Final report

International Growth Centre

Recruitment, motivation, and retention effects of teacher performance pay

Andrew Zeitlin
Clare Leaver
Owen Ozier
Pieter Serneels

March 2017

When citing this paper, please use the title and the following reference number:
S-38408-RWA-1

DIRECTED BY
LSE
UNIVERSITY OF OXFORD

FUNDED BY
UKaid
from the British people

Recruitment, motivation, and retention effects of teacher performance pay
Final Report to the International Growth Centre
15 March 2017

Andrew Zeitlin (Georgetown University), Clare Leaver (University of Oxford), Owen Ozier (The World Bank), and Pieter Serneels (University of East Anglia)

1. Problem statement and research objectives

Low primary school learning outcomes remain a central concern across several countries of sub-Saharan Africa. Focusing on Rwanda, impressive progress has been made in expanding access to education for all. In spite of this, the country is still facing a number of constraints related to quality of teaching and student learning outcomes.

Repetition of classes in primary schools increased from 12.5% in 2012 to 18.3% in 2013 while primary completion rates decreased from 75.6% in 2010 to 61.3% in 2014, and the transition from primary to secondary school fell from 93.8% in 2010 to 73.4% in 2013 (MINEDUC 2014).

As highlighted in the Education Sector Strategic Plan (ESSP) 2010-2015, a number of factors contribute to this slowdown in the improvement of education quality, ranging from lack of an adequate system to assess learning outcomes, to shortages of pedagogical materials and examples for practice, while financial restrictions that limit potential recruitment and retention of the best qualified teachers has also been identified as a key limitation to improving education quality (MINEDUC 2010)[1].

Having championed a leading study of pay for performance in the health sector in Rwanda with encouraging results (Basinga et al., 2011), the current ongoing study, conducted in collaboration with the Ministry of Education and Rwanda Education Board (REB), aims to evaluate not only the incentive effect (on effort) but also the selection effects (on skill and intrinsic motivation) of pay-for-performance (P4P) contracts. The study corresponds to the government's priorities of promoting a culture of results-based management within the Rwanda Public Service.

To this extent, the study is designed on a two-tiered experiment to answer three primary research questions:

1. Can P4P improve teacher performance, and so contribute to student learning gains?
2. How effective are P4P contracts at recruiting effective (skilled and intrinsically motivated) teachers, particularly in rural areas?
3. Do P4P contracts help to retain effective teachers?

A key step in answering these questions is to evaluate and identify the most effective contracts for teachers. If the incentives incorporated in teachers' contracts can both 1) attract more motivated and

[1] The concerns about the education sector's ability to attract and retain skilled and motivated workers is the subject of national debate (see, for instance, New Times 2013)

better teachers, and 2) induce teachers to expend their effort in a way that maximizes student learning gains, then the design of these civil servants' contracts becomes a central part of improving the quality of state-delivered services.

Understanding if P4P contracts can achieve improvements in state-delivered services aligns closely with the research objectives of the IGC theme of state effectiveness. The IGC's (2014) Evidence Paper on State Effectiveness, Growth, and Development, under the sub-theme of "Public Sector Organization", Azulai and coauthors identified two key research topics: "Recruitment strategies and civil servants' traits", and "Job attributes and civil servants' performance". This project joins together these two strategies for improving the quality of state-delivered services: it examines precisely how job attributes typically thought of as tools to improve the performance of existing civil servants may impacts the composition of those called to public service.

In Section 2 below, we detail the study design that has been deployed to answer these questions. Section 3 documents progress toward the implementation of this design.

Against the backdrop of the broader question of performance-based contracting for civil servants in general, and teachers in particular, there are important, practical questions regarding the design of measurement tools that can provide the basis of accountability metrics. Research by the Gates Foundation released in 2013 (Kane et al.) shows that classroom observation, in combination with student performance measures and student feedback, is a valid and useful tool to identify teacher effectiveness and enables top-down accountability in state schools. Measuring and incentivising teacher inputs, over and above teacher outputs, may also be attractive from a contract theoretical perspective, specifically when teacher control over student learning outcomes has its limits. The teacher and student evaluation instruments funded through this project, detailed in Sections 4 and 5 below, provide the core data to test the effectiveness of P4P contracts that incentivise both inputs and outputs, and to produce more robust top-down accountability within Rwanda's system of basic education.

Built in collaboration with Dr. David Johnson of the University of Oxford's Department of Education, the teacher evaluation instrument discussed in Section 4 combines teacher-input and surprise classroom monitoring visits to arrive at a more comprehensive picture of teacher effectiveness. Section 4 discusses the development of these teacher evaluation instruments and the initial findings that both lesson planning audits and classroom observations produce differentiation in teacher performance that can be used to assess teacher impact on student performance.

Towards those ends, Section 5 discusses the development of student assessment instruments for this project and the finding that these instruments are strongly correlated with state-administered measures of student performance (Primary Leaving Exams) across cohorts at the same school, indicating their validity as measures of student performance. With valid data in hand on teacher-inputs, classroom observations, and student performance, we will be able to bridge the question of the effectiveness of P4P contracts to improve the quality of state-delivered services.

2. Intervention design

This study is designed to test the selection and incentive effects of P4P using a two-tiered RCT focusing on actual recruitment of civil-service teachers for teaching jobs. Both tiers of this experiment are built around the comparison of two contracts.

The first of these is a P4P contract, which pays RWF 100,000 (approximately 15% of annual salary) to the top 20% of upper-primary teachers as measured by a composite of teacher input metrics (presence, preparation, and pedagogy) and a student learning outcome metric based on Barlevy and Neal's (2012) pay-for-percentile metric. The second is a *fixed-wage* contract that provides RWF 20,000 to all upper-primary teachers.

Following a pilot in 2015, the two-tiered experiment started in the run-up to the 2016 academic year. The first tier randomized advertised jobs to contracts. New primary teaching posts in five core curricular subjects across six districts were allocated to P4P or fixed-wage status (a total of 594 posts). The randomization took place at district-by-qualification level,[2] and was followed by an advertising campaign to increase awareness of the new posts and their associated contracts, including organization of 'job fairs' at Teacher Training Colleges.[3]

The second tier randomized schools to contracts. A school was included in the sample if the new post was filled and assigned to upper primary grades (total 164 schools, 313 recruits). Following a full baseline survey in February 2016, *s*ample schools were assigned to either P4P or fixed wage. All upper primary teachers within each school received the new contract.[4] Teacher payments, determined in P4P schools by multiple teacher-input observations as well as beginning- and end-of-year student assessments, are occurring in February and March 2017.

Our two-tiered experimental design distinguishes between the actual contract *experienced* by hired teachers and the contract *advertised* to potential applicants. This design allows estimation of impacts of: (a) experienced P4P on the student learning gains achieved by 1,600 teachers in 164 schools—thereby answering Research Question 1; (b) advertised P4P on application volumes and applicant characteristics among 1,881 applicants to positions across the 6 districts, and on characteristics and performance of actual hires among the 313 successfully placed applicants—answering Research Question 2; and (c) experienced P4P on retention rates of both new and incumbent teachers—answering Research Question 3. Our approach enables us to isolate the selection effect in Research Question 2 by estimating the impact

[2] Since few teachers apply for jobs in multiple districts and prospective teachers are only eligible for posts if they have the relevant subject qualification, we view district-qualification pairs as distinct labour markets. To illustrate two outcomes of the assignment, in the district of Kayonza new TML-eligible posts (English and Kinyarwanda) were assigned to P4P and new TSM-eligible posts (Maths and Science) were assigned to fixed-wage contracts. The reverse was true in the district of Kirehe.

[3] Extensive data on *potential* applicants were collected at these job fairs. Advertisements also took place over the radio, in person at District Education Offices, and through dissemination of printed materials in district capitals of the study districts.

[4] At the individual applicant level, this amounted to re-randomization and hence a change to the initial assignment for some new recruits. All new recruits were paid a 'retention bonus' of 80,000 RWF to ensure that new contracts strictly dominated those advertised at the first tier.

of *advertised* P4P, holding constant *experienced* P4P.[5] Using the second-tier re-randomization, we can therefore pinpoint the mechanism behind the impact of P4P.

3. Implementation progress

Following the advertisement and recruitment process, 164 schools were successfully enrolled in the study, covering 313 new recruits and approximately 1,600 teachers in total. Schools were randomized into treatment arms, with the P4P treatment introduced in March of 2016. In P4P schools, two visits were undertaken to assess teachers' *inputs* into the classroom. Learning outcomes were measured at endline in the full set of schools in order to determine eligibility for awards under P4P.

Following the collection of teacher bank details in October 2016 during endline survey, a brief phone survey was carried out in January 2017 to confirm bank details and teacher status as of October 2016 in both P4P and fixed-wage schools. Based on this verification process, fixed-wage teacher payments of RWF 20,000 were made via electronic bank transfer in early March 2017 to 802 upper-primary teachers. It is expected that P4P teacher payments of RWF 100,000 and retention bonus payments of RWF 80,000 will be issued to qualifying teachers in mid-March 2017.

4. Teachers' classroom activity under Payment for Performance

P4P incentives are based on an evaluation of teacher performance across two domains: student performance and teacher inputs. Teacher inputs are evaluated using a combination of assessments of written lesson plans and surprise classroom visits, as well as teacher presence. Here we focus on the first two components. These were measured twice during 2016 – once shortly after an initial teacher training session, and once near the end of the academic year. These lesson plan and monitoring data are used to calculate a single teacher-input score for teachers in the P4P arm.[6] This score, along with a teacher's student performance scores, and the teacher presence at time of inspection, serve as the percentile basis for teacher bonus payments, and serve as a primary outcome in this research study. This section provides a detailed description of how teacher preparation and pedagogy inputs are assessed, through evaluation of their lesson plan and classroom observation.

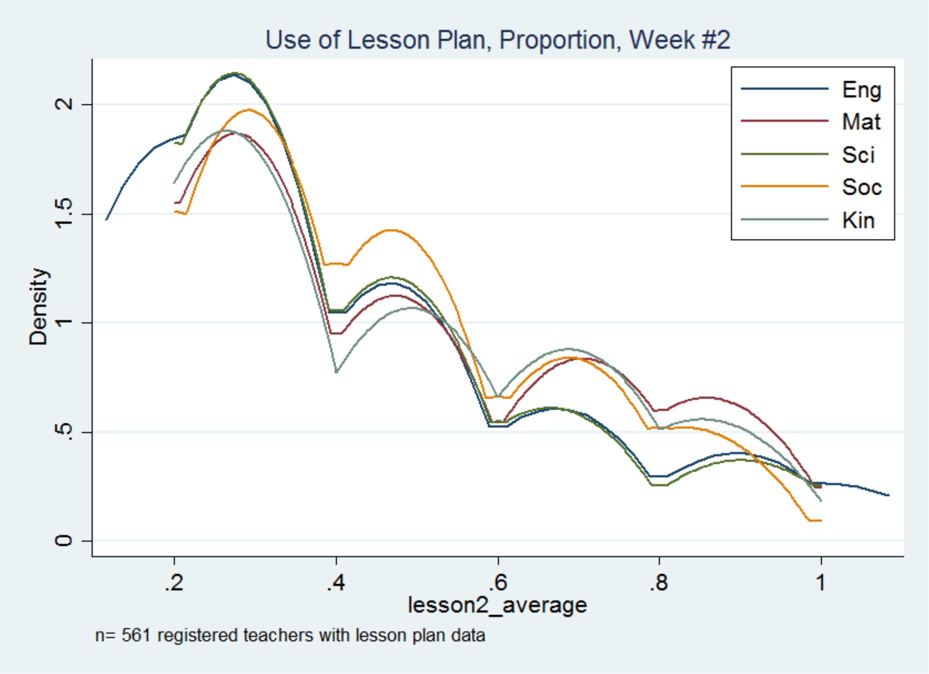
Teacher preparation: lesson plan

Figure 1 below shows the distribution of lesson plan completeness for those teachers who were recorded as having used the lesson plan at all, by school subject. Even for teachers that do use lesson plans, its use is incomplete – 43% of teachers filled out a lesson plan just once in the week of auditing, while the average completeness is 2.11 days per week. Mathematics teachers are most likely to regularly use lesson plans, while Kinyarwanda teachers are the least likely. Because usage is distributed across different levels of usage, the metric does well to differentiate teacher performance along lesson plan use, a fundamental component of teacher quality.

[5] Note that the school-level randomization allows us to compare the performance of teachers who applied under different contracts, but who ended up in the same contractual assignment. The *total effect* of P4P is identified here by the comparison between those who applied under P4P and whose school was assigned to P4P with those who applied under fixed wages and whose school was assigned to the fixed-wage arm.

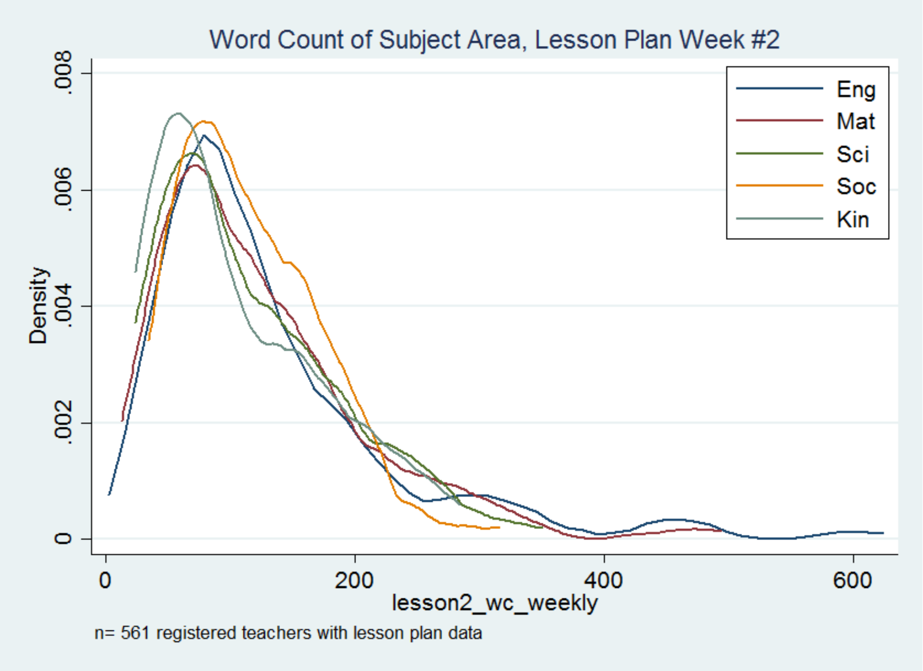
[6] The lesson plan and classroom observation data are not collected for teachers who are receiving fixed wage increases

Figure 1: Proportion of teachers using lesson plans partially/fully, by subject



We can also observe lesson plan word count, another rough metric of lesson plan quality. For teachers with lesson plans, the mean weekly word count is 84 words. This number is higher for Social Studies and English teachers than for Mathematics, Science, or Kinyarwanda teachers. Word count is not well suited to be used as an objective measure of teacher quality because it is not clear that a higher word count is associated with higher quality lesson plans or better teacher performance.

Figure 2: Weekly lesson plan word count by teacher, by subject



Teacher Pedagogy: Classroom Monitoring

In addition to lesson plan grading, teachers were subject to unannounced classroom monitoring visits. Enumerators evaluated teaching pedagogy and quality in four domains: Lesson Objective, Teaching Quality, Assessment, and Student Engagement. Each domain consisted of a set of 3-9 component activities that are thought to contribute to teaching quality. Enumerators then assigned each teacher a subjective quality score of 0 to 3 for each domain. Seventy-six percent of registered teachers were observed during the first round of classroom monitoring, while 75% were observed during the second round.

Tables 1-4 below show the proportion of teachers engaging in each pedagogic activity in a given domain, along with the associated impact on subjective score. Since there is a subjective component to the aggregate score, these estimated impacts show which pedagogic activities are considered most important to the enumerators when judging overall quality.

Table 1 shows the pedagogic components of the lesson objective teaching domain. Teachers are marked for their use of lesson plans, communication of those plans to their students, and adherence to the lesson plan throughout the lesson. Nearly all teachers are recorded as having used some sort of lesson plan (92%), while slightly less actually used the that plan in their teaching (87%). Enumerators judged having a lesson plan and using it as the most important factors in a teacher's adherence to lesson objectives, while it was less important for teachers to explicitly explain those plans to students at the beginning of each lesson. Figure 3a shows the distribution of aggregate Lesson Objectives scores by school subject taught.

Table 1: Components of Lesson Objectives classroom monitoring

Teacher Activity	% Exhibiting Behavior	Associated Increase in Score
Q1. A lesson planning form is present with a written lesson objective.	0.92	1.86
Q2. Teacher communicates lesson objective to students early in the lesson.	0.69	0.94
Q3. Teacher teaches the same lesson objective as is written in the lesson planning.	0.87	1.45

Proportion of teachers exhibiting behavior at any frequency. Associated increase in score determined by bivariate regression

Figure 3a: Aggregate scores for Lesson Objectives classroom monitoring

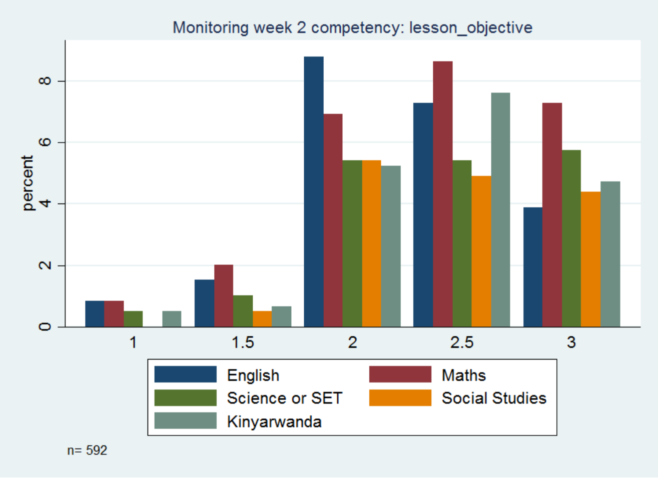


Table 2 shows the component pedagogic activities for a general Teaching quality measure. This domain covers teacher preparation and teaching style, as well as student involvement. Enumerators judge the most important aspects of teaching quality to be whether a teacher reviews previously taught material, uses prepared charts and diagrams, or uses chants, songs, or claps to advance learning objectives.

Table 2: Components of Teaching classroom monitoring

Teacher Activity	% Exhibiting Behavior	Associated Increase in Score
Q4. Teacher reviews previously taught material.	0.72	0.30
Q5. Teacher lectures.	0.64	0.00
Q6. Students work individually.	0.72	0.14
Q7. Students work in pairs.	0.20	0.21
Q8. Students work in groups.	0.70	0.31
Q9. Teacher writes on the board.	0.98	0.21
Q10. Teacher draws on the board.	0.33	0.19
Q11. Teacher uses prepared charts or diagrams.	0.31	0.33
Q12. Teachers uses chants, songs, or claps.	0.64	0.34

Proportion of teachers exhibiting behavior at any frequency. Associated increase in score determined by bivariate regression

Figure 3b: Aggregate scores for Teaching classroom monitoring

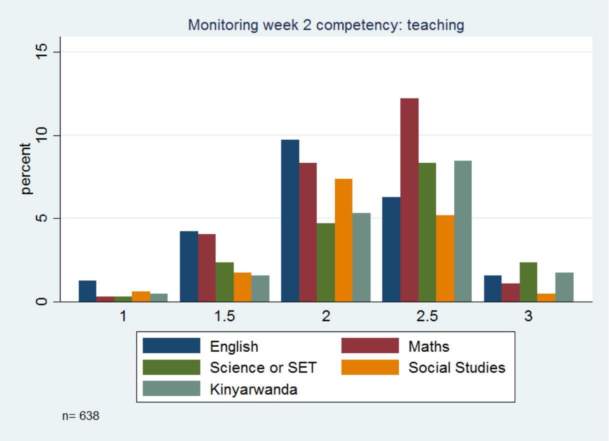


Table 3 shows the component pedagogic activities for the Assessment domain of teaching. This domain measures student involvement through solicited and forced participation, as well as teachers' use of homework and other exercises. Enumerators judge having students answer questions on the board or in notebooks and teacher review as the most important signs of teaching quality in classroom assessment.

Table 3: Components of Assessment classroom monitoring

Teacher Activity	% Exhibiting Behavior	Associated Increase in Score
Q13. Students answer questions on the board.	0.81	0.36
Q14. Students answer questions in their notebooks	0.92	0.35
Q15. Teacher calls students randomly, i.e. 'cold call'	0.64	0.24
Q16. Teacher calls students who raise their hand.	0.92	0.12
Q17. Teacher checks students' exercises for accuracy.	0.92	0.39
Q18. Teacher gives homework at the end of the lesson.	0.90	0.02

Proportion of teachers exhibiting behavior at any frequency. Associated increase in score determined by bivariate regression

Figure 3c: Aggregate scores for Assessment classroom monitoring

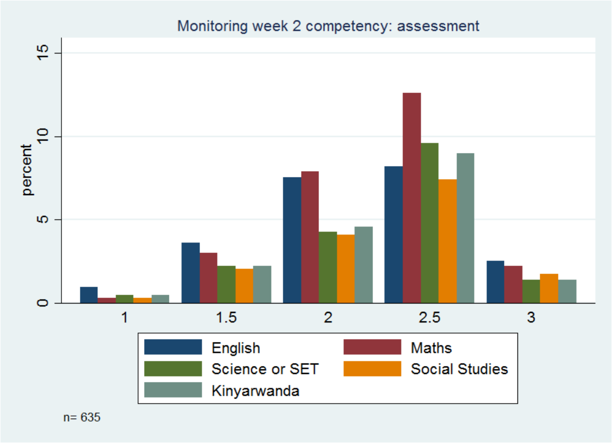


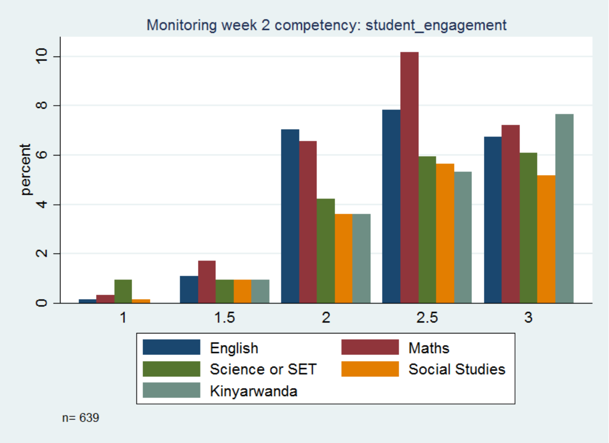
Table 4 shows the components of teacher quality in the Student Engagement domain. This domain measures student activities including using notebooks, following instructions, and classroom engagement. The prevalence of all three activities is very high, with more than 90% of teachers meeting each criteria. Enumerators judge continual engagement as the most important factor.

Table 4: Components of Student Engagement classroom monitoring

Teacher Activity	% Exhibiting Behavior	Associated Increase in Score
Q19. Students use a notebook to copy notes or exercises.	0.90	0.18
Q20. Students follow teacher instructions for classroom behavior.	0.95	0.39
Q21. Students are engaged in an activity throughout the lesson	0.94	0.40

Proportion of teachers exhibiting behavior at any frequency. Associated increase in score determined by bivariate regression

Figure 3d: Aggregate scores for Student Engagement classroom monitoring



Out of 3 possible points, teachers averaged a score of 2.20 on Lesson Objective, 2.16 on Teaching Quality, 2.23 on Assessment, and 2.45 on Student Engagement in the second round of classroom observation data collection. Each of these scores does show improvement from the first round of collection, which was conducted just after teacher training sessions. Classroom monitoring scores are distributed across the scoring spectrum, thus allowing these evaluations to contribute to a scoring metric capable of differentiating teacher quality.

5. Measures of Student Learning

Assessment tools were developed for the dual purpose of measuring teacher effectiveness and assessing learning outcomes in Primary 4, 5, and 6 at the school, grade, and subject level. What was needed was an instrument that could be administered efficiently and at scale, while capturing key features of the curriculum. Dr. David Johnson, an educationalist from the University of Oxford, led the team's work in this regard.

For each grade, these tools cover five core academic subjects – English, Kinyarwanda, Mathematics, Science, and Social Studies – in a brief instrument that is administered to a classroom of up to 40 pupils in under one hour, or 12 minutes on average per subject. Questions were read aloud by proctors, while pupils had paper copies of each question, on which they wrote their answers.

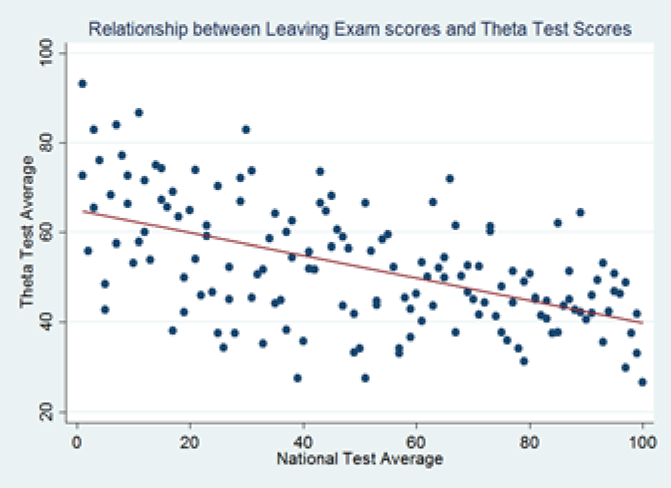
Clearly, there are limits to the number of curricular competencies that can be covered in such a short time. The essential goal of the exercise was not to aim for complete coverage, therefore, but rather to provide an externally valid snapshot of learning levels.

One measure of whether the snapshot provide by the STARS assessment tools captures a meaningful measure of student learning outcomes is whether the outcomes of the STARS assessments are correlated with official measures of learning outcomes. To assess whether this is the case, we compare results in a set of school-level average scores on the Primary Leaving Exam, which were kindly shared with our research team by REB for purposes of sampling schools in the six districts of the second phase of our activities (in practice we are only in a fraction of sampled schools from these districts). To provide a measure of the validity of the STARS instruments, we estimate the strength of the relationship between school-level average scores on the endline assessment undertaken at the end of the 2016 school year with PLE scores from the end of the 2015 academic year.

It should be noted that this is a stringent test of the quality of our instruments because of timing differences. Clearly, students in P6 at the end of 2016 are not the same individuals who sat the PLE in 2015. If P6-level scores, from a different cohort, are predictive of P6 PLE exam outcomes, then this provides strong support for the predictive validity of our assessments.

The extent of this association is documented in Figure 4. Here, we derive a measure of learning performance on the STARS assessment ('theta'), using a two-parameter Item Response Theory model. Its association with assessment outcomes is strong: the STARS assessments explain 37.1 percent of the variation in PLE scores even within a district; cross-district explanatory power, school-level explanatory power when the test-takers are assessed by both instruments, and explanatory power at the pupil are all likely to be even higher.

Figure 4: Correlation between PLE and STARS assessment outcomes

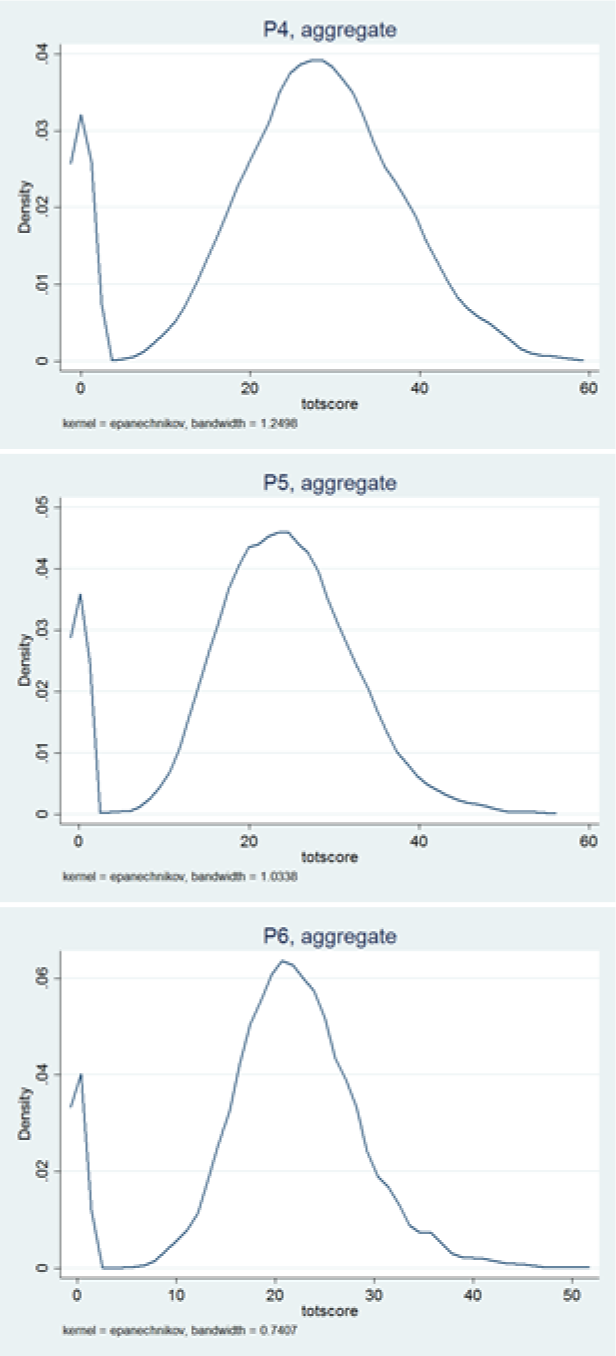


The fact that even across grade cohorts there remains a strong association between performance on the STARS assessments and performance on the PLE suggests that these assessments can be useful as a diagnostic tool.

Another key question for the quality of the assessment instruments is whether they could effectively discriminate variation in achievement levels for both high-performing and lower-performing pupils. If a test is too easy, many students will achieve 100%, and the only variation detected by the instrument will be among lower-performing pupils. On the other hand, if an assessment instrument is too difficult then many students will achieve very lower marks, and the only variation in scores that remains will be driven by a few high-performing pupils. These phenomena are known as the ‘ceiling effect’ and ‘floor effect’.

Based on the distribution of pupil’s scores, the STARS instruments have no ceiling effects and floor effects are mild. The majority of pupils’ scores are approximately normally distributed with a mean response close to 50% of questions answered correctly. Results are shown in figure 5, below.

Figure 5: Raw score distributions for pupils, by grade level

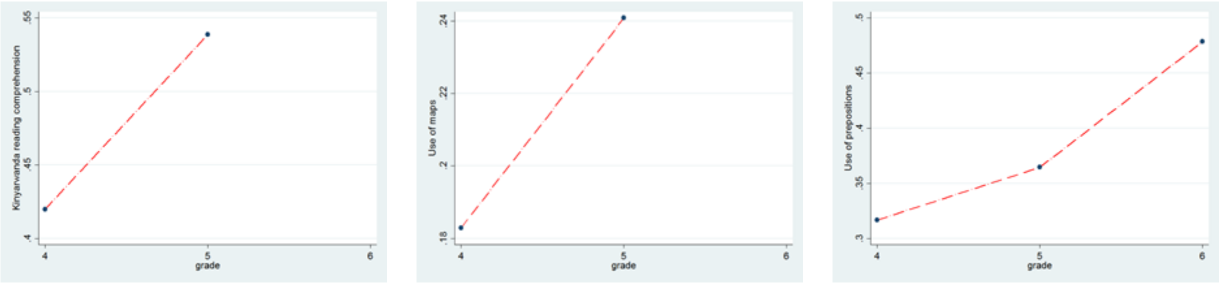


In addition to providing a picture of performance at any one point in time, a further value of these assessments is that they allow tracking learning outcomes across grades by comparing different cohorts or (if the same school is assessed in consecutive years) by comparing the same student's performance at different points in time.

In a short assessment instrument such as the STARS measures, scope for doing so is limited. With just over 10 minutes per subject, repeating many questions across grades would substantially detract from the assessment's ability to detect performance on grade-level material. Nonetheless, a small set of comparable questions were built into the STARS assessment. Figures 6a, 6b, and 6c summarize outcomes for some of these items in the representative first-year sample of schools and pupils. For demonstration purposes we focus on key competencies in Kinyarwanda, English, and Social Studies.

Two key lessons are observable from such analyses. First, grade-level performance is not as strong as would be hoped. Second, and more optimistically, for each of these measures, continued gains in student performance are observed across years.

Figure 6: Evolution of specific competencies across grades



6a. Kinyarwanda Reading Comprehension

6b. Social studies competencies (map use)

6c. English competencies (preposition use)

6. References

Azulai, M., Bandiera, O., Blum, F., Kleven, H., La Ferrara, E., Padro, G., Tejada, C.M., 2014. IGC Evidence Paper – State Effectiveness, Growth, and Development.

Basinga P., P. J. Gertler, A. Binagwaho, A. L.B. Soucat, J. R. Sturdy, C. M.J. Vermeersch, 2011, Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation, *Lancet*. 377(9775), pp. 1421-8.

Barlevy, G. D. Neal, 2012, Pay for Percentile. *American Economic Review,* Vol. 102, No. 5 (2012), pp. 1805–1831.

Kane, T.J., McCaffrey, D.F., Miller, T., Staiger, D., 2013, Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Seattle, WA: Bill & Melinda Gates Foundation.

MINEDUC (2010) "Education Sector Strategic Plan 2010 – 2015". July-Kigali

MINEDUC, 2014, The 2014 Education Statistical Yearbook, Rwanda Ministry of Education, Kigali.

The International Growth Centre (IGC) aims to promote sustainable growth in developing countries by providing demand-led policy advice based on frontier research.

Find out more about
our work on our website
www.theigc.org

For media or communications
enquiries, please contact
mail@theigc.org

Subscribe to our newsletter
and topic updates
www.theigc.org/newsletter

Follow us on Twitter
@the_igc

Contact us
International Growth Centre,
London School of Economic
and Political Science,
Houghton Street,
London WC2A 2AE

IGC
International
Growth Centre

DIRECTED BY
LSE
UNIVERSITY OF
OXFORD

FUNDED BY
UKaid
from the British people

Designed by soapbox.co.uk