student-teacher pair, across three different subjects: math, science and Indonesian language. If a teacher's subject-matter knowledge were truly important we would observe that students tend to do better in one subject (say math) than in another subject (say Indonesian language) if his or her teacher is better in math than languages. The idea of using variation across different subjects in an attempt to isolate the causal effects of teachers' subject-matter knowledge on student learning is not new. Metzler and Woessman (2012) already used the idea on Peruvian data and find that a 1.0 standard deviation increase in the subject-matter knowledge of teachers leads to a 0.10 standard deviation increase in student learning. Metzler and Woessman (2012) however estimate their model on a cross section of data only.

3.2 Measurement error in test scores: From true achievement to observed test scores

The x^*'s and y^*'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 measure

[13]If the subject-matter knowledge scores used in the analysis only proxy for differences in the level of motivation, government programs to improve the levels of subject-matter knowledge of teachers would not be effective as they do not change levels of motivation.

"true" achievement scores only with some level of error. Differences between observed scores and the theoretical true scores are due to good/bad luck with guessing if test-takers are not fully convinced about an answer. Test-takers might perform better or worse depending on the time-of-day of the field visit, or depending on whether they are having problems at home, etc. The fact that observed test scores do not in general completely reflect a student's level of true achievement is important in empirical value-added modeling. Much of the literature acknowledges the need to correct for measurement error in empirical operalizations of value-added models (see Andrabi, Das, Khwaja, and Zajonc (2011) and many others).

Measurement error however in testing data yields two, somewhat different, problems. The first problem relates to the scaling of test scores. Prior to using test score data in value-added regressions, choices must be made on *how to scale* the test results. Simply using standardized (raw, percentage correct) test scores is not always appropriate, as gains on more noisy standardized scores would appear smaller than gains on cleaner standardized scores. This is because the variance of the "signal" in noisy standardized test score data is smaller than the variance of the "signal" in cleaner standardized test score data. This measurement issue appears to be somewhat neglected in the literature. The second is a problem more commonly associated with (classical) measurement error, i.e. the attenuation bias due to an automatic correlation between the noisy regressor and the measurement error component (which is part of the error term of the regression equation).

3.2.1 Noisy test scores and test scaling

I assume throughout that the measurement error in test scores is "classical" in the sense that it has mean 0 in the population, and is uncorrelated with the true scores. The standard model for (classical) measurement error is the following:

$$y_{sit} = y^*_{sit} + e_{sit} \tag{9}$$

where e_{sit} is the measurement error term, i.e. the difference between the observed score y_{sit} and the true score y^*_{sit}. Based on this relationship we would be able to write the value-

added specification in terms of the observed scores y_{sit}. One complexity however is that we do not know the variance of y_{sit}, so that we do not know a priori how to construct it properly. The value-added model is defined in terms of standardized true scores, and the parameters of interest β and γ are interpreted with respect to this normalization. To maintain the normalization in estimation we should therefore construct the true scores so that $E[y_{sit}] = E[y^*_{sit}] + E[e_{sit}] = 0$ and $V(y_{sit}) = V(y^*_{sit}) + V(e_{sit}) = 1 + V(e_{sit})$. In other words the variance of the observed score used in the analysis, should have a variance that is *greater than one*. How large it should be exactly, indeed, depends on the variance of the measurement error component.

Without some further analysis we do not know how large the variance of the measurement error component is, and, consequently, how to set the variance of the observed score y_{sit}. It is straightforward to show that, in general, the following conditions hold for the observed score y_{sit}, i.e the observed score that maintains the model normalizations $E[x^*] = 0$ and $V(y^*_{sit}) = 1$, and the assumption of classical measurement error $E\left[e\middle|x^*\right] = 0$:

$$E[y_{sit}] = 0 \tag{10}$$

$$V(y_{sit}) = 1/\rho_{y_{sit}} \tag{11}$$

where $\rho_{y_{sit}} = \frac{V(y^*_{sit})}{V(y_{sit})}$ is the well-known coefficient of reliability.

The test scores y_{sit} used in the analysis is therefore constructed by the following transformation of the raw scores (percentage correct) R_{sit}:

$$y_{sit} = \frac{R_{sit} - \mu_{R_{sit}}}{\sigma_{R_{sit}}\sqrt{\rho_{y_{sit}}}} \tag{12}$$

The transformation applies to all test scores used in the analysis, where the mean and the standard deviation of the raw scores $\mu_{R_{sit}}$ and $\sigma_{R_{sit}}$ can be calculated using the data.[14] The reliability coefficient $\rho_{y_{sit}} = \rho_{R_{sit}}$ is estimated based on the correlation between two splits of the raw test score data, one split is based on the even-numbered test items 2, 4, 6, etc.

[14] In the analysis the student scores are standardized by grade level and subject and the teacher scores are standardized by subject only.

while the other split is based on the odd-numbered test items 1, 3, 5, etc.[15] The even-odd split-half reliability score is a common measure of reliability used in the literature, similar to Cronbach's alpha for example.[16] Appendix B derives how this works exactly, and under which assumptions.

It should be noted however that Cronbach's alpha and the even-odd split-half reliability coefficients tend to *overestimate* the true reliability of the test score data. More specifically, such estimators tend to assume that measurement error is independent across the difference items of the test. This is a strong assumption and most probably incorrect in most applications, including ours. If test-takers are tired for example, they will perform below their true potential on multiple items on the test, so that measurement error is positively correlated across the different test items. It is important to keep this caveat in mind.

Incorporating observed scores into equation (2) yields the value-added specification in terms of observed (scaled) test scores:

$$y_{sit} = \beta x_{sit} + \gamma y_{sit-1} + [e_{sit} - \beta a_{sit} - \gamma e_{sit-1} + u_{sit}] \tag{13}$$

where e_{sit}, a_{sit} and e_{sit-1} are the measurement error components associated with y_{sit}, x_{sit} and y_{sit-1} respectively.

3.2.2 Noisy test scores and attenuation bias

While section (3.2.1) deals with the issue of scaling in the presence of measurement error, this section deals with the second measurement error problem. The measurement error components, now part of the composite error term of equation (13), correlate with the structural components of the model, and cause biases. We can broadly distinguish two types of solutions to this problem. One is to first estimate the parameters *with bias*, and correct estimates afterwards. The second is to account for potential biases directly in estimation. In this paper

[15] It is straightforward to show that the reliability of the raw (fraction correct) scores R_{sit} is the same as the reliability of the scaled scores y_{sit}.

[16] Cronbach's alpha is a construct that is very similar to the even-odd split half reliability coefficient, both in nature as in results. Cronbach's alpha is the average of the correlation of all possible halve splits of the test. In our data, the values we obtain for Cronbach's alpha and for the even-odd split half reliability coefficient are practically the same.

I follow the second approach.

The observed equivalent of the term Δy^*_{sit-1}, which was introduced as an instrumental variable in section (3.1), is now no longer a good instrument. The observed lagged change score $\Delta y_{sit-1} = y_{sit-1} - y_{sit-2}$ is positively correlated with the measurement error term e_{sit-1}, which is a component of the error term in (13). Using Δy_{sit-1} anyway, would yield attenuated parameter estimates of the γ parameter in equation (13). The solution for this I propose in this paper makes use of the even-odd splits again. Instead of using y_{sit-1} as a right-hand-side regressor, we can also use $y_{E,sit-1}$ (the scaled test score, using only the even-numbered items 2, 4, 6, etc. of the test). This changes the measurement error component in the error term:

$$y_{sit} \quad = \quad \beta x_{sit} + \gamma y_{E,sit-1} + [e_{sit} - \beta a_{sit} - \gamma e_{E,sit-1} + u_{sit}] \tag{14}$$

The idea is now to use $\Delta y_{O,sit-1} = y_{O,sit-1} - y_{O,sit-2}$ (based on the odd-numbered testing items 1, 3, 5, etc.) as an instrument for $y_{E,sit-1}$. $\Delta y_{O,sit-1}$ is a good instrument under the assumptions introduced before, i.e. that Δy^*_{sit-1} is a good instrument, and that measurement error components are uncorrelated across the different items of the test.

The next problem is that observed teacher's test scores x_{sit} correlate positively with a_{sit}. During the time span of the study however, the teachers were tested a maximum of three times, i.e. if they were present during each of the three field visits. In the analysis we use a one period lagged teacher score x^{t-1}_{sit} as an instrumental variable for x^t_{sit}, where the superscript indicates when test was fielded. It is important to notice that the lagged teacher's test score x^{t-1}_{sit} is not the test score of last year's teacher, but the test score of student i's current teacher, of which the score was obtained from a test that was fielded last year, at $t-1$. By contrast x^t_{sit} is the test score of student i's current teacher, obtained from a test that was fielded at t.

4 The data, stylized facts and results from empirical value-added models

The data used in this study was collected for a randomized controlled field experiment with the objective to establish the causal effects of Indonesia's teacher certification program, which included a doubling of a teacher's base pay (De Ree, Muralidharan, Pradhan, and Rogers 2015). 360 schools were sampled using a two-step sampling strategy. First, 20 districts were selected from 10 broader geographical strata, and second, within each district, 12 primary and 6 junior secondary schools were selected.[17] The data on the 240 primary schools are used in this paper. Within these 240 schools, all class teachers ($N \approx 1700$) were tested and interviewed.[18] In addition, all students ($N \approx 45,000$) in the 240 sample schools were tested as well. The multiple choice tests for teachers and students consisted of a mathematics, a science and an Indonesian language component.[19] At the back of the answering sheets, students also filled out a brief questionnaire on wealth indicators and on parent education.[20]

Further selections on the data were made prior to the analysis. First, the data from one of the districts (*Maluku Tenggara Barat*) were dropped from the analysis, because baseline data was collected about 6 months later as for the rest of the sample. Second, students who repeated class at least once are not used in the analysis. Third, because the statistical model requires three periods of measurement on test score data, we only rely on the grade levels which we can track for three consecutive years, i.e. $2 \to 3 \to 4$, $3 \to 4 \to 5$ and $4 \to 5 \to 6$. Fourth, we only use classrooms for which we can match students to teachers, for which we have testing information for both students and teachers, and for which the class has a single teacher for

[17] See World Bank (2015) and De Ree, Muralidharan, Pradhan, and Rogers (2015) for more details on the selection of schools.

[18] In general, all "core subject teachers" were tested an interviewed, which are class teachers in primary schools, and mathematics, Indonesian language, Physics, Biology, and English language teachers in Junior secondary schools.

[19] Tests for teachers were different from tests for students, and grade appropriate tests were used to test students.

[20] The information on parent education was not used in this paper as it was found that the wealth indicators correlate more strongly to the student's test scores than the information on parental education. One reason for this might be that young children might not know the education levels of their parents. About 90% of the students in our data reported to have a TV at home. They are however much less likely to have a fridge (50%) or a car (15%). Similarly, parents are quite likely to have some level of education, but generally not more than secondary school.

all three subjects: mathematics, Indonesian language, and science. A minority of the primary schools in the data organize instruction differently, for example, by having teachers teaching a single subject to grade 4, 5 and 6, similar to the common way of instruction in secondary schools.

4.1 Stylized facts

In a preliminary analysis we find that teacher test scores do not correlate particularly strongly with some of the other observable teacher characteristics. Frequencies of some teacher characteristics are presented in figure (4).[21]

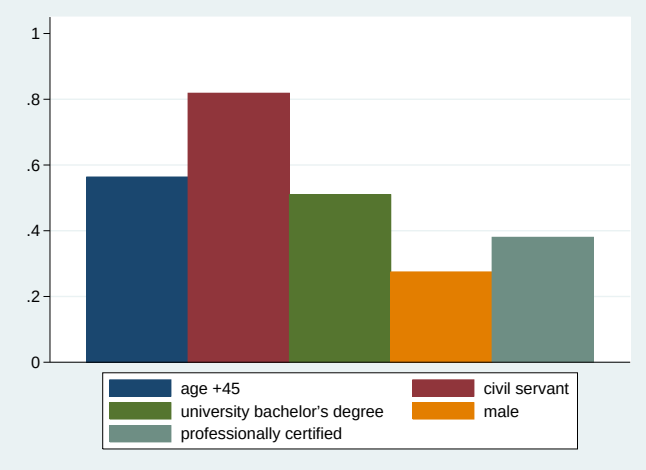


Figure 4: Background characteristics of primary school teachers in the sample

Table 1 presents results of regressions of the scaled teacher test score x_{sit} on five background characteristics. The columns marked "pairwise", columns (1), (3), (5), and (7), present results from univariate regressions, where the scaled test score x_{sit} is regressed on one background characteristic at the time. The columns marked "pooled", the columns (2), (4), (6), and (8) present results from multivariate regressions where all five regressors are included at the same time. The results show that older teachers do worse and teachers with university bachelor's degrees do better. Teachers in the civil service do not score better or worse, and the same is true for certified teachers.[22] An interesting pattern is also that male teachers tend to score

[21] Notice that our sample of teachers is not fully representative for the total population of teachers in Indonesia.

[22] Especially the latter could be seen as a problem, as it indicates that the roll-out of the certification program

better at the math and science components, while women score score better at Indonesian language.

Are these point estimates large or small? What would we expect a priori of a bachelor's education for example? Of those without a bachelor's degree, about two-thirds has a 2-year post-secondary diploma, and the rest, one-third, has only a secondary education. This heterogenous group scores .4 of a standard deviation below the group with a bachelor's degree on average. Not very much perhaps. Another way of looking at this is by looking at the R^2 of these regressions. They are around 0.05, showing that only 5% of the variance of the (noisy) test score data is explained by these characteristics. Discounting the fact that test scores are noisy, we can estimate an alternative R^2, measuring what the R^2's would have been if the testing data would have been noise-free.[23] Even these alternative R^2's are not very big, indicating that these background characteristics do not explain much of the variation in subject-matter knowledge of teachers. Apparently there are many teachers without university diplomas who do very well on the test, and, similarly, many teachers with university diplomas who do badly. This suggests that Indonesia's teacher training programs – on average – do not manage to lift teacher subject-matter knowledge to a much higher level.

With more flexible regression models we can explain much more of the variation in test scores than the models presented in table 1. School fixed-effects models (where the teacher test score is regressed on a full set of school dummy variables) have R^2's in the neighborhood of 0.60, indicating that 60% of the variance of (noisy) test scores can be explained at the school level. (We note that these school fixed-effects models are prone to overfitting, as the number of teachers per school is probably too small. I ran a bootstrap analysis to investigate the extent of the overfitting problem in this setup.[24] The findings indicate that a more conservative estimate of the R^2 of the school fixed-effects model would be 0.45 or 45%. Still much higher

does not favor teachers with higher levels of subject-matter knowledge. See also World Bank (2015).

[23]The standard R^2 is defined as the variance of the prediction, divided by the variance of the left hand side variable. The alternative R^2 reported in table 1 is calculated as the variance of the prediction, divided by 1, the variance of the true score.

[24]For the bootstrap, the teacher's test scores are randomly, reallocated across teachers after which we estimate the school fixed-effects model again. This procedure is repeated 1,000 times and the R^2's of these regressions are stored. I find that on average, these artificial regressions have R^2's of around 0.15. The overfitting alone therefore already accounts for 15% of the variance in the test scores. These R^2's however are much lower than the R^2 of 0.60 we find in the data, suggesting that there is a lot of real clustering of subject-matter knowledge in schools left.

Table 1: explaining teachers' test scores – by subject

	math score		Indonesian score		science score		pooled score	
	(1) pairwise	(2) pooled	(3) pairwise	(4) pooled	(5) pairwise	(6) pooled	(7) pairwise	(8) pooled
age +45 years	-0.216***	-0.292***	-0.429***	-0.419***	-0.249***	-0.366***	-0.325***	-0.393***
civil servant	-0.046	0.075	-0.387***	-0.208*	0.002	0.152	-0.154*	0.010
university degree	0.387***	0.359***	0.304***	0.273***	0.499***	0.457***	0.437***	0.400***
male teacher	0.200**	0.199**	-0.088	-0.083	0.242***	0.245***	0.134*	0.137**
teacher is certified	0.087	0.082	0.002	0.161*	0.144	0.114	0.086	0.129
R^2	.	0.036	.	0.041	.	0.049	.	0.061
alternative R^2	.	0.059	.	0.078	.	0.095	.	0.081

NOTE. Regressions of teacher test scores on background characteristics. The test scores are standardized, then rescaled as described in equation (12). By subject, the "pairwise" column presents results of univariate regressions (the test score regressed on a single characteristic). The "pooled" column presents the results of a multivariate regression with five regressors, age +45 years, civil servant, teacher has a university degree, male, and whether the teacher is professionally certified. These five regressors are all binary variables. Standard errors are robust to clustering at the school level. * significant at 10%; ** significant at 5%; *** significant at 1%.

than the R^2's reported in table 1.) These findings indicate that there is a lot of clustering of high, and low levels of subject-matter knowledge in schools. Where in some schools all teachers tend to score higher, regardless of their degree or age, while in other schools all teachers tend to score lower. This is an indication that there is some important inequality of opportunity for Indonesian students.

Of particular interest for the empirical analysis is whether children from higher socio-economic groups are more likely to attend these high-performing schools. Table 2 presents the relationship between socioeconomic characteristics of students and the mean test scores of their teachers. In column (1) the scaled teacher score is regressed on a student asset index[25], and column (2) presents the results of a school fixed-effects model. Column (1) shows clearly that schools with the better, i.e. more knowledgeable, teachers also cater to wealthier families. Within schools, the column (2) results, we no longer observe this association, indicating

[25]The asset index counts how many of the following eight assets are available to the household: tv, fridge, hand phone, bicycle, motorcycle, car, computer, and children's books. The parameter on the asset index, therefore, measures how many standard deviations a teacher's subject-matter knowledge is higher on average, if its student has one more of these assets. The mean of the asset index is about 4.5 with a standard deviation of just below 2.

that there are no selection effects within schools.[26] The results here provide support for the relevance of the fixed-effects approaches discussed in the theoretical section (3). Students are certainly not randomly distributed across schools (or teachers) which is a clear threat to the validity of value-added models in levels: correlations between teachers' subject-matter knowledge and student learning gains may be due, in part, to the fact that better teachers teach children with wealthier parents.

Table 2: testing for endogenous selection into schools

	pooled	school FE
asset index	0.067***	-0.004
p-value	0.003	0.359

NOTE. Student level regressions where the teacher's scaled test score (see equation (12)) is regressed on a student asset index. In the analysis I use the same data as for the value-added regressions presented in table 3, that is, endline data and grade level's 4, 5 and 6. The asset index counts how many of the following eight assets are available to the household: tv, fridge, hand phone, bicycle, motorcycle, car, computer, and children's books. The parameter on the asset index, therefore, measures how many standard deviations a teacher's subject-matter knowledge is higher on average, if its student has one more of these assets. The mean of the asset index is about 4.5 with a standard deviation of just below 2. The "pooled" column (1) presents the pairwise regression of the teacher's score on the asset index. The "school FE" column (2) present the results of a school fixed-effects model. Standard errors are robust to clustering at the school level. * significant at 10%; ** significant at 5%; *** significant at 1%.

4.2 Results from empirical value-added models

Table 3 presents results of the three types of value-added models we proposed in section (3). Column (1) and (2) are the level models, column (3) and (4) are the school fixed-effects models, and column (5) and (6) are the student-teacher fixed-effects models. The column (1), (3) and (5) regression models include the lagged student score based on the even-numbered items in the test, i.e. $y_{E,sit-1}$, while the change score $\Delta y_{O,sit-1}$ based on the odd-numbered items, is one of the two excluded instruments. For the column (2), (4) and (6) results this is reversed, i.e. $y_{O,sit-1}$ is used as a regressor and $\Delta y_{E,sit-1}$ is the excluded instrument. The two sets of results should be fairly similar, and we find that they are in general.

[26] The argument here is that if there are no selection effects based on observables, it is plausible that there are also no selection effects based on unobservables.

Table 3: Results from value-added models

	Level				School fixed-effects				Student-teacher fixed-effects			
	$y_{E,sit-1}$ included		$y_{O,sit-1}$ included		$y_{E,sit-1}$ included		$y_{O,sit-1}$ included		$y_{E,sit-1}$ included		$y_{O,sit-1}$ included	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)	(6a)	(6b)
	parameter	se	parameter	se	parameter	se	parameter	se	parameter	se	parameter	se
Teacher's subject-matter knowledge (x^t_{sit})	0.291***	0.062	0.256***	0.058	0.222***	0.078	0.195**	0.082	0.201**	0.092	0.149*	0.090
Teacher has a bachelor's degree	0.003	0.050	0.025	0.048	0.021	0.050	0.052	0.050	.	.	.	.
Teacher's age	0.014***	0.004	0.012***	0.003	0.008*	0.005	0.007*	0.004	.	.	.	.
Teacher is certified	-0.039	0.062	-0.042	0.061	-0.022	0.082	-0.015	0.082	.	.	.	.
Teacher is civil servant	-0.031	0.090	-0.015	0.083	-0.077	0.104	-0.091	0.097	.	.	.	.
Class size	0.001	0.002	0.000	0.002	-0.004	0.005	-0.003	0.005	.	.	.	.
Student asset index	0.051***	0.014	0.052***	0.012	0.038***	0.008	0.038***	0.007	.	.	.	.
Student asset index is missing	0.165	0.213	0.251	0.227	0.023	0.168	0.123	0.180	.	.	.	.
Lagged student test score $y_{E,sit-1}$ (even items)	0.507***	0.049			0.424***	0.049			0.282***	0.075		
Lagged student test score $y_{O,sit-1}$ (odd items)			0.565***	0.040			0.512***	0.040			0.328***	0.077
F-stat underidentification test of first stage regressions	77.7		79.9		20.4		20.5		45.2		37.2	
p-value underidentification test of first stage regressions	0.000		0.000		0.000		0.000		0.000		0.000	
Number of clusters (schools)	186		186		186		186		186		186	
Number of obs. (student-subject)	28425		28425		28425		28425		28392		28392	
excluded instruments	$\Delta y_{O,sit-1}$ x^{t-1}_{sit}		$\Delta y_{E,sit-1}$ x^{t-1}_{sit}		$\Delta y_{O,sit-1}$ x^{t-1}_{sit}		$\Delta y_{E,sit-1}$ x^{t-1}_{sit}		$\Delta y_{O,sit-1}$ x^{t-1}_{sit}		$\Delta y_{E,sit-1}$ x^{t-1}_{sit}	

NOTE. Standard errors are robust to clustering at the school level. * significant at 10%; ** significant at 5%; *** significant at 1%. In the analysis, student test scores are standardized (through formula (12)) by grade level and subject, and teacher test scores are standardized by subject only.

In the **level specifications** we find strong relationships with the teacher's test scores. A standard deviation increase in teacher's subject knowledge is associated with 0.25 to 0.29 standard deviations of additional learning on a year-to-year basis. Also we find that more senior teachers appear to do better. In this multivariate model there are no significant relationships with a teacher's level of education, whether the teacher is certified, whether the teacher is a civil servant, and with the size of the class. Instead there is a strong relationship between learning, and a student's socio-economic background. The parameter on the student's lagged score, i.e. the persistence parameter γ in equation (14), is estimated at around 0.5 in the level specification (this estimate is similar to estimates presented in the literature, e.g. by Andrabi, Das, Khwaja, and Zajonc (2011)).

The analysis in the previous section suggests that there is important endogenous selection within schools (Table 2). Children from wealthier backgrounds are more likely to enroll in schools where teachers have higher levels of subject-matter knowledge. And controlling for the socioeconomic background characteristics alone is probably not sufficient to account for the entire selection effect. In the school fixed-effects model, however, we still obtain sizable and statistically significant parameter estimates associated with teachers' subject-matter knowledge. Because the data does not suggest that there are selection effects within schools, the school fixed-effects results are important. They suggest that *teachers with more subject-matter knowledge tend to be better teachers.*

The same seems to be the case for older, more experienced teachers, although the age effect in the school fixed-effects model is only significant at the 10% level. Older teachers appear to do a bit better. It is beyond the scope of this paper to further explore this result. One important question is whether this "age effect" is truly an age effect, i.e. with age and experience, teachers become better. Or whether it is really a cohort effect, where older cohorts are simply better because, for example, the quality of teacher training has decreased over time. The latter explanation would be especially worrisome of course as it would suggest that when older more experienced teachers retire, the quality of instruction would decrease. More research is needed in this area.

Where the results of the school fixed-effects model indicate that teachers with more subject-

matter knowledge are better teachers, it does not prove that teachers are better *because* they have more subject-matter knowledge. Based on videotaping Indonesian classrooms, Chang, Shaeffer, Al-Samarrai, Ragatz, De Ree, and Stevenson (2013), Ragatz (forthcoming) (and related work) find that teachers with higher levels of subject-matter knowledge also teach differently on average. And interestingly, it appears that the way teachers with more knowledge teach differently, relates directly to the additional knowledge. For example, Ragatz (forthcoming) argues: "techniques such as investigation, open-ended questioning, the use of mathematical language and symbols, and the use of non-routine problems and applications were found more often in classrooms with higher-knowledge teachers. This appears to be in part because the techniques themselves require a greater amount of subject mastery to be conducted effectively." This finding lends weight to the idea that increases in the level of subject-matter knowledge among Indonesian teachers could lead to improved student outcomes. The student-teacher fixed-effects model evaluates the hypothesis empirically.

The **student-teacher fixed-effects** model attempts to disentangle the subject-matter knowledge of teachers from other, unobserved, teacher attributes, such as general intelligence, motivation, etc. The idea of the model is to only employ variation across subjects for the same student-teacher pair. Would it be, for example, that students with teachers who know more math than science, learn more math than science? The results from the student-teacher fixed-effects model confirms that subject-matter knowledge is important, even if we control for all non-subject specific student and teacher attributes. The point estimates are smaller in magnitude than in the level and school fixed-effects specifications. This is consistent with the idea that part of the reason why teachers with more knowledge are better, is because they are also more skillful along other dimensions. Perhaps naturally, the standard errors are larger for the student-teacher fixed-effects results, which, in combination with the size of the estimated effects, makes that the results are somewhat less robust overall.[27]

The point estimates presented in this paper tend to be larger than comparable parameter estimates presented by Metzler and Woessman (2012). The point estimates we present here are closer to 0.2, while Metzler and Woessman (2012) present point estimates of around 0.1.

[27]Parameters for all other other fixed teacher and class attributes (age, education level, etc.) are not identified in the student-teacher fixed-effects model as they do not vary across subjects.

Another difference between this paper and Metzler and Woessman (2012) is that we also estimate the persistence parameter γ based on variation across subjects only. The size of the point estimates for γ tend to decrease from the level model to the school fixed-effects model to the student-teacher fixed-effects model.

4.2.1 Projected compounding effects of improvements in teacher subject-matter knowledge

The primary interest of this research is to measure the counterfactual learning effects of improvements in teacher subject-matter knowledge. The parameter estimates presented in table 3 measure the input effects on a year-to-year basis. But nation-wide efforts to increase teacher subject-matter knowledge will imply that students have better teachers for two, three years in a row, or throughout their entire schooling career. The overall effect on student-learning outcomes, therefore, depends on the strength of the input effect but also on the rate at which past achievement gains persist into the future.

Within the context of our model we predict what would happen to student outcomes in response to an increase in teachers' subject-matter knowledge of 1.0 and of 2.0 standard deviations across-the-board. Figure (5) shows projections based on the results from the student-teacher fixed-effects model. For the projections I took the average of the parameter estimates of column (5) and (6), i.e. $\hat{\beta} \approx 0.175$, and $\hat{\gamma} \approx 0.3$. Students benefit most (as measured relative to the distribution of performance in the current population) in the first year after an improvement in teachers' subject-matter knowledge. The increase is smaller in the second year as the second year's teacher receives the student already at a higher starting level. Students who benefit year-after-year from teachers with higher levels of subject-matter knowledge practically reach the new equilibrium level at the end of grade 3. (Note that you need teachers with higher levels of subject-matter knowledge also in grade 4, 5 and 6, to maintain higher relative test scores. Otherwise the student would fall back again.)

At the end of the primary cycle, test scores have increased by 0.25 and 0.50 student-level standard deviations respectively. These effects are substantial. PISA reports for example that the standard deviation of PISA scores for Indonesian 15-year-olds is roughly 70 PISA points in

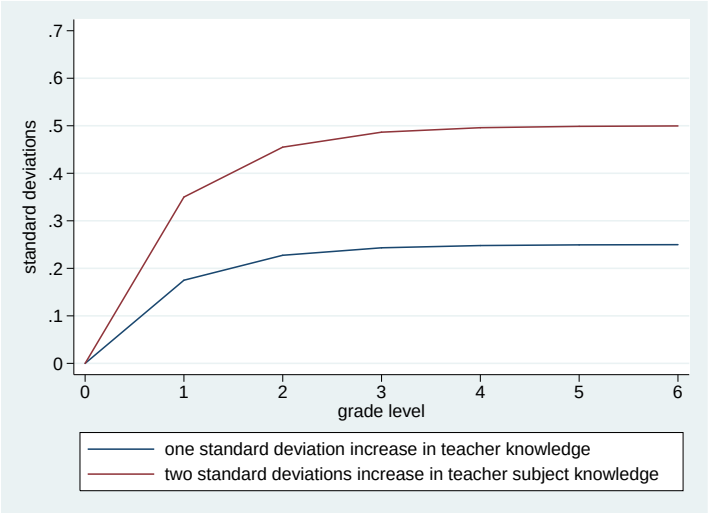


Figure 5: Projected responses across grade levels in response to 1.0 and 2.0 standard deviation increases in teachers' subject-matter knowledge

the population. 0.25 or 0.50 of that would correspond to effects of 17.5 and 35.0 PISA points respectively.[28] Especially, an increase of 35.0 PISA points would be remarkable considering Indonesia's past performance on PISA. Indonesia's long term performance on the mathematics component of PISA is about 375. With an additional 35 points Indonesia would break through the 400-point barrier, and close more than half of the current learning gap between Indonesia and its regional peers in PISA, Malaysia (with a score of 421) and Thailand (with a score of 427). Notice that section (2) shows that both 1.0 and 2.0 standard deviation increases in subject-matter knowledge of primary teachers appear to be achievable medium-term goals for education policy making in Indonesia.

5 Conclusion

This paper makes the case that subject-matter knowledge of Indonesian primary school teachers is low, in the sense a large percentage of them has difficulties with mathematics problems from the high-school curriculum. Based on a school fixed-effects model and a (more general) student-teacher fixed-effects model, the paper also finds empirical support for the idea that

[28] One caveat here is for finding these differences in performance at age 15, the year when students finish junior secondary school, we also need junior secondary teachers to improve relative to where they are now.

realistic improvements in teachers' subject-matter knowledge could lead to meaningful improvements in the quality of instruction in Indonesia and the learning levels of its students. It is estimated that a 1.0 (2.0) standard deviation increase in teacher's subject knowledge can translate into a 0.25 (0.50) student-level standard deviation increase in the learning levels of students by the time they leave primary school.

The results of this paper suggest that policies which incentivize improvements in subject-matter knowledge can work in Indonesia. This finding starkly contrasts results from experimental research about Indonesia's recent, and very costly, teacher certification program (De Ree, Muralidharan, Pradhan, and Rogers (2015) and World Bank (2015)). Indonesia's certification program was implemented in 2005/06. In a nutshell, the certification program required a university bachelor's degree as minimum academic qualifications for primary teachers, and promised a generous *doubling of take-home pay* after passing the certification program successfully. While the certification program led to hundreds of thousands of teachers starting course work to obtaining bachelor's degrees, no important improvements in education quality was observed. The stability of the PISA scores is an example of this, but World Bank (2015) provides additional *micro level* evidence.

The 2012 PISA results perhaps marked the starting points of redirecting policies more explicitly geared towards quality. Indonesia's current minister of Education and Culture Anies Baswedan for example reflected on Indonesia's most recent 2012 PISA results in *Republika*, one of Indonesia's mainstream daily newspapers, arguing that "Indonesia's education is in a state of emergency" (Novia 2014). The research presented in this paper indicates that future policies to promote quality education could benefit from incorporating subject-matter tests in teacher evaluation systems.

A Binary choice models with noisy regressors

A.1 The model with clean regressors

Consider the following model for the continuous latent variable q_k^*:

$$q_k^* = \alpha_k + x^* \beta_k + \varepsilon_k \tag{15}$$

where $\varepsilon_k | x^* \sim N\left(0, \sigma_{\varepsilon_k}^2\right)$, where x^* is a test-taker's true level of subject-matter knowledge, standardized to have mean 0 and a variance of 1 in the population.

Test-takers answer correctly to item k ($Q_k = 1$) if the latent variable $q_k^* > z_k$. The probability of answering correctly, given a certain level of subject-matter knowledge x^* is then:

$$\begin{aligned} P\left(q_k^* > z_k \middle| x^*\right) &= P\left(a_k + x^* \beta_k + \varepsilon_k > z_k | x^*\right) && (16) \\ &= P\left(\varepsilon_k \leq \alpha_k - z_k + x^* \beta_k \middle| x^*\right) && (17) \\ &= \Phi\left(\frac{\alpha_k - z_k}{\sigma_{\varepsilon_k}} + \frac{\beta_k}{\sigma_{\varepsilon_k}} x^*\right) && (18) \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal CDF. A probit model based on the clean regressor x^*, therefore provides estimates of the following composite parameters: $\frac{a_k - z_k}{\sigma_{\varepsilon_k}}$ and $\frac{\beta_k}{\sigma_{\varepsilon_k}}$. The next section A.2 how we can recover these parameters from a probit analysis based on the noisy regressor x.

A.2 The model with noisy regressors

The measurement error model is:

$$x = x^* + e \tag{19}$$

I assume that measurement error is "classical", i.e. $E\left[e \middle| x^*\right] = 0$.

The noise term e can be divided into two orthogonal components. One that has a corre-

lation of 1 with the noisy variable x and one that is orthogonal to it:

$$e = x\gamma + u \tag{20}$$

It is not difficult to show that $\gamma = \frac{C(e,x)}{V(x)} = \frac{C(e,x^*+e)}{V(x)} = \frac{V(e)}{V(x)} = 1 - \rho$, where $\rho = \frac{V(x^*)}{V(x)}$ is the well-known coefficient of reliability. Equation (21) can therefore be written as:

$$e = x\left(1 - \rho\right) + u \tag{21}$$

where $E\left[x\right] = E\left[u\right] = 0$, and $C\left(x, u\right) = 0$, by definition.

Incorporating (19) and (21) in the original latent variable model (15) yields:

$$\begin{aligned} q_k^* &= \alpha_k + x^*\beta_k + \varepsilon_k & (22)\\ &= \alpha_k + x\rho\beta_k - u\beta_k + \varepsilon_k & (23) \end{aligned}$$

The derivations so far suggest that with binary choice models with noisy regressors we obtain similar attenuation biases as what we typically get in linear models with noisy regressors (i.e. parameters on the noisy regressor are attenuated at a rate equal to ρ). But with binary choice models something else is happening which relates to the normalization of the model.

Based on (23), the new model for the latent variable q_k^*, we can derive a new probit model:

$$P\left(q_k^* > z\middle|x\right) = P\left(\alpha_k + x\rho\beta_k - u\beta_k + \varepsilon_k > z_k\middle|x\right) \tag{24}$$

where I make the additional assumption that $-u\beta_k + \varepsilon_k\big|x \sim N\left(0, \beta_k^2\sigma_u^2 + \sigma_{\varepsilon_k}^2\right)$.

$$\begin{aligned} P\left(q_k^* > z_k\middle|x\right) &= P\left(\alpha_k + x\rho\beta_k - u\beta_k + \varepsilon_k > z_k\middle|x\right) & (25)\\ &= \Phi\left(\frac{\alpha_k - z_k}{\sqrt{\beta_k^2\sigma_u^2 + \sigma_{\varepsilon_k}^2}} + \frac{\rho\beta_k}{\sqrt{\beta_k^2\sigma_u^2 + \sigma_{\varepsilon_k}^2}}x\right) & (26) \end{aligned}$$

Using (19) and the definition of reliability, we can derive that $V\left(e\right) = \left(1 - \rho\right)V\left(x\right)$, and by using (21), we can derive that $V\left(e\right) = V\left(x\right)\left(1 - \rho\right)^2 + V\left(u\right)$. Combining these yields the

following expression for $V(u)$:

$$V(u) = \rho(1-\rho)V(x) \tag{27}$$

which can be simplified, because in our normalization $V(x) = 1/\rho$.

$$V(u) = (1-\rho) \tag{28}$$

Combining (28) with (26) yields:

$$\begin{aligned} P\left(q_k^* > z_k \middle| x\right) &= \Phi\left(\frac{\alpha_k - z}{\sqrt{\beta_k^2(1-\rho) + \sigma_{\varepsilon_k}^2}} + \frac{\rho\beta_k}{\sqrt{\beta_k^2(1-\rho) + \sigma_{\varepsilon_k}^2}}x\right) & (29) \\ &= \Phi(\theta_k + \zeta_k x) & (30) \end{aligned}$$

The constant and the parameter of a regression based on the noisy variable x are:

$$\begin{aligned} \theta_k &= \frac{\alpha_k - z_k}{\sqrt{\beta_k^2(1-\rho) + \sigma_{\varepsilon_k}^2}} & (31) \\ \zeta_k &= \frac{\rho\beta_k}{\sqrt{\beta_k^2(1-\rho) + \sigma_{\varepsilon_k}^2}} & (32) \end{aligned}$$

The two equations can be solved for for $\frac{\alpha_k - z_k}{\sigma_\varepsilon}$ and $\frac{\beta_k}{\sigma_\varepsilon}$.

Solving for $\frac{\beta_k}{\sigma_\varepsilon}$ yields:

$$\frac{\beta_k}{\sigma_{\varepsilon_k}} = \frac{\zeta_k}{\sqrt{\rho^2 - \zeta_k^2(1-\rho)}} \tag{33}$$

Then solving for $\frac{\alpha_k - z_k}{\sigma_\varepsilon}$ yields:

$$\frac{\alpha_k - z}{\sigma_\varepsilon} = \frac{\theta_k\rho}{\sqrt{\rho^2 - \zeta_k^2(1-\rho)}} \tag{34}$$

B Even-odd split-half reliability coefficients

First the totality of the test items are divided into two sets, those based on the even-numbered test items 2, 4, 6, etc. and those based on the odd-numbered items 1, 3, 5, etc. Based on the two sets we can construct two raw scores for the same test-taker R^E_{sit} and R^O_{sit}. It is now assumed that both scores approximate the true raw score R^*_{sit}, each with its own measurement error component:

$$R^E_{sit} = R^*_{sit} + e_{R^E_{sit}} \tag{35}$$

$$R^O_{sit} = R^*_{sit} + e_{R^O_{sit}} \tag{36}$$

It is furthermore assumed that $E\left[e_{R^E_{sit}} \middle| R^*_{sit}\right] = E\left[e_{R^O_{sit}} \middle| R^*_{sit}\right] = 0$ and $E\left[e_{R^E_{sit}} \times e_{R^O_{sit}} \middle| R^*_{sit}\right] = 0$. Especially the latter (both measurement error components are uncorrelated with each other) is a strong assumption.

Based on these assumptions we can derive:

$$\begin{aligned} corr\left(R^E_{sit}, R^O_{sit}\right) &= \frac{C\left(R^*_{sit} + e_{R^E_{sit}}, R^*_{sit} + e_{R^E_{sit}}\right)}{SD\left(R^E_{sit}\right) SD\left(R^O_{sit}\right)} && (37)\\ &= \frac{V\left(R^*_{sit}\right)}{SD\left(R^E_{sit}\right) SD\left(R^O_{sit}\right)} && (38)\\ &\approx \frac{V\left(R^*_{sit}\right)}{V\left(R^E_{sit}\right)} = \rho_{R^E_{sit}} && (39) \end{aligned}$$

The correlation of the scores based on even items and odd items, is approximately equal (under assumptions) to the reliability of a test score based on half of the test items (either the even or the odd ones). Tests based on less items are more noisy, and less reliable. We should therefore "upscale" the estimate to represent the reliability of the totality of test items (even

and odd-numbered items combined).

$$
\begin{aligned}
V\left(R_{sit}\right) &= V\left(\frac{1}{2}\left(R_{sit}^{E}+R_{sit}^{O}\right)\right) && (40)\\
&= \frac{1}{4}\left[V\left(R_{sit}^{E}\right)+V\left(R_{sit}^{O}\right)+2C\left(R_{sit}^{E},R_{sit}^{O}\right)\right] && (41)\\
&= \frac{1}{4}\left[V\left(R_{sit}^{E}\right)+V\left(R_{sit}^{O}\right)+2V\left(R_{sit}^{*}\right)\right] && (42)\\
&\approx \frac{1}{2}\left[V\left(R_{sit}^{E}\right)+V\left(R_{sit}^{*}\right)\right] && (43)
\end{aligned}
$$

This can be rewritten:

$$
2V\left(R_{sit}\right)-V\left(R_{sit}^{*}\right) \approx V\left(R_{sit}^{E}\right) \tag{44}
$$

which can be rewritten into a formula for the the reliability of the complete test:

$$
\rho_{R_{sit}}=\frac{2\rho_{R_{sit}^{E}}}{1+\rho_{R_{sit}^{E}}} \tag{45}
$$

This formula is also known as the Spearman-Brown prophecy formula.

References

Ammermüeller, A., and P. Dolton (2006): "Pupil-Teacher Gender Interaction Effects on Scholastic Outcomes in England and the USA," *ZEW Discussion Paper No. 06-060*.

Andrabi, T., J. Das, A. I. Khwaja, and T. Zajonc (2011): "Do value-added estimates add value? Accounting for learning dynamics," *American Economic Journal: Applied Economics*, 3, 29–54.

Blundell, R., and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 87, 115–143.

Chang, M. C., S. Shaeffer, S. Al-Samarrai, A. B. Ragatz, J. De Ree, and R. Stevenson (2013): "Teacher Reform in Indonesia: The Role of Politics and Evidence in Policy Making," Directions in development, The World Bank.

Chu, J. H., P. Loyalka, J. Chu, Q. Qu, Y. Shi, and G. Li (2015): "The impact of teacher credentials on student achievement in China," *China Economic Review*, 36.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2010): "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects," *Journal of Human Resources*.

De Ree, J., S. Al-Samarrai, and S. Iskandar (2012): "Teacher Certification in Indonesia: a Doubling of Pay, or a Way to Improve Learning?," Policy Brief 73264, The World Bank.

De Ree, J., K. Muralidharan, M. Pradhan, and H. Rogers (2015): "Double for Nothing? Experimental Evidence on the Impact of an Unconditional Teacher Salary Increase on Student Performance in Indonesia," *NBER working paper w21806*.

Dee, T. S. (2005): "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?," *American Economic Review*, 95(2).

——— (2007): "Teachers and the Gender Gaps in Student Achievement," *Journal of Human Resources*.

Guarino, C. M., M. D. Reckase, and J. M. Wooldridge (2015): "Can Value-Added Measures of Teacher Performance Be Trusted?," *Education Finance and Policy*, 10(1), 117–156.

Hanushek, E. A., M. Piopiunik, and S. Wiederhold (2014): "The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance," *NBER working paper*, (20727).

Kingdon, G., and F. Teal (2010): "Teacher unions, Teacher Pay and Student Performance in India: A Pupil Fixed Effects," *Journal of Development Economics*.

Metzler, J., and L. Woessman (2012): "The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation," *Journal of development economics*, 99(2).

Novia, D. R. (2014): "Pendidikan Indonesia Gawat Darurat," in *Republika*.

OECD (2010): "PISA 2009 Results: What Students Know and Can Do (Volume 1)," `http://dx.doi.org/10.1787/9789264091450-en`, OECD.

Ragatz, A. (forthcoming): "The Importance of Teacher Knowledge in Student Learning Outcomes," *World Bank Policy Brief*.

Rothstein, J. (2007): "Do value-added models add value? Tracking, Fixed effectsm and causal inference," *unpublished*.

The Jakarta Globe (2012): "Indonesian Teachers Score Low on Competency Test: Big Surprise?," News paper article.

Todd, P. E., and K. I. Wolpin (2003): "On the specification and estimation of the production function for cognitive achievement," *The Economic Journal*, 113.

World Bank (2013): "Indonesia: Spending more or spending better: improving education financing in Indonesia," Report 73050-ID, World Bank.

——— (2015): "Indonesia: Teacher Certification and Beyond: an empirical evaluation of the teacher certification program and education quality improvements in Indonesia," Report 94019-ID, The World Bank.