Teacher Development Programme

Endline Evaluation (Volume II)

Stuart Cameron, Gunilla Pettersson Gelander, Jana Harb, Mehjabeen Jagmag, Paul Jasper, Alexandra Doyle, Hanna Laufer, Madhumitha Hebbar, Lucia Barbone

May 2018

EDOREN
Education Data, Research & Evaluation in Nigeria

No 2, 16 Mafemi Crescent
Utako
Abuja, Nigeria

Tel +234 810 727 8718
Tel +234 817 667 8243
Email info@edoren.org
Website www.edoren.org

Table of contents

List of figures, tables and boxes

List of abbreviations

ATET	Average treatment effect on the treated
BL	Baseline
CAPI	Computer-assisted personal interview
DFID	Department for International Development (UK)
DID	Difference-in-differences
EDOREN	Education Data, Research and Evaluation in Nigeria
EQUALS	Evaluation Quality Assurance and Learning Service (of DFID)
ESSPIN	Education Sector Support Programme in Nigeria
FE	Fixed effects
FMOE	Federal Ministry of Education
GEP3	Girls' Education Project 3
INSET	In-service training
IRT	Item response theory
ITC	Intertemporal correlation
ITT	Intention-to-treat
IV	Instrumental variable
JSS	Junior secondary school
LASSO	Least absolute shrinkage and selection operator
LATE	Local average treatment effect
LGA	Local Government Authority
LGEA	Local Government Education Authority
MDE	Minimum detectable effect
ML	Machine learning
MSC	Most significant change
NCE	National Certificate of Education
NEDS	National Education Data Survey
NERDC	National Education Resource Development Council
NGN	Nigerian Naira
OECD-DAC	Organisation for Economic Co-operation and Development Development Assistance Committee
OLS	Ordinary least squares
OPM	Oxford Policy Management
P1–P6	Primary 1–6
PSU	Primary sampling unit
RANA	Reading and Numeracy Activity
RCT	Randomised control trial
RE	Random effects
SBMC	School-based management committee
SD	Secure digital (memory card used in mobile phones)
SLM	School leadership and management
SSV	School support visit
SUBEB	State Universal Basic Education Board
TDNA	Teacher development needs assessment
TDP	Teacher Development Programme
TDT	Teacher Development Team
TF	Teacher facilitator
TOC	Theory of change
TOR	Terms of reference
UBEC	Universal Basic Education Council
UNICEF	United Nations Children's Fund

1 Introduction

The Teacher Development Programme (TDP) evaluation design was set out in an evaluation framework (Education Data, Research and Evaluation in Nigeria (EDOREN), 2014) at baseline. A number of developments made it necessary to look again at the original evaluation design prior to the endline research. These included: the decision not to have a midline survey in 2015, as originally planned; the need to check the extent to which the programme had been rolled out as intended and to deal with potential issues, such as contamination and attrition; and changes in the programme's design and theory of change (TOC) as it adapted to new learning. The EDOREN evaluation team therefore created an endline plan (Cameron *et al*., 2017), in consultation with staff of the UK Department for International Development (DFID) and TDP. The endline plan also drew on the results of initial research, including an implementation review (Durrani *et al*., 2018) and a validation telephone survey (Cameron and Pettersson, 2017). The plan was reviewed by DFID's Evaluation Quality Assurance and Learning Service (EQUALS) in November 2017, and revised in response to the EQUALS reviewer's comments in January 2018. The endline plan effectively constitutes the agreed terms of reference for the endline evaluation, and should be consulted where more detail is needed on the background of this evaluation.

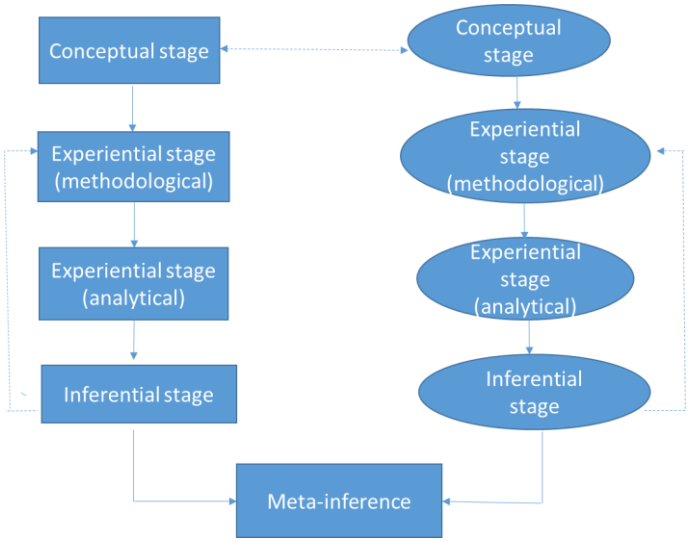
This volume complements the Endline Evaluation Report Volume I – which sets out the main results and recommendations – and the earlier methodological and background documents referenced above, to provide detailed descriptions of the methodology used in the TDP impact evaluation endline. Chapter 2 describes the overall mixed methods approach employed in the evaluation. Chapter 3 describes the quantitative research design, including the quantitative impact evaluation approach and the sampling strategy. Chapter 4 describes how the quantitative data were collected through a survey. Chapter 5 describes both the approach and data collection for the qualitative research. Chapter 6 describes how the pupil tests were designed and analysed. Chapter 7 describes how the classroom observation instrument was updated for the endline and how the data from this were analysed. Chapter 8 describes the development of a teacher motivation scale. Chapter 9 describes issues around permits for the research, consent, confidentiality, and sharing of data. Chapter 10 describes the processes around stakeholder engagement and impact. Finally, a set of annexes present the impact evaluation matrix around which this endline evaluation was designed, further detailed statistical tables, and the EQUALS matrix for this evaluation.

2 Mixed methods approach

The impact evaluation endline research employs a parallel mixed methods design, where the design is mixed at the methodological (conceptual) stage of the study and integrated at the data interpretation (inferential) stage, as illustrated in Figure 2.1 (Teddlie and Tashakkori, 2006; Leech and Onwuegbuzie, 2009).

For the endline study the qualitative and quantitative methods answer different aspects of the same question. The purpose of mixing qualitative and quantitative methods is to provide robustness and depth to the research findings; increase understanding of how changes occurred (or failed to occur) as expected; and allow each method to enquire into areas that cannot be investigated with the other method.

Figure 2.1: Parallel mixed methods design at the endline



2.1 Design of impact evaluation endline study

The design for the impact evaluation is structured around the programme TOC for the in-service training (INSET) component and the TDP logframe, which were developed during the inception of the programme.

The quantitative research is used to measure programme impact and changes in teacher effectiveness and provide some contextual analysis at the school level. The key features of the quantitative design were determined at the baseline and there was limited scope to change this design at the endline. However, the quantitative endline instruments were revised to incorporate learning from the previous qualitative midline and quantitative baseline research.

The qualitative research explores perceptions of how changes occurred, and factors that facilitated or hindered the programme's achievement of its intended impact and outcomes. It also addresses key questions raised during the baseline and midline (formative) research rounds. The selection of schools for the qualitative research was nested within the quantitative sample of treatment schools and informed by quantitative baseline data, to better explain changes since the baseline (see Chapter 4).

During the implementation review, quantitative data were analysed in relation to the actual and planned programme implementation, in terms of quantity and timing. The qualitative research explores reasons for observed departures from the implementation plan in order to answer questions relating to programme efficiency.

To assess programme sustainability qualitative methods were used to measure perceptions of sustainability held by key stakeholders.

2.2 Timing of the mixed methods endline research

Due to programmatic and timeline constraints, the main quantitative survey and qualitative research had to be conducted in parallel. Consequently, each of these two components were not able to inform the other's design or provide detail to the other's findings.

However, the implementation review, which also involved both quantitative and qualitative parts, and the validation survey, were both conducted before the main study, and able to inform both quantitative and qualitative parts of the main study. This allowed information on changes to the TDP model, progress in programme implementation, and the movement of head teachers, teachers and students, to feed into the main study.

2.3 Mixing the research teams

To enable mixing of the qualitative and quantitative research in practice it is important that the qualitative and quantitative research teams are as integrated as possible. For this reason, the draft *Impact Evaluation Endline Plan* was developed by the qualitative and quantitative analysts working together to assess the relative strengths of different research methods in regard to answering the different evaluation questions at endline.

During the impact evaluation team workshop in Oxford in June 2017 the analysts discussed, revised and agreed on the draft *Impact Evaluation Endline Plan*, the impact evaluation workplan outlining the sequencing of different activities, and the manner of cooperation and communication between qualitative and quantitative analysts for the remainder of the impact evaluation to ensure integration of research findings.

Once all data had been made available to the analysis teams (in November 2017–January 2018), the teams conducted preliminary analyses separately. An analysis workshop was then held in Oxford in March 2018 in order to bring together insights from the quantitative and qualitative analyses and ensure that the report would properly address both. A division of labour for report-writing was also agreed whereby some sections would be led by a quantitative researcher (e.g. TDP's impact on pupil learning) and others by a qualitative researcher (e.g. programme sustainability), in each case staying in close touch with, and incorporating inputs or insights from, other team members.

3 Quantitative research design

3.1 Quantitative methods: impact assessment design

3.1.1 Original design

Table 3.1 below lists the indicators for which this evaluation aimed to estimate the causal impact that TDP has had. That is, the evaluation team used the data collected via the school survey at endline and baseline in treatment and control areas to assess whether TDP has affected pupil learning outcomes in English, maths, and science, and intermediate outcomes related to teacher effectiveness in treatment schools.

Table 3.1: TDP Quantitative impact indicators

Evaluation matrix reference	Evaluation indicator	TOC level
Im-1	• **Indicator Im-1a:** Percentage change in mean scores in English, maths, and science for pupils in TDP schools (pupils tested in Primary (P3) at baseline and in Primary 6 (P6) at endline) • **Indictor Im-1b:** Change in the proportion of pupils in the bottom and top performance bands in English, maths, and science , respectively (pupils tested in P3 at baseline and in P6 at endline)	Final impact
Effe-1, Effe-2	• **Indicator Effe-1:** Percentage change in time teacher involves pupils in positive interaction during lesson (% of total lesson time) • **Indicator Effe-2:** Percentage change in average daily absence from school (% of teachers)	Intermediate impact

The original design of the impact evaluation component of this evaluation specified that the preferred strategy to estimate the impact of TDP on the indicators above was to exploit the fact that TDP treatment was assigned randomly across schools, and having two data points (baseline and endline) for each unit of observation. In such a setting, comparison of changes over time in the impact indicators across the treatment and control groups can be used to estimate the impact of TDP. Average differences in such changes can be causally attributed to TDP, because TDP treatment is randomly allocated across schools. Technically speaking, this randomisation process creates a counterfactual group (the control schools) to the treatment group in order to identify the effect of TDP.

It is important to emphasise here that the method this impact evaluation adopted to create a counterfactual was to *randomly* assign *clusters* of schools to either the treatment or the control group. This was because, by design, TDP combined schools into groups of 12 schools in each Local Government Authority (LGA), based on their geographical proximity to each other. Three teachers and the head teacher were, then, based on pre-defined criteria, selected within each treatment school (in the cluster) to participate in TDP, while no teacher from schools in the control clusters received TDP's training.

On the basis of the assumption that the random assignment of treatment worked as intended, any potential difference in teacher effectiveness and pupil learning in the treatment and control group over time can be attributed to TDP. The TDP baseline report provided extensive balancing checks to give evidence that this was indeed the case.

Brief summary of the randomisation process

The randomisation process was described in detail in the TDP baseline report (De *et al*., 2016a). This section provides a brief summary for purposes of completeness.

For the selection of the treatment and control groups the EDOREN evaluation team recommended that within each of the 14 LGAs in each of the three TDP Phase 1 states TDP should select two clusters of 12 schools each, based on guidelines provided. To prevent bias in the selection of teachers, the identification of teachers who would participate in TDP had to occur before the treatment and control school clusters were selected. Thus, within each LGA, schools in both treatment and control groups were required to select four teachers each (before knowing if the school would participate in TDP or not), who would potentially benefit from TDP. In every school (treatment and control) the programme always selected the head teacher (whether they teach or not), as well as three teachers.

The EDOREN evaluation team, after receiving lists of school clusters and teachers from TDP, assigned clusters in each LGA to either the treatment or the control group using a simple random-number generator.

As described in Section 3.2, a sample of four schools per cluster (both treatment and control clusters) were visited for data collection purposes.

3.1.2 Threats to the original design

As described in the TDP endline plan (Cameron *et al*., 2017), implementing a simple comparison of changes over time to identify programme impact would have required the design conditions from baseline to be retained at endline. In particular, it would have relied on design compliance – TDP training being implemented according to the study design in treatment schools only – and panel data collection to be satisfactory, with low levels of non-differential attrition. The following sections present analysis on both of these issues and how they affected the final analytical choices taken by the impact evaluation team.

Contamination and non-compliance analysis

The risk of sample contamination and non-compliance with treatment assignment in this impact evaluation arises from two main sources:

- contamination of the control schools and non-compliance with implementation in the treatment schools by TDP itself; and
- contamination of the treatment and control schools by other education interventions.

As described above, this impact evaluation follows an experimental design which entails the random assignment of schools to treatment or control status (participation or non-participation in TDP).

The first risk is that if implementation is non-compliant with this design – that is, some control schools receive TDP training (contamination by TDP) and/or some treatment schools do not receive TDP training (non-compliance) – this may affect the impact estimates.

The second risk is contamination of treatment and control schools by other education interventions, which may also affect the impact estimates. A set of donor-funded programmes (in addition to TDP) have been working in Jigawa, Katsina, and Zamfara during the evaluation period from October 2014 to October 2017 to improve the quality of primary education. There has also been State Universal Basic Education Board-(SUBEB)-led training. Although these interventions differ, they share the overall common objective of improving pupil learning, and include some types of teacher training aimed at improving teaching skills. The

main relevant programmes are listed in Table 3.2. For more details on these programmes see the 2017 TDP impact evaluation endline plan.

Table 3.2: Other main education programmes operating during the evaluation period

Programme	Objective	Main activities	States covered by the education intervention	Timing
Education Sector Support Programme in Nigeria (ESSPIN)	• Improve learning outcomes for children of basic education school age in the programme's six states. • Increase access to and completion of basic education for Nigerian children of primary school age, especially girls.	• In-service teacher training on Primary 1 to Primary 3 literacy and numeracy. • School visits. • Head teacher leadership training.	Jigawa	2009/10 to 2016
Reading and Numeracy Activity (RANA), which is part of Girls' Education Project Phase 3 (GEP3) output 2	• Improve the early learning skills of children in P1–P3 in the mother tongue, while also preparing children to learn with English as a language of instruction by the time they transition to P4.	• In-service training on Hausa-based literacy instruction, time on task, lesson planning, and effective preparation and use of materials for P1–P3 teachers and head teachers. • Provision of teaching and learning materials in Hausa for P1–P 3. • School support visits (SSV). • Head teacher capacity development training.	Katsina and Zamfara	February/April 2016 to February/April 2019
Jolly Phonics	• Improve early grade literacy skills.	• In-service training for P1 and P2 teachers to teach the 42 major sounds of the English language, how to form/write these sounds, how to blend the sounds together to read words, how to segment the sounds in words to write them, and irregular 'tricky' words that do not fit within this sound system. • Provision of teaching and learning materials for P1 and P2 teachers and pupils. • To sensitise head teachers on the use of the Jolly Phonics programme and equip them with resources to support teachers.	Jigawa, Katsina, and Zamfara	Jigawa training: P1 teachers January to March 2015 and P2 teachers October 2016. Katsina training: P1 teachers October 2018. Zamfara training: P1 teachers November 2013 and refresher training January 2015, P2 teachers January 2015.

		• To provide a Jolly Phonics Monitoring Team and SUBEB officials with the knowledge, skills, and resources to support teachers and monitor implementation of the programme.		
SUBEB-led training loosely based on TDP	• Improve teacher effectiveness and pupil learning levels.	• Jigawa: teacher training using elements of the TDP model[2], selected teachers only received some of the TDP training modules, not all the content was covered, and there are no SSVs. • Zamfara: teacher training based on the TDP model, provision of all TDP materials except the Trainer in the Pocket and secure digital (SD) cards, and SSVs.	Jigawa, (Katsina[1]) and Zamfara	Jigawa: July/August 2017. Zamfara: November 2016 to July 2017.

Note: (1) There is no overlap between the sample LGAs and the LGAs in Katsina where this SUBEB-led training is planned to be implemented. (2) For a description of the TDP training model see Chapter 3 in Volume I.

Source: ESSPIN (2016) for ESSPIN; TDP Impact Evaluation Implementation Review fieldwork July 2017 for SUBEB-led teacher training; EDOREN (2016d) for GEP3 and RANA; and Jolly Phonics (2015) for Jolly Phonics.

From an impact evaluation perspective, a differential implementation of non-TDP education interventions in the sampled treatment and control schools may lead to an overestimation or underestimation of the impact of TDP if it is not taken into account appropriately. For instance, if RANA is implemented in more control schools than treatment schools, and the impact indicators in control schools consequently improve, the evaluation would underestimate the impact of TDP. Conversely, if more treatment schools received RANA than control schools, with related improvements in outcomes, the evaluation would overestimate the impact of TDP and partly assign RANA's effects to TDP. Both of these situations would hence result in endogeneity – or selection bias – affecting the estimation of TDP's effects.

The objective of this section is to examine the extent of contamination by TDP in control schools and non-compliance in treatment schools, and differential contamination by non-TDP education interventions. The section also summarises the approaches that this evaluation took to address the resulting risk of selection bias.

How did contamination and non-compliance affect TDP treatment and control schools?

Given the sources of sample contamination outlined above, the evaluation team collected data from the programme implementers and through the endline quantitative survey to assess the extent of contamination and non-compliance. Although the endline plan proposed the augmentation of the survey instruments to collect detailed self-reported and head teacher-reported information on training received, fieldwork implementation indicated that head teachers and teachers were unable to distinguish reliably between the different training programmes, given similarities between the different training programmes. In particular, teachers and head teachers in control schools had difficulties differentiating between the SUBEB-led training loosely based on TDP and the actual TDP training as these used the same materials and often the same master trainers.

The evaluation team therefore used programme records data to identify and control for sample contamination, as well as TDP non-compliance with treatment assignment. The programme records data contain school-level, and in most cases also teacher-level data, on the training provided.

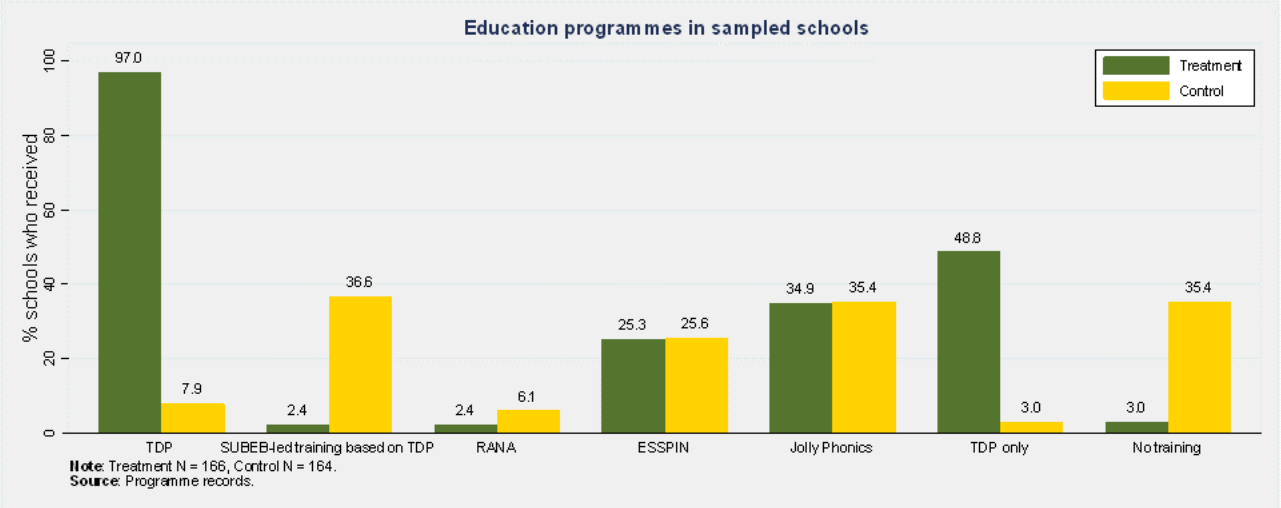
For the purposes of the analysis, schools were classified as contaminated by non-TDP education interventions if at least one teacher in a given school received training by an education intervention other than TDP. Control schools were classified as contaminated by TDP itself if at least one teacher in a given school received TDP training. Treatment schools were classified as non-compliant with treatment assignment if no teacher in a given school received TDP training. Figure 3.1 presents the results on sample contamination and non-compliance.

Implementation compliance of TDP in the **treatment group** is high, as 97% of the schools have received TDP training. Given that this indicator is defined at the school level, it is important to note here that this does not necessarily mean that all teachers who were supposed to receive training in these schools did. An impact estimation challenge arises, however, from the implementation of other education interventions in the treatment schools: only 49% of these schools received only TDP training.

According to the programme records data, TDP implementation is less compliant with assignment in the **control group**, as in 8% of the schools in the control group there were teachers that received TDP training. Further, the number of 'pure' control schools – schools in which teachers received no training at all – is low, at about 35%.

As discussed earlier, differential contamination of the evaluation sample is the primary concern for impact estimation. Differential receipt of other education interventions is not an issue in the case of Jolly Phonics and ESSPIN, as both treatment and control schools were comparably exposed to them (35% and 25%, respectively). Differential exposure may have a bearing on the impact estimates in the case of RANA (2% of treatment schools and 6% of control schools), and especially in the case of SUBEB-led training loosely based on TDP (2% of treatment schools and 36% of control schools).

Figure 3.1: Education programmes in sampled schools



3.1.3 Implications for the impact analysis

The results above indicate that contamination and non-compliance could be problematic for the impact analysis if they are not appropriately accounted for.

The approach taken by the evaluation team to address contamination by non-TDP programmes is to include variables that capture whether schools were exposed to other education interventions in each

impact analysis conducted in the context of this evaluation. Practically speaking, this meant creating school-level exposure indicators (binary variables based on programme records as defined above) and including them in the regression specifications listed below.

Non-compliance among treatment schools and contamination by TDP of control schools is dealt with via an instrumental variables (IV) approach, where randomised treatment assignment is used as an instrument for actual treatment receipt. Treatment receipt in this context is defined based on programme records data as specified above – at the teacher level for teacher-level analyses and at the school level for all other analyses.

While it is not possible to systematically control for non-compliance or contamination in the descriptive statistics presented throughout Volumes I and II of this report, the findings are caveated wherever these factors are deemed relevant.

3.1.4 Differential attrition

Section 3.2 below presented results on the overall level of attrition between baseline and endline surveys in this impact evaluation. It also presented the background characteristics of teachers and pupils that were correlated with the likelihood of individuals dropping out between both data collection rounds.

As described above, however, it is differential attrition that is of most concern to the impact analysis component of this evaluation. Differential attrition refers to situations where the background characteristics of individuals who drop out between the survey rounds differ significantly between control and treatment groups. This means that, after attrition, the two groups are not comparable anymore and that the original assumption of the control group being an appropriate counterfactual to the treatment group for impact identification purposes is not correct anymore. Such differential attrition could, hence, introduce selection bias into simple comparisons of outcomes across the treatment and control groups between baseline and endline, if not appropriately controlled for.

This section presents results that indicate that differential attrition in the TDP surveys is generally not problematic. To establish this, the evaluation team used baseline data to compare individuals who did not drop out of the sample over time across the treatment and control groups. Nevertheless, the evaluation team took steps to control for any potential biases introduced by attrition, which are presented below as well.

Pupil differential attrition

As discussed above, to assess whether differential attrition could be problematic in this impact analysis, the evaluation team compared non-attrited individuals at baseline across the TDP treatment and control groups. The purpose of this analysis is to assess whether, after attrition, the two groups are still appropriate counterfactuals. Table 3.3 presents the results of this analysis for pupils.

Estimates suggest that there are no significant differences between the control and treatment pupils who remained in the sample, when taking multiple hypotheses testing into account. When this correction is not taken into account, there is some indication of significant differences across treatment and control groups, associated with their wealth status and certain school characteristics.

Table 3.3: Results for differential pupil attrition since baseline

Variables	Pupils in the sample at both baseline and endline					
	Weighted estimates					
	1		2		3	4
	Control		Treatment		Diff (1-2)	Diff (1-2)
	Mean	N	Mean	N		
Pupils' age in years	8.88	597	8.74	587	0.14	0.14
Pupil is female (%)	40.7	782	43.14	783	-2.44	-2.44
State-wise pupil household asset index	0.65	776	0.68	783	-0.03	-0.03
Asset index – Quintile 1 (%)	10.59	776	15.15	783	-4.56**	-4.56
Asset index – Quintile 2 (%)	13.5	776	11.49	783	2.01	2.01
Asset index – Quintile 3 (%)	18.52	776	19.1	783	-0.58	-0.58
Asset index – Quintile 4 (%)	28.9	776	22.31	783	6.59**	6.59
Asset index – Quintile 5 (%)	28.48	776	31.95	783	-3.47	-3.47
Rasch score: Literacy	501.09	781	497.63	782	3.46	3.46
% questions correct: Literacy	16.9	782	16.64	783	0.26	0.26
Literacy Rasch level 0 (%)	57.97	781	61.98	782	-4.01	-4.01
Literacy Rasch level 1 (%)	39.39	781	35.77	782	3.62	3.62
Literacy Rasch level 2 (%)	2.64	781	2.25	782	0.39	0.39
Rasch score: Science	503.16	781	492.21	782	10.95	10.95
% questions correct: Science	46.27	782	43.58	783	2.69	2.69
Science Rasch level 0 (%)	16.4	781	20.02	782	-3.62	-3.62
Science Rasch level 1 (%)	67.56	781	66.22	782	1.34	1.34
Science Rasch level 2 (%)	16.04	781	13.76	782	2.28	2.28
Rasch score: Numeracy	502.87	781	497.35	782	5.52	5.52
% questions correct: Numeracy	27.49	782	26.53	783	0.96	0.96
Numeracy Rasch level 0 (%)	80.41	781	79.5	782	0.91	0.91
Numeracy Rasch level 1 (%)	14.52	781	16.07	782	-1.55	-1.55
Numeracy Rasch level 2 (%)	5.07	781	4.43	782	0.64	0.64
Num. of Primary 1–6 teachers currently employed	19.21	782	22.14	783	-2.93	-2.93
Num. of Primary 1–6 pupils currently enrolled	1258.63	782	1143.71	781	114.92	114.92
Pupil–teacher ratio	70	782	60.84	781	9.16*	9.16
Average daily teacher absenteeism (% of teachers absent)	12.11	782	9.74	783	2.37**	2.37
School has electricity supply (%)	21.1	782	31.57	783	-10.47	-10.47
School needs major repairs (%)	88.75	782	85.63	783	3.12	3.12

Source: TDP baseline survey. Notes: (1) Base population: non-attrited pupils. (2) Standard errors clustered at the school level. (3) ***, ** and * correspond to 1%, 5% and 10% significance levels. (4) Column 3 uses unadjusted p-values and Column 4 adjusts p-values for multiple hypothesis testing as described in (Sankoh *et al.*, 1997)

Teacher differential attrition

Table 3.4 presents results from the same type of analysis for teachers. Again, there is very limited indication of differential attrition. The only indicator for which a significant difference between the treatment and control groups can be identified is teacher absenteeism, a school-level indicator. This means that treatment teachers who remained in the sample were more likely to work at schools with lower levels of teacher absenteeism.

Table 3.4: Results for differential teacher attrition since baseline

Variables	Teachers in the sample at both baseline and endline					
	Weighted estimates					
	1		2		3	4
	Control		Treatment		Diff (1-2)	Diff (1-2)
	Mean	N	Mean	N		
Teachers' age	36.91	195	37.16	248	-0.25	-0.25
Teacher is female (%)	23.04	197	18	250	5.04	5.04
Total teaching experience in ANY school in 2014 (years)	12.22	196	11.87	247	0.35	0.35
Teacher has National Certificate of Education (NCE) qualification or above (%)	64.7	197	64.99	250	-0.29	-0.29
Teacher attended teaching-related training in last two years (%)	51.94	196	46.72	250	5.22	5.22
Teacher owns a mobile phone (%)	96.6	196	98.38	250	-1.78	-1.78
Raw teacher development needs assessment (TDNA) score: maths	42.53	191	43.3	235	-0.77	-0.77
Fully or near-sufficient maths subject knowledge (%)	37.82	191	38.3	235	-0.48	-0.48
Emerging maths subject knowledge (%)	41.13	191	41.2	235	-0.07	-0.07
Limited maths subject knowledge (%)	21.05	191	20.51	235	0.54	0.54
Raw TDNA score: English	21.54	191	20.53	235	1.01	1.01
Fully or near-sufficient English subject knowledge (%)	4.03	191	2.77	235	1.26	1.26
Emerging English subject knowledge (%)	36.3	191	38.36	235	-2.06	-2.06
Limited English subject knowledge (%)	59.67	191	58.87	235	0.8	0.8
Raw TDNA score: Science	20.68	191	20.15	235	0.53	0.53
Fully or near-sufficient science and technology subject knowledge (%)	2.29	191	2.67	235	-0.38	-0.38
Emerging science and technology subject knowledge (%)	34.55	191	31.34	235	3.21	3.21
Limited science and technology subject knowledge (%)	63.16	191	65.99	235	-2.83	-2.83
Raw TDNA score: Measuring pupil progress	12.85	191	13.25	235	-0.4	-0.4
Measuring pupil progress: Fully or near-sufficient	4.29	191	2.23	235	2.06	2.06
Measuring pupil progress: Emerging	10.39	191	14.41	235	-4.02	-4.02
Measuring pupil progress: Limited	85.32	191	83.36	235	1.96	1.96

Num. of Primary 1-6 teachers currently employed	12.97	197	12.78	250	0.19	0.19
Num. of Primary 1-6 pupils currently enrolled	712.72	196	649.5	249	63.22	63.22
Pupil–teacher ratio	56.77	196	57.82	249	-1.05	-1.05
Average daily teacher absenteeism (% of teachers absent)	15	197	11.71	250	3.29***	3.29**
School has electricity supply (%)	12.15	197	12.72	250	-0.57	-0.57
School needs major repairs (%)	88.84	197	85.83	250	3.01	3.01
Class size during lesson observation	46.78	193	41.85	242	4.93	4.93

Source: TDP baseline survey. Notes: (1) Base population: non-attrited teachers. (2) Standard errors clustered at the school level. (3) ***, ** and * correspond to 1%, 5% and 10% significance levels. (4) Column 3 uses unadjusted p-values and Column 4 adjusts p-values for multiple hypothesis testing as described in (Sankoh *et al.*, 1997).

Implications for the impact analysis

As discussed, the above analysis suggests that there is little evidence of differential attrition affecting the sample used for impact evaluation purposes in this evaluation. However, this evaluation dealt with any remaining risk in two main ways. First, characteristics that were slightly associated with differential attrition were included in econometric models to control for observable characteristics that might introduce selection bias. Second, extensive robustness checks via alternative estimation models were implemented, which also made it possible to control for unobservable background characteristics of individuals which might, in turn, be related to any systematic differences between the treatment and control groups that could have been introduced by differential attrition. The methodological basis for those robustness checks and results is presented in more detail below.

3.1.5 Quantitative impact analysis methods

This section presents the main quantitative analysis methods used in this impact evaluation to assess whether TDP has had a causal effect on the indicators listed in Table 3.1. It starts by establishing the potential outcomes framework that forms the basis of these approaches and how it relates to the original impact identification strategy intended to be used in this evaluation, presents the main alternative used by the evaluation team to produce results presented in Volume I, and then explains how different alternative analytical approaches were used as supplementary analyses and for robustness purposes.

The original impact identification strategy

As described above, the original strategy for this impact assessment relied on random assignment of schools, teachers, and pupils to either a treatment or control group to identify treatment effects of TDP.

One can think of this using the potential outcomes framework (Angrist and Pischke, 2009). In this framework, the observed outcome for any individual can be written as:

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0} \quad (1)$$

where Y_i is the observed individual outcome for individual *i*, $T_i$is a binary indicator for the treatment status which has value 1 for individuals if they are treated and 0 for individuals if they are not treated. Hence, Y_{i1} is the outcome for the individual *i* if the individual is treated and Y_{i0} if the individual is not treated. The treatment effect for individual *i* can then be written as:

$$\beta_i = Y_{i1} - Y_{i0} \quad (2)$$

The problem that any impact assessment tries to solve is that for any individual *i* it is never possible to directly observe the two outcomes Y_{i1} and Y_{i0} at the same time, which means that it is also not possible to directly observe the treatment effects directly. The two outcomes are potential or hypothetical outcomes, because any individual is either observed as being part of the treatment or the control group. Hence, average treatment effects need to be estimated by comparing individuals observed in those two groups.

Randomised control trials (RCTs) solve this problem by ensuring that the treatment status – that is, whether any individual is either in the treatment or in the control group – is randomly allocated and hence, on average, not associated with the background characteristics of individuals in those two groups that might themselves be associated with the outcome Y. Basically, RCTs ensure that the average treatment effect can be estimated by simply comparing average outcomes across the treatment and control groups. In a regression context, this can be expressed as follows:

$$y_i = \alpha + \beta_i T_i + u_i, \quad (3)$$

where y_i is the individual observed outcome, T_i is a dummy for observed treatment status, and u_i is the individual error term. Random allocation of treatment T_i ensures that it is exogenous in this context and uncorrelated with the error term: $E(u_i|T_i) = 0$. This means that estimating this equation using ordinary least squares (OLS) yields an unbiased estimate of the average treatment effect: $E[\hat{\beta}_{OLS}] = ATE = E[Y_{i1} - Y_{i0}]$. Note that the sampling strategy used in this evaluation (see Section 3.2) would have ensured the representativeness of this estimate for the population from which this sample was drawn overall.

It should be noted here that, in practice, the equation that would be estimated here would include a set of covariates (X_i) in order to increase the precision with which $\hat{\beta}_{OLS}$ could be estimated. Following (McKenzie, 2011), equation (3) would include outcome variables measured at endline and covariates measured at baseline.

However, as described above, TDP treatment was not implemented as planned. This means that, even though groups were randomly *assigned* to the treatment and control group, actual *receipt of* TDP training was not randomly allocated. This means that the randomised treatment implementation argument is not correct. In addition, there was a differential presence of non-TDP education interventions in TDP evaluation areas, which implies that outcomes might have been affected by these interventions differentially across the two groups. This means that $E(u_i|T_i) \neq 0$. In other words, it can no longer be assumed that treatment is fully exogenous. Instead, results relying on estimating (3) above using OLS would be affected by endogeneity, a selection bias. Thus, any identification strategy to retrieve an estimate of the programme effect needs to address this endogeneity.

The preferred alternative: the IV approach

An IV estimator can tackle such selection bias through the use of an instrument, i.e. a variable that is directly related to the regressor of interest in the equation above (T_i) but that does not have a direct relationship with the outcome y_i, other than via T_i. In the context of an impact evaluation originally relying on random treatment assignment but where non-compliance in the treatment group and contamination with the treatment in the control group is problematic (sometimes also referred to as two-sided non-compliance), treatment *assignment* can be thought of as an instrument for actual treatment *receipt*.

In the TDP context, treatment *receipt* was defined using programme records data, as described above. These data contain school- and teacher-level data on training actually provided by TDP, obtained from the programme implementers. The data were used at the most appropriate level depending on the outcome analysed: for school- and pupil-level outcomes (lm-1 and Effe-2 in Table 3.1), treatment *receipt* was defined at the school level. If any teacher was trained by TDP, the school was classified as having received TDP

training. For teacher-level outcomes (Effe-1 in Table 3.1), a teacher was classified as having received the training if the programme records reported that this was the case. It is important to emphasise here that, if a school is classified as having *received* TDP, that would not imply necessarily that all teachers in that school would have received the treatment. Given that it was not possible to uniquely match pupils to teachers in the sample, this means that not all pupils tested in a school classified as having *received* TDP would necessarily have been taught by TDP-trained teachers.

As explained above, treatment *receipt* cannot be considered to be randomly allocated in the context of this evaluation. However, given the original TDP impact evaluation design, treatment *assignment* can. The idea behind an IV approach is that *assignment* can therefore be used to explain random variation in the treatment *receipt* indicator. This random variation can then, in turn, be used to estimate the effect that TDP training implementation had on the outcome indicators.

The IV estimator works as a two-step process: in the first step the instrumented variable is regressed on the instrument and other explanatory variables and in the second step the outcome variable is regressed on the predicted value obtained from the first step and the other explanatory variables. To explain this in the regression context set out above, consider the following equation:

$$y_i = \alpha + \gamma X_i + \beta T_i + u_i, \tag{4}$$

where y_i represents the outcome variable for the individual *i*, X_i is a vector with individual covariates, and T_i is defined as the TDP treatment *receipt*. Finally, u_i is the error term. The IV approach then assumes the existence of an additional reduced-form model, a 'first-stage' equation, such that:

$$T_i = \pi_0 + \pi_1 X_i + \pi_2 Z_i + v_i, \tag{5}$$

where $Z_i$represents the instrument used. In the present case, Z_i is defined as the TDP treatment *assignment*. The two-step approach estimates first equation (7) using OLS and then uses the fitted values from this regression to estimate equation (6) to retrieve $\hat{\beta}_{LATE}$, again using OLS. Outcome variables would be measured at endline, while the covariates included would be baseline (or fixed – such as geographical location of schools) measurements.

For this approach to be appropriate for impact identification purposes, three assumptions generally need to hold (Greene, 2011; Imbens and Rubin, 2015):

The first assumption, **instrument relevance**, in the context of this evaluation refers to whether treatment assignment has a strong association with treatment receipt. Essentially, this is a statement about the strength of the relationship of Z_i and T_i in equation (7). Therefore, this assumption can be tested through the examination of the explanatory power of the first step of the estimations, and of the significance of the instrument Z_i with respect to the instrumented variable T_i.

The 'rule of thumb' for a satisfactory explanatory power is an F-statistics greater than 10 when estimating this explanatory power in a specification as in (7) (Staiger and Stock, 1997). Table 3.5 shows the F-statistics for the main specification of each outcome indicator that this evaluation is estimating impact on. The results show that treatment *assignment* has good explanatory power with respect to treatment *receipt*. It is also positive, as expected, and highly significant.

Table 3.5: Statistical tests on instrument relevance

Outcome	Instrument	Estimate	p-value	F-statistics
Pupil learning – maths	Treatment assignment	0.9029	0.000	F(55,157)=118.38
Pupil learning – English	Treatment assignment	0.9029	0.000	F(55,157)=122.39
Pupil learning – science	Treatment assignment	0.9028	0.000	F(55,157)=126.40
Teachers' positive interaction	Treatment assignment	0.7499	0.000	F(38,141)=64.20
Teacher absenteeism	Treatment assignment	0.8983	0.000	F(46,176)=313.93

Note: The F-statistics are the results of the standard F-test for the significance of the instrument in the first-stage regression. Please refer to Greene (2011) for further details.

The second assumption, **instrument exogeneity**, refers to the relationship between treatment *assignment*, *receiving* treatment, and treatment effects on the outcomes of interest. For treatment assignment to be truly exogenous, and therefore to be a valid instrument, causal effects of treatment assignment should only materialise via receiving treatment, and not via any other channels.

In other words, treatment assignment (Z_i) should affect the probability of actually receiving treatment (T_i), which in turn might affect the outcome of interest (Y_i), but there should not be any other channel via which treatment assignment might affect the outcome of interest. This assumption is also sometimes referred to as the exclusion restriction and cannot be tested formally. It is the key assumption that needs to hold in order to identify programme impact and, in the context of the regression specifications mentioned above, can be specified as $E\left(u_i \middle| Z_{i,}\right) = 0$.

It should be noted here that identifying treatment effects in the present context in fact relies on the conditional version of this assumption, where the exclusion restriction holds conditional on other covariates included in (6) and (7): $E(u_i|Z_i, X_i) = 0$. A key set of covariates included are indicators that control for parallel implementation of non-TDP education interventions, in order to ensure that any contamination introduced by those is also taken care of.

As said, this assumption cannot be formally tested: 'A key feature of these exclusion restrictions is that they are, at their core, substantive assumptions, requiring judgment regarding subject-matter knowledge.' (Imbens and Rubin, 2015, p. 550). As described above, treatment assignment was random in this impact evaluation, which implies that, *a priori*, it was not dependent on school-, teacher-, or pupil-specific characteristics, hence ensuring that there should not be any correlation between those characteristics, treatment assignment, and the outcomes of interest.

A condition that could invalidate this assumption, however, is if treatment assignment itself had a behavioural effect on the units of observation, and hence the outcome indicators of interest, other than via TDP training implementation. For example, if schools were assigned to the treatment group and head teachers or teachers changed their behaviour because of that, even in the absence of an actual implementation of TDP training, then this would invalidate the exclusion restriction.

There is no evidence for such effects in this impact evaluation. Teachers and head teachers were aware of being in the TDP treatment group only because of TDP training implementation, which is captured via the treatment *receipt* indicator. It is difficult to therefore conceive of a way in which TDP treatment assignment could have affected outcome variables other than via the actual implementation of TDP training.

The third assumption needed here, **monotonicity,** implies that there are no units of observation in the sample that in any situation have a treatment receipt status that is exactly the opposite of what they were assigned to, i.e. they do not receive the training programme because they were assigned to it or they do receive the training because they were not assigned to it. There is no evidence among the studied population of such 'defier' behaviour and so the authors hold this assumption to be true.

Given all three assumptions, using this IV approach and implementing the two-steps estimator yields an unbiased estimate of the impact of *receipt* of TDP treatment. **These are the estimates presented in Volume I of this report.**

It is important to emphasise, however, that the coefficient obtained ($\widehat{\beta_{LATE}}$), represents an estimate of the local average treatment effect (LATE), rather than the average treatment effect. The LATE is the effect of receiving the treatment for those individuals who comply with the treatment assignment, i.e. effectively would change their treatment implementation status because of a change in the treatment assignment. This means that estimates presented in this report cannot directly be extended to the full population from which the sample used here is drawn. The results from supplementary analyses indicate, however, that the results presented in this report do not vary substantially when other estimation techniques are used to estimate slightly different values.

Supplementary analyses: robustness checks

As discussed above, in the context of this evaluation, IV estimations provide a robust estimate of the impact of the implementation of TDP training on the various outcomes considered. It is important to emphasise that this impact relates to TDP training *receipt* as defined by programme records data, as explained above.

However, the evaluation team implemented various robustness checks to compare the LATE parameter estimated using this IV approach specified above with alternative treatment estimates, such as the intention-to-treat (ITT) effect and an estimate of the average treatment effect on the treated (ATET).

Estimating these effects relies on a different set of assumptions and allows the use of different types of estimation techniques. The purpose of these robustness checks is, hence, to verify how sensitive results are to changes in underlying assumptions and estimation techniques, and to thereby provide an assessment of the robustness of the headline findings presented in Volume I.

Robustness check 1: ITT analysis

ITT analyses estimate the impact of being assigned to the treatment group on outcomes of interest, as opposed to actually *receiving* the treatment. Such estimations are typically estimated in contexts where there are individuals in treatment or control groups that did not comply with their treatment status, as in the context of the present evaluation. This estimate is often used, for example, in situations where medicine is offered as a treatment but where it is not clear whether individuals actually took the medicine or not.

The identification of the ITT is based on the assumption that treatment *assignment* is random, which was the case in this evaluation, as discussed above. The regression specification used to estimate the ITT here is defined as follows:

$$y_i = \alpha_i + \gamma X_i + \beta Z_i + u_i\,, \qquad (6)$$

where, as above, y_i represents the outcome of interest, X_i is a vector of covariates, and Z_i is the treatment *assignment* indicator.[1] Finally, u_i is the error term, which is assumed to be uncorrelated with Z_i as in the standard RCT context presented earlier. Estimating this equation using OLS yields an estimate of $\widehat{\beta_{ITT}}$, an estimate of the ITT. As before, outcome values are measured at endline, while the covariates included are baseline measurements.

The ITT estimate can therefore be interpreted as the difference in the outcome of those who were intended to receive the programme compared to those who were not intended to receive the programme, without addressing the issue of two-sided non-compliance. While both ITT and LATE are causal impact estimates, they therefore capture two slightly different concepts.

In the context of this evaluation, the ITT estimates the average difference in outcomes between schools assigned to TDP treatment and schools not assigned to the TDP treatment – irrespective of whether TDP training was actually implemented in a school or not. On the other hand, the LATE estimates the treatment effect for only those individuals from schools who complied with the treatment assignment.

Non-compliance is the reason why it is reasonable to expect the ITT estimate to generally have a lower magnitude than the LATE estimate: among the schools assigned to the treatment group, there are some schools in which TDP was not implemented, while in some schools assigned to the control group TDP was implemented. Assuming that, for example, TDP might affect pupil learning outcomes positively, this means that averages in the assigned treatment group might therefore be lower than among compliers in the treatment group, and in the assigned control group higher than compliers in the control group. The difference between the two averages might therefore be smaller than when compliers only are compared.

Robustness check 2: ATET analysis relying on conditional independence – including automated covariate selection

A second robustness check was implemented to directly estimate the ATET. This is the average impact for those individuals who actually received the treatment. This involves comparing units of observation that did receive TDP training, as defined above, to units of observations that did not. Practically speaking, this means estimating equation (6) directly, without the use of treatment *assignment* as an instrument:

$$y_i = \alpha_i + \gamma X_i + \beta T_i + u_i \tag{7}$$

where, as above, y_i represents the outcome variable, X_i is a vector with the individual covariates included, and T_i is the treatment *receipt* indicator. As described above, however, treatment receipt cannot be assumed to directly be exogenous here.

Retrieving an unbiased ATET (not affected by endogeneity or selection bias) therefore requires the conditional independence (or selection on observables) assumption to hold (Cameron and Trivedi, 2005): all those variables that are correlated with treatment *receipt* and also correlated with outcomes of interest, can be observed and controlled for. If this assumption holds, all differences between treatment and control groups can be attributed to the programme implementation. Formally, this assumption can be expressed as:

$$E[u_i|T_i, X_i] = 0. \tag{8}$$

[1] For consistency purposes, the same notation for equations (3) and (4) is kept. *Z* here represents the same variable as above, even though in this case it is not used as an instrument.

The assumption implies that, conditional on covariates X_i, the treatment receipt indicator T_i is randomly allocated and estimating equation (9) using OLS yields an unbiased estimate of the treatment effect: $\widehat{\beta_{ATET}}$.

This last equation exemplifies the importance of controlling for the right covariates in this estimation procedure. If variables that might be driving selection bias are not controlled for, then the conditional independence assumption does not hold and the estimate of β_{ATET} in equation (11) will be biased. As described above, the main ATET estimation using OLS in this evaluation is implemented using a comprehensive set of covariates, as illustrated further below. As before, outcomes are measured at endline, while covariates are baseline measurements.

In addition, however, the evaluation team implemented additional analyses using an approach that included automated and principled covariate selection using machine learning (ML) following recent methodological advances in this area. This approach is explained in more detail in Box 1.

It is important to emphasise here the 'selection on observables' connotation of the conditional independence assumption. This means that if any of the possible endogeneity in (9) is due to unobservable characteristics of the units of observation, then even controlling for X_i will not yield an unbiased estimate of β_{ATET}. The following section explains how this was addressed by implementing a set of further robustness checks.

Box 1: Double ML for causal inference (least absolute shrinkage and selection operator (LASSO) model selection technique)

The problem of covariate selection

All of the estimation techniques presented in this section rely, at least partly, on some form of the conditional exogeneity assumption, or include a set of baseline covariates for precision purposes. The conditional exogeneity assumption states that in order to identify programme impact, all observable covariates that might be driving selection bias or endogeneity in the relationship between treatment and the outcome variable are controlled for. Including covariates for precision purposes means that variation in the outcome that is due to these covariates is controlled for, which means that treatment effects can be estimated more efficiently.

In these contexts, researchers generally need to make decisions about which covariates to include in their models in order to be able to claim that the conditional exogeneity assumption holds or that they have fitted their model appropriately to increase estimate precision. These decisions are typically based on theoretical knowledge, *a priori* assumptions about the relationships at hand, or substantive knowledge of the programme effects being analysed. This is how the evaluation team decided on the set of covariates to control for in the majority of the models presented in this report.

However, making this decision is not a trivial task given that researchers are often faced with a large set of possible covariates to select from and a variety of ways of controlling for them. For instance, in the case of this impact evaluation, the baseline data related to pupil learning outcomes potentially would have allowed controlling for over 200 variables.

Over the past few years, new approaches based on ML methods have been developed to transform covariate selection into a principled, algorithm-driven process. In order to ensure that the results presented in this report are not driven by researcher discretion in selecting the sets of covariates included in their models, the evaluation team employed one such ML approach as an additional robustness check to the main findings presented in Volume I.

How was ML used in this impact evaluation?

The approach used in this impact evaluation follows the emerging double selection and double ML literature. (Belloni, Chernozhukov, and Hansen, 2014; Chernozhukov *et al*., 2018). Both approaches build on the idea that, when the ultimate objective is to estimate causal effects by controlling for selection bias, good covariate selection means that one needs to build a model with variables that are related to treatment status indicators and that are also related to the outcome of interest.

For an IV setting, ML algorithms should be applied to three separate estimation problems in a first step: how covariates are related to the instrument, the treatment receipt, and how they are related to the outcome variable. For a simple one-stage OLS setting, only two of those problems need to be taken into account: the relationship between covariates and the treatment receipt, and between covariates and the outcome. In a second step, the results of these first analyses can be used to estimate the impact of the treatment.

The fundamental insight driving this approach is that the first step in this process can be interpreted as a prediction problem that ML can help to address, given its strength in predictive analysis even when facing a very large set of potential predictors or covariates. In this evaluation, in the first step, regularised regression via LASSO is used to predict both outcomes and the different TDP treatment statuses. Prediction errors, that is to say residuals, are recorded and these ae then used in the second step to estimate the effect of the programme.

The intuition behind this approach is that if the relationships between covariates and the outcome, on the one hand, and between covariates and the treatment status, on the other, are modelled well in the first step, the remaining prediction errors, or residuals, will capture information that cannot be explained by the covariates controlled for. Hence, this information should reveal whether once covariates are taken into account, exogenous variation (i.e. variation that is not related to the covariates) in treatment status can explain the remaining variation in the outcome variable. In other words, once ML is used to explain all variation that is due to the background characteristics of treated and non-treated observations, one can assess whether differences in outcomes between these groups persist that can be explained by the remaining treatment status variation. Note that this is just a restatement of the conditional exogeneity assumption: once all endogeneity by covariates is taken into account, treatment status variation is orthogonally related to variation in the outcome variable.

More formally, the evaluation team employed a double partialling out approach using LASSO regularisation to estimate both the LATE and the ATET in the context of the non-panel models presented in this section. (See Fonti (2017) for a description of the LASSO operator.) This involved implementing the following steps:

First, applying LASSO to the following three equations:

$$(a)\ y_i = \alpha X_i + u_i,$$
$$(b)\ T_i = \beta X_i + u_i,$$
$$(c)\ Z_i = \gamma X_i + u_i.$$

The notation is as before, where T_i is TDP treatment *receipt*, Z_i is TDP treatment *assignment*, and $y_i$is the outcome of interest. In contrast to the other models presented in this section, the matrix of covariates X_i, however, contains an extended list of possible baseline covariates collected via the TDP survey (including geographical controls and indicators capturing contamination by non-TDP interventions) interacted with each other and, in addition, including second order polynomials of continuous variables. For the case of pupil learning outcomes, this means that this matrix contains, for example, a set of several hundred covariates. The reason for creating this large matrix is that there is no *a priori* reason for assuming that endogeneity is only linearly related to covariates. Inflating the

matrix, as described here, means that LASSO regularisation can also take into account potential non-linear relationships that might be driving selection bias.

For each of the equations above, LASSO regularisation produces a prediction model that uses only a subset of these covariates, deemed to be relevant predictors of the left-hand side of each equation. For the purposes of this robustness check, the evaluation team used the LASSO approach developed by Chernozhukov, Hansen, and Spindler, (2016), implemented in R, that relies on a theoretically grounded choice of the penalty term for LASSO to select a valid prediction model for (a), (b), and (c). This approach prevents regressions on (a), (b), and (c) to overfit, and hence to produce valid predictor models of each of the left-hand side variables. These models are therefore then used to predict each of the left-hand side variables above and to calculate residuals:

$$(d)\ \tilde{y}_i = y_i - \hat{\alpha}_{LASSO} X_i,$$
$$(e)\ \tilde{T}_i = T_i - \hat{\beta}_{LASSO} X_i,$$
$$(f)\ \tilde{Z}_i = Z_i - \hat{\gamma}_{LASSO} X_i.$$

In a second step, residuals from (d), (e), and (f), are used to calculate either the ATET using OLS or the LATE using an IV approach. For the ATET, this involved estimating the following equation using OLS, where the estimated coefficient $\hat{\delta}$ can be interpreted as the ATET estimate:

$$(g)\ \tilde{y}_i = \delta\, \tilde{T}_i + \epsilon.$$

For the LATE, this involved using the two-step estimator on the following set of equations, where $\tilde{Z}_i$ is the instrument and $\tilde{T}_i$ the variable instrumented for:

$$(g)\ \tilde{y}_i = \delta\, \tilde{T}_i + \epsilon_i,$$
$$(h)\ \tilde{T}_i = \theta \tilde{Z}_i + \epsilon_i.$$

For inference purposes, this last step of the process is implemented taking the full survey settings into account, including weights, stratification, and clustering of standard errors at the school level.

Estimation results from this process are presented in Section 3.1.6 below. They are clearly highlighted as OLS or IV LASSO estimations and presented next to estimations derived from other models implemented in the context of this impact assessment. The automated nature of this ML-driven process removes researcher discretion from the crucial step of covariate selection in these models. These results therefore allow for an important check regarding how sensitive other results are with respect to changes in the covariates included.

Robustness check 3: using panel estimation techniques and difference-in-differences

As discussed above, all of the estimation procedures presented so far make use of data from both the endline and baseline surveys implemented in the context of this evaluation. In particular, outcome indicators were measured at endline while the covariates used were measured at baseline. There are several benefits to this approach, including potential increases in statistical power and the fact that baseline covariates cannot be affected by the treatment, and hence controlling for them is not fraught with the potential danger of reintroducing endogeneity into the estimation (McKenzie, 2011).

This means, however, that the measurements of both the outcomes and the covariates used so far were limited to one time-period only (endline for outcomes, baseline for other variables). However, it is also

possible to take the panel structure of the data more fully into account, exploiting the fact that there are repeated measures of both outcomes and covariates in the data.

Further robustness checks implemented here exploit this structure using panel estimation models. Such panel data methods can control for unobserved individual heterogeneity, providing an alternative identification strategy to the ones described so far. Using such panel data methods, it is possible to retrieve an estimate of all three parameters described above (LATE, ITT, and ATET). The following methodological summary generally follows Cameron and Trivedi (2005)). A general model (sometimes called the linear unobserved effects model) for panel data in the context of TDP can be written as:

$$y_{iw} = \alpha_w + \gamma X_{iw} + \beta T_{iw} + u_{iw}, \quad (9)$$

where y_{iw} represents the outcome for each individual *i* at survey wave *w* (either baseline or endline), X_{iw} is a vector with the individual covariates included at each wave, and T_{iw} is the TDP treatment indicator (either assignment or treatment receipt). In the context of panel methods, one can then think of the error term u_{iw} as being composed of an individual time-invariant error term and an idiosyncratic error: $u_{iw} = c_i + \varepsilon_{iw}$, which yields the following full equation:

$$y_{iw} = \alpha_w + \gamma X_{iw} + \beta T_{iw} + c_i + \varepsilon_{iw}. \quad (10)$$

Depending on the assumptions made about how the time-invariant (unobserved) error term (c_i) is correlated with regressors (including the treatment indicator), different estimation techniques will yield a consistent and unbiased estimate of the treatment effect β. Similarly, this will be affected by whether T_{iw} is considered to be the exogenous treatment assignment or the treatment receipt that is potentially endogenous.

Following the approaches mentioned above, and the methodological guidance provided in Cameron and Trivedi (2005), the results from the following estimations are presented below:

- First, assuming that, as above, treatment assignment (Z_i) is a valid instrument for treatment receipt (T_{iw}) in the context of equation (13), one can use panel IV estimation techniques (either fixed effects (FE IV) or random effects (RE IV)) to estimate the **LATE** under slightly weaker assumptions than in 0. Basically, the unobserved time-invariant error term in (13) is generally allowed to be correlated with the instrument (Z_i). In the present case, estimates of all outcome indicators in Table 3.1 are based on FE IV, except for the indicator on teachers' positive interaction.[2]

- Second, assuming that, as in the previous robustness check section, treatment receipt (T_{iw}) could be considered to be exogenous when controlling for observable background characteristics and the unobserved error term (c_i), one can use standard panel estimation techniques (fixed effects (FE) or random effects (RE)) to estimate the **ATET**, under slightly weaker assumptions than above. Basically, the conditional independence assumption now holds conditional on the fact that panel estimation techniques take this unobserved term into account. As above, RE methods are only used for the indicator on teachers' positive interactions, and otherwise the results presented below use FE analysis techniques where possible.

- Finally, for completeness purposes, the sections below present the results from a difference-in-difference (DID) estimation of the treatment *receipt* indicator (T_{iw}). The regression specification here is as follows:

[2] This decision was based on a Hausman test, implemented following Cameron and Trivedi (2005). This tests the null hypothesis that individual-specific effects are uncorrelated with regressors. A rejection of the null hypothesis is evidence of the presence of individual fixed effects, which helps guide the choice of fixed effect model.

$$y_{iw} = \phi T_{iw} + \delta W_w + \beta T_{iw} * W_w + \gamma X_{iw} + u_{iw} \quad (11)$$

The indicator W_w is a dummy that indicates whether an observation is from the baseline or endline survey wave. The identifying assumption here is a form of conditional independence assumption, that implies that controlling for all covariates in (14), and given the treatment assignment, survey wave indicator, and their interaction $(T_{iw} * W_w)$, the estimated coefficient on this interaction is an unbiased estimate of the **ATET**. This assumption is sometimes also called the (conditional) common trend assumption.

Comparing supplementary analyses to the main estimation

As mentioned above, the preferred analytical approach in this impact evaluation to identify the impact of TDP was to estimate the LATE using IV methods. Results derived from supplementary analyses and robustness checks need to be interpreted taking the below points into account, which establish a hierarchy among these results:

- First, analytical approaches that exploit the random assignment of TDP treatment can generally be assumed as providing more robust causal inference. The original design of this study provides a strong argument for exogeneity to this assignment indicator compared to the actual treatment receipt indicator, conditional on controlling for any differential contamination introduced by non-TDP programmes using programme records data.

- Second, following (McKenzie, 2011), it is assumed that approaches that rely on endline measurement of outcomes while controlling for baseline covariates can provide more efficient estimates of treatment effects than pooled panel estimates.

- Third, this evaluation aims to estimate the effect of TDP treatment implementation, rather than the effects of being assigned to the treatment group. Hence, getting around the issue of two-sided non-compliance is important and ITT estimates do not precisely estimate the value that is of interest to this evaluation.

- Fourth, controlling for covariates at endline is difficult, given that endline characteristics of pupils and teachers might have been affected by the TDP treatment. Panel specifications generally therefore include a limited set of covariates only. This means that arguments of conditional exogeneity based on controlling for observables (i.e. where all relevant covariates are taken care of) are less strong in these specifications. At the same time, however, panel specifications control for unobserved fixed characteristics that might be introducing endogeneity and hence provide an important robustness check to the main specifications used to identify causal effects.

- Finally, automated covariate selection and double ML (see box above) provides a robust way of selecting baseline covariates, increasing the credibility of conditional independence arguments in those specifications.

Supplementary analyses: changing the definition of the outcome variable

It should be noted here that for pupil learning outcomes, different types of outcome variables could be used to estimate the impact of TDP. In Volume I, results are presented for the scaled standardised mean scores in English, maths, and science. However, as described in Chapter 6, these scores are transformed versions of raw scores that can be derived from pupil tests. The results presented below are generally robust to changing the outcome measure used from the scaled version to raw scores.

Analysis implementation: data used, survey context, and covariate selection

As described above, this impact assessment was implemented using data from the baseline and endline rounds of the TDP school survey. In all procedures implemented to estimate the models specified above, the sampling structure of this survey was taken into account, which means that survey weights (see Section 3.3) were included, and standard errors were clustered at the school level. All estimations were performed using Stata (version 14).

All the estimations specifications described above include a set of covariates, indicated by the matrix X_i. Under perfect randomisation, including baseline covariates would serve the purpose of increasing the precision with which treatment effects can be identified, because these covariates can explain some variation in the outcome variable that is not due to treatment. In the present case, including covariates in X_i also serves the purpose of ensuring appropriate identification of TDP programme effects. Indeed, the conditional exclusion restriction described above relies on this.

Hence, following a principled approach to covariate selection is a key component of this evaluation. Table 3.6 below summarises the different categories of covariates included in the models estimated here. The underlying idea is that covariates that might be correlated with treatment status and the outcome variable should be included here as that might be correcting for any potential endogeneity. It should be noted again that the set of covariates that can be included in the panel data models is limited by the fact that one should not include variables in the model that might have been affected by the programme itself, and that any covariates that do not vary over time will be automatically excluded from fixed effects estimation procedures (which also prevents inclusion of baseline-only values). Note that this last point does not apply to the DID estimation. Finally, it should be noted that covariates with large sets of missing values were excluded from the estimation procedure.

The set of covariates included in the non-panel estimations can be summarised as follows:

- **Baseline measurement of outcome variable.** Controlling for baseline outcomes makes it possible to capture pre-intervention variation in the outcome, and hence increases the power and precision with which effects can be identified. (McKenzie, 2011).
- **Pupil-level, school-level, and head teacher characteristics.** These variables included demographic characteristics and household characteristics where available. They also included infrastructure characteristics of schools.
- **Non-TDP contamination indicators.** As discussed above, these control for the implementation of other, non-TDP interventions in schools that were surveyed as part of this evaluation.
- **Geographical controls.** These include state and LGA fixed effects.

A full list of variables is available in Annex E.

Table 3.6: Main groups of covariates

Outcome variable	Covariates included in non-panel estimations	Covariates included in panel estimations
Pupil learning outcomes (maths, English, science)	• Outcome baseline measurement • Pupil background characteristics (demographics, household)	• Head teacher background characteristics • School infrastructure • Non-TDP contamination

	• School background characteristics (infrastructure, staff, checks from authorities) • Head teacher background characteristics • Non-TDP contamination • Geographical controls	
Positive classroom interaction	• Outcome baseline measurement • Teacher background characteristics (demographics, educational attainment) • School background characteristics (infrastructure, staff, checks from authorities) • Non-TDP contamination • Geographical controls	• Head teacher background characteristics • School infrastructure • Non-TDP contamination
Teacher absenteeism	• Outcome baseline measurement • School background characteristics (infrastructure, staff, checks from authorities) • Head teacher background characteristics • Non-TDP contamination • Geographical controls	*NA*

In models that aim to estimate the effect of TDP on pupil learning outcomes, some variables based on information reported by pupils, such as age and number of children present at home, have a high number of missing observations. In particular, pupils at baseline were likely to not know their own age. In this case, if the age was instead reported at endline, the baseline value was calculated accordingly. Still, relatively high numbers of missing values were still present for those variables. Estimations have been conducted with and without these variables, to assess robustness of the findings to changes in these specifications: the results do not vary significantly in terms of the impact estimated. (Regression results are available upon request.)

3.1.6 Results

How results are presented in this volume

For each outcome variable defined in Table 3.1 the results are summarised in this section using both a graph and a regression output table, to allow for a detailed comparison between estimates derived from the different methodological approaches presented above. The following paragraphs use the example of Figure 3.2 and Table 3.7 to explain how the results can be interpreted.

First, a figure comparing the different estimates is presented (Figure 3.2), with point estimates being represented by a small circle, and its confidence interval represented by a blue line. This is similar to the graphs used in Volume I. In each graph, the red line marks zero. Roughly speaking, if the confidence interval overlaps with this line, the coefficient is not statistically significant.

The y-axis indicates what estimate each of the points represents and specifies the estimation method used to generate this value. Estimates are presented as follows:

- The first item represents the LATE estimate obtained using IV methods, as specified above (equations (4) and (5)). Results related to this estimation are also shown in column (1a) of the subsequent table.

- The second item represents the ATET estimate obtained using OLS, as specified in the robustness check 2 above (equation (7)). Results related to this estimation are also shown in column (2a) of the subsequent table.

- The third item represents the ITT estimated obtained using OLS, as specified in robustness check 1 above (equation (9)). These results are also shown in column (3) of the related table.

- The fourth item represents the ATET estimate obtained with panel estimation techniques, as specified in robustness check 3 above (equations (9) and (10)). This is also shown in column (4) of the related table.

- The fifth item represents the panel LATE estimate obtained with panel estimation techniques, as specified in robustness check 3 above. Results are also shown in column (5) of the related table.

- Finally, the sixth item represents the ATET estimate obtained using DID estimation, as specified in equation (11) above. This result is also shown in column (6) of the related table.

The table (Table 3.7) presents a summary of estimated coefficients, with their standard errors in parenthesis and the number of observations included in each estimation specification. The table also differentiates on whether treatment assignment or treatment receipt was used in the specification. It additionally reports the results obtained using double ML for automated covariate selection for the LATE and ATET estimate (columns 1a and 2a). (See Box 1 for an explanation.)

Regression tables that show coefficient estimates for the full set of covariates included in these estimations are presented in Annex F.

Teacher effectiveness: teachers' positive interaction during lessons

Figure 3.2 and Table 3.7 below present results related to the impact of TDP on the time teachers involve pupils in positive interactions during lessons. Both the graph and the table show that, in this case, receiving TDP training had a positive and significant (at the 5% level) effect on the proportion of time that teachers spent in positive interactions with their pupils. This result is confirmed by the main identification strategy (LATE IV) and the robustness checks. Therefore, TDP increased the proportion of time spent in positive interactions for those teachers who received TDP training.

It is important to emphasise that the confidence intervals for all estimates presented below overlap significantly. Statistically speaking, the estimates are therefore indistinguishable from each other. Irrespective of whether one looks at the LATE, ATET, or the ITT, positive impacts of TDP can be confirmed, but there are no significant differences between them. Small differences in point estimates can be expected though. For example, the fact that the ITT estimate is smaller than the LATE estimate can be expected due to the nature of the two-sided non-compliance, as explained above. Note that the results do not change when using automated covariate selection: LASSO-driven results are presented in the table below and confirm the overall finding.

Regression results including all covariates can be found in Annex F.1 and F.2.

Figure 3.2: Teachers' positive interaction (comparison of results)

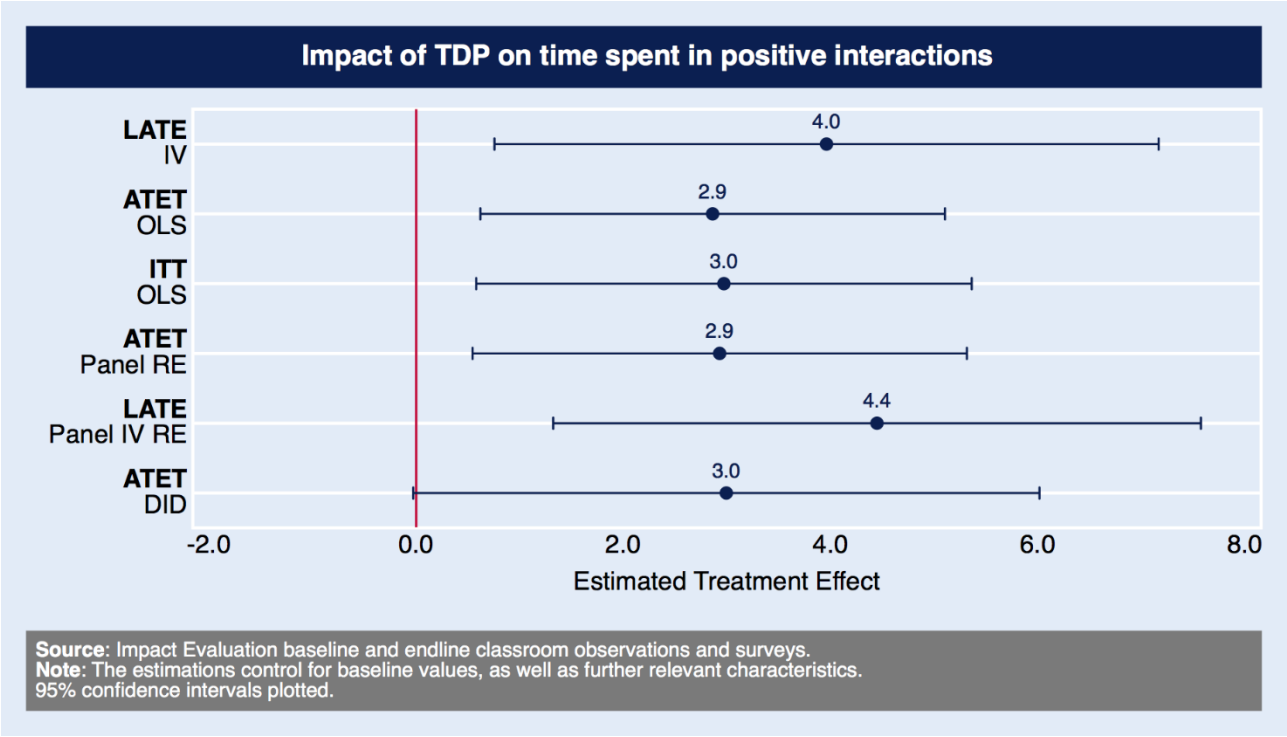


Table 3.7: Teachers' positive interaction (summary of results)

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Estimation technique	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel RE**	**Panel IV RE**	**DID**
Estimate	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	0.040*	0.030*	0.029*	0.049**	-	0.029*	0.044**	0.006
	(0.016)	(0.013)	(0.011)	(0.018)		(0.012)	(0.016)	(0.012)
Treatment assignment	-	-		-	0.030*	-	-	-
					(0.012)			
DID (treatment receipt)	-	-		-	-	-	-	0.030
								(0.015)
Constant	0.256***	0.000	0.263***		0.270***	0.248***	0.246***	0.183***
	(0.054)	(0.006)	(0.055)		(0.058)	(0.025)	(0.025)	(0.039)

Observations	443	443	443	443	443	909	909	857
R-squared	0.174	-	0.177	-	0.177		-	0.152
	Note: Estimates include covariates as specified above, survey weights, and robust standard errors clustered at the school level. Standard errors are reported in parentheses. The complete tables with the results are available in the annex.							
	*** p<0.001, ** p<0.01, * p<0.05							

Teacher effectiveness: teacher absenteeism from school

Figure 3.3 and Table 3.8 present the results of impact estimations on the second teacher effectiveness indicator: teacher absenteeism from school. The main strategy suggests that TDP did not have a significant effect on teacher absenteeism. This is also confirmed by the analysis of the ATET and the ITT: no parameter is statistically significant. Note that baseline data on this indicator could not be used in a panel setting because measurement was not deemed to be comparable across time. The low number of observations for this outcome, which is at the school level, did not allow for the implementation of the LASSO covariate selection.

Regression results including all covariates can be found in Annex F.3.

Figure 3.3: Teacher absenteeism

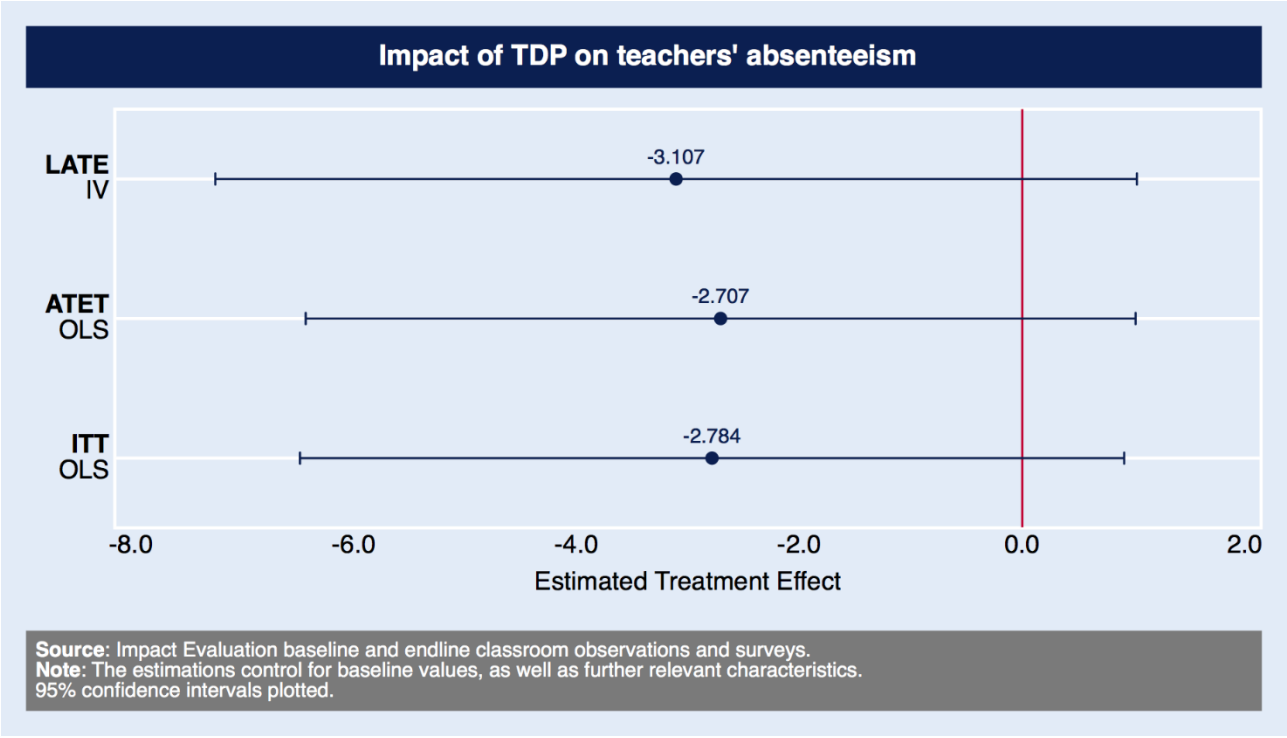


Table 3.8: Teacher absenteeism – summary of results

	(1)	(2)	(3)
Model	**IV**	**OLS**	**OLS**
Parameter estimated	**LATE**	**ATET**	**ITT**
Treatment receipt	-3.107	-2.707	-

	(2.099)	(1.890)	
Treatment assignment	-	-	-2.784
			(1.877)
Constant	19.174	18.901	19.139
	(10.958)	(10.891)	(10.984)
Observations	305	305	305
R-squared	0.182	0.184	0.184
Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights. The complete tables with the results are available in the annex.			
*** p<0.001, ** p<0.01, * p<0.05			

Pupil learning – maths

Figure 3.4, Table 3.9, and Table 3.10 present results related to the impact of TDP training on pupil learning outcomes in maths. Figure 3.4 shows that all estimation procedures except one indicate that there is no impact of TDP on pupil learning outcomes, as confidence intervals overlap with zero. The results are, however, weakly significant (at 5% level) in the DID estimation. This is also confirmed when looking at Table 3.9.

To better understand this last result, further DID estimations were implemented using two alternative measurements of the outcome variable:

- an estimation using non-standardised scores measured as the raw sum of scores each pupil obtained in the test; and
- the percentage of the raw scores with respect to the highest raw score possible.

Table 3.10 reports the coefficient estimated with a DID model, using these two alternative measures. The DID ATET estimates obtained with these alternative measures are not statistically significant.

Taking into account these results, and the hierarchy of comparisons established above that emphasises the strength of inference based on the LATE IV and ITT estimations, these results do not point towards an impact of TDP training on pupils' learning in maths.

Regression results including all covariates can be found in Annex F.4 and F.5.

Figure 3.4: Pupil learning – maths

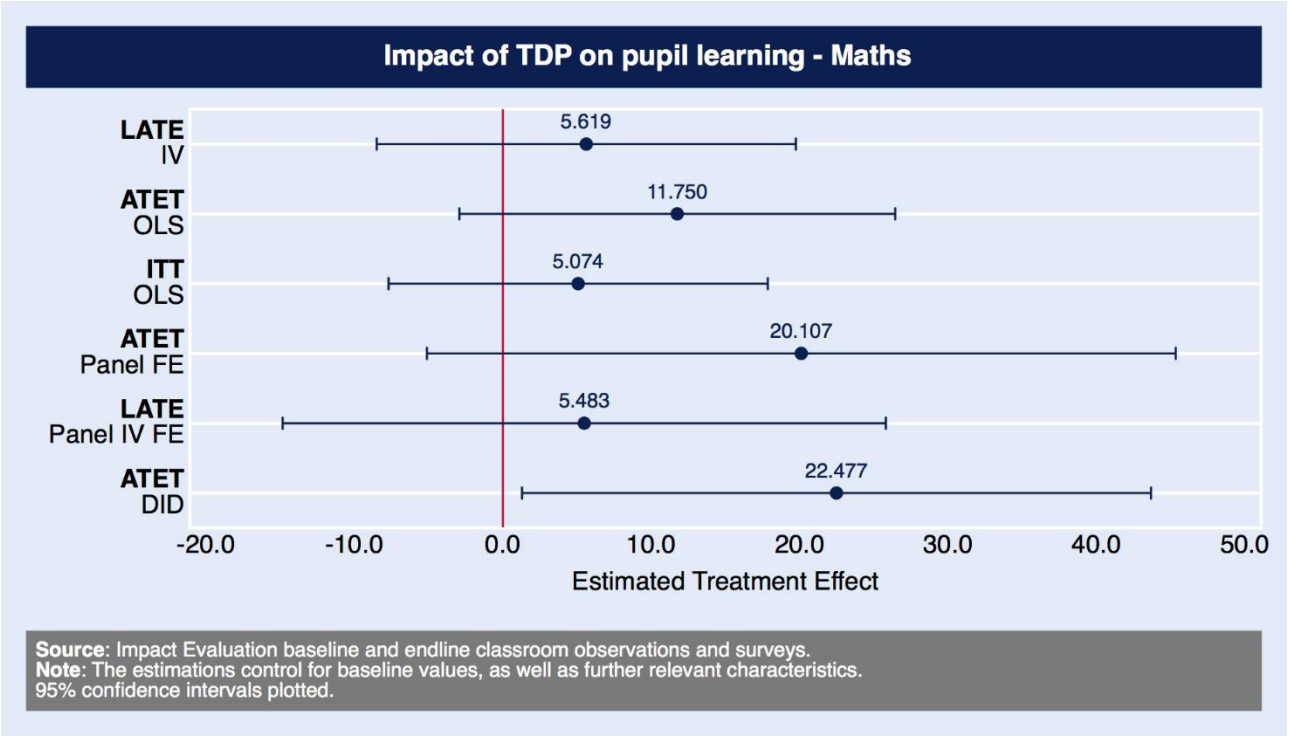


Table 3.9: Pupil learning – maths: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Model	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel FE**	**Panel IV FE**	**DID**
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	5.619	15.737	11.750	7.714	-	20.107	5.483	-5.186
	(7.163)	(12.105)	(7.447)	(10.652)		(12.818)	(10.367)	(8.080)
DID	-	-	-	-	-	-	-	22.477*
								(10.754)
Treatment assignment	-	-	-	-	5.074	-	-	-
					(6.481)			
Constant	234.926***	-2.059	131.754	-1.910	200.413**	471.083***	459.501***	459.812***
	(60.099)	(3.536)	(67.164)	(3.522)	(64.016)	(29.794)	(24.663)	(64.360)
Observations	1,378	1,378	1,378	1,378	1,378	3,125	3,119	2,871
R-squared	0.221	-	0.222	0.001	0.220	0.062	-	0.155

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex.

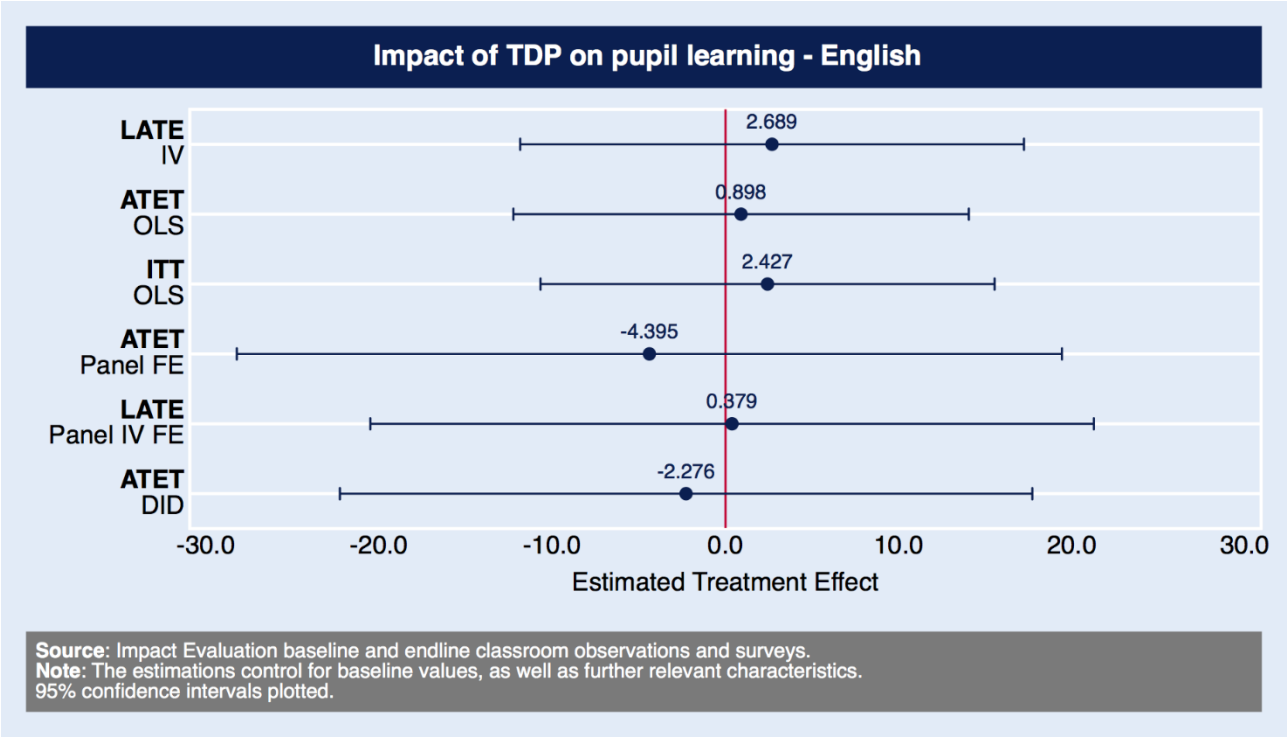
*** p<0.001, ** p<0.01, * p<0.05

Table 3.10: DID results with non-standardised test scores

Outcome measure	Sum of raw scores	Percentage of raw score in relation to total score possible
Model	**DID**	**DID**
Parameter estimated	**ATET**	**ATET**
Treatment receipt	0.188	0.233
	(0.645)	(1.487)
DID	0.719	2.253
	(0.738)	(1.774)
N	2,721	2,721
R-squared	0.162	0.313
Note: Estimates include the full set of covariates, state and LGA fixed effects, and survey weights, with standard errors clustered at school level.		
*** p<0.001, ** p<0.01, * p<0.05		

Pupil learning – English

Figure 3.5: Pupil learning – English



As can be seen from Figure 3.5 and Table 3.11, the main strategy suggests that there is no significant improvement in pupils' test scores in English. This is also confirmed by all robustness checks: no estimate is statistically significant.

Regression results including all covariates can be found in Annex F.6 and F.7.

Table 3.11: Pupil learning – English: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)

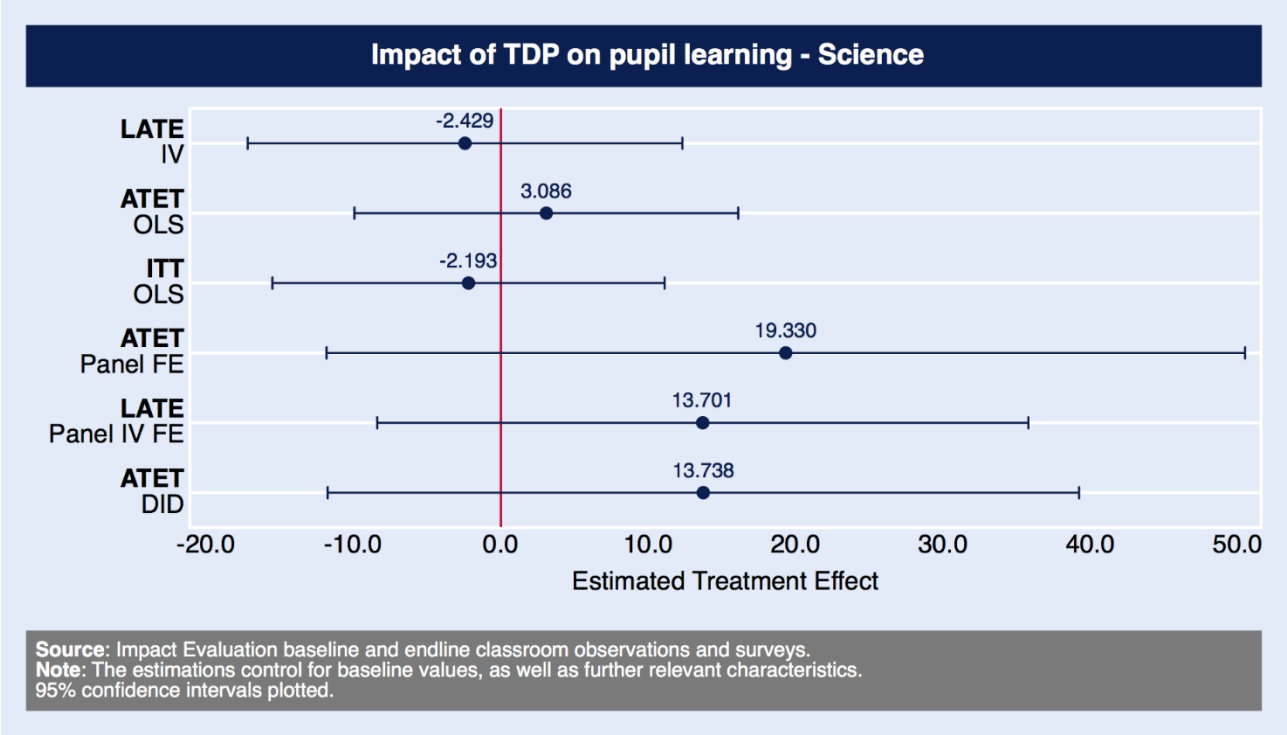
Model	IV	IV LASSO	OLS	OLS LASSO	OLS	Panel FE	Panel IV FE	DID
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	2.689	3.581	0.898	-3.639	-	-4.395	0.379	7.441
	(7.379)	(11.481)	(6.671)	(9.926)		(12.108)	(10.659)	(7.985)
DID	-	-	-	-	-	-	-	-2.276
								(10.146)
Treatment assignment	-	-	-	-	2.427	-	-	-
					(6.652)			
Constant	178.984***	-1.963	102.769	-1.851	101.696	448.321***	455.461***	401.338***
	(51.415)	(3.348)	(58.079)	(3.315)	(57.546)	(27.837)	(26.645)	(74.388)
Observations	1,375	1,375	1,375	1,375	1,375	3,114	3,114	2,867
R-squared	0.232	-	0.232	0.0002	0.232	0.075	-	0.134

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex.

*** p<0.01, ** p<0.05, * p<0.1

Pupil learning – science

Figure 3.6: Pupil learning – science



Similarly to the results for the English test scores, it can be seen from Figure 3.6 and Table 3.12, that the main strategy suggests that there is no significant improvement. This is also confirmed by all robustness checks: no estimate is statistically significant.

Regression results including all covariates can be found in Annex F.8 and F.9.

Table 3.12: Pupil learning – Science: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Model	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel FE**	**Panel IV FE**	**DID**
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	-2.429	7.291	3.086	4.742	-	19.330	13.701	-3.462
	(7.478)	(12.059)	(6.604)	(10.196)		(15.833)	(11.270)	(8.577)
DID	-	-	-	-	-	-	-	13.738
								(12.934)
Treatment	-	-	-	-	-2.193	-	-	-
					(6.749)			
Constant	319.411***	-0.040	227.342***	-0.014	232.348***	510.193***	480.816***	496.289***
	(68.565)	(3.531)	(76.823)	(3.515)	(74.993)	(34.876)	(24.976)	(90.481)
Observations	1,380	1,380	1,380	1,380	1,380	3,121	3,121	2,873
R-squared	0.238	0.0002	0.239	0.0004	0.238	0.070	-	0.140

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex.

*** p<0.01, ** p<0.05, * p<0.1

Proportion of pupils in bottom and top rank – maths

Figure 3.7, Table 3.13, Figure 3.8, and Table 3.14 present results related to the impact of TDP training on the proportion of pupils in the bottom and top bands of performance in maths. Figure 3.7 shows that, when looking at estimations of LATE and ITT derived from estimations using *a priori* covariate selection, TDP seems to have slightly increased the proportion of pupils in the bottom performance band.

However, results derived from other estimation procedures and, in particular, from IV estimations using automated covariate selection presented in (1b) in Table 3.13, do not confirm this finding. Following the hierarchy of evidence established in Section 3.1.5, noting that automated covariate selection lends additional robustness to identifying assumptions, this means that the authors find that there is no conclusive evidence for adverse effects of TDP training implementation in treatment schools. Taking robustness checks into account, the conclusion is that TDP did not have an effect on the proportion of pupils in the bottom performance band in maths.

It should be noted here that, in contrast to other estimation models, the results presented in Volume I on this indicator therefore correspond to the regression specified in (1b) below.

There is also no evidence of an impact on the top performance band (Figure 3.8). This is also confirmed when looking at Table 3.13 and Table 3.14.

Regression results including all covariates can be found in Annexes F.10, F.11, F.12, and F.13.

Figure 3.7: Proportion of pupils in bottom band – maths

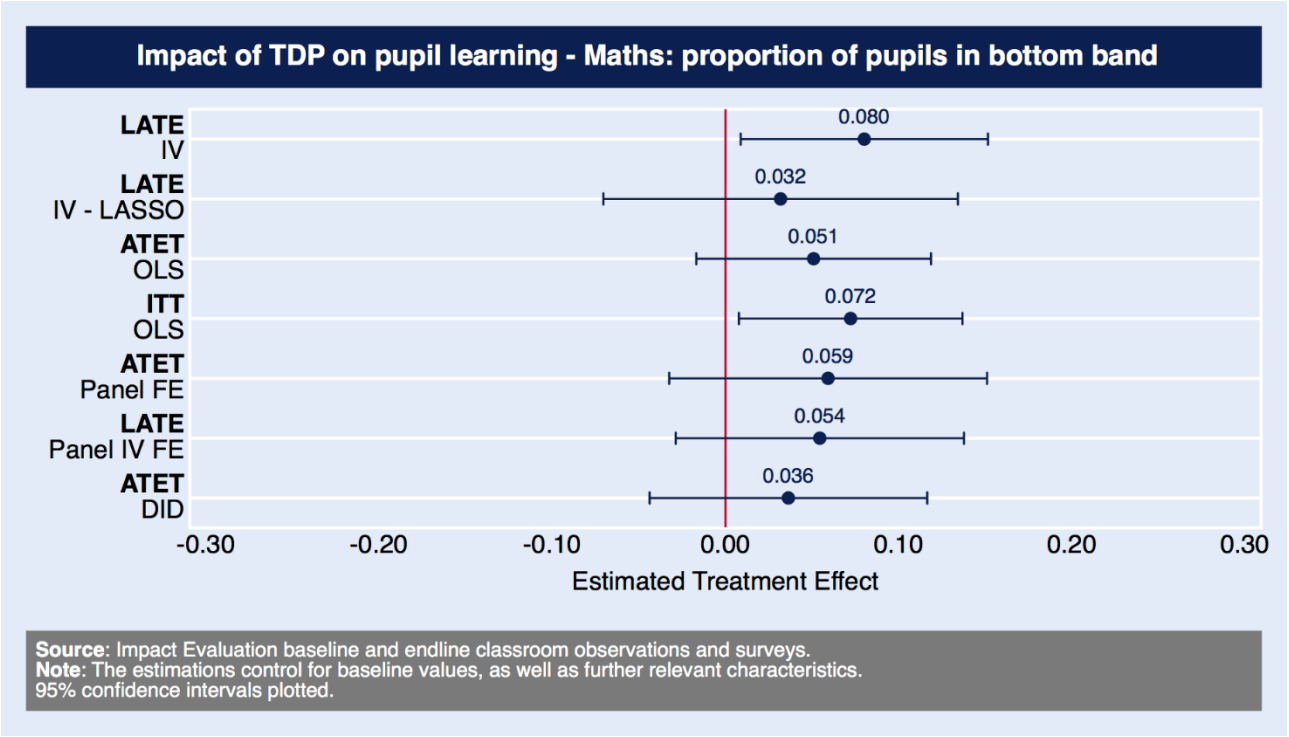


Table 3.13: Proportion of pupils in bottom band – maths: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Model	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel FE**	**Panel IV FE**	**DID**
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	0.080*	0.032	0.051	0.016		0.059	0.054	-0.021
	(0.036)	(0.052)	(0.034)	(0.043)		(0.047)	(0.042)	(0.034)
DID	-	-	-	-	-	-	-	0.036
								(0.041)
Treatment assignment	-	-	-	-	0.072*	-	-	-
					(0.033)			
Constant	0.447*	0.017	0.641**	0.017	0.627**	0.809***	0.813***	0.630**
	(0.210)	(0.015)	(0.225)	(0.015)	(0.225)	(0.116)	(0.087)	(0.201)
Observations	1,377	1,377	1,377	1,377	1,377	3,112	3,106	2,861
R-squared	0.172		0.172		0.174	0.452		0.334

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex.

*** p<0.01, ** p<0.05, * p<0.1

Figure 3.8: Proportion of pupils in top band – maths

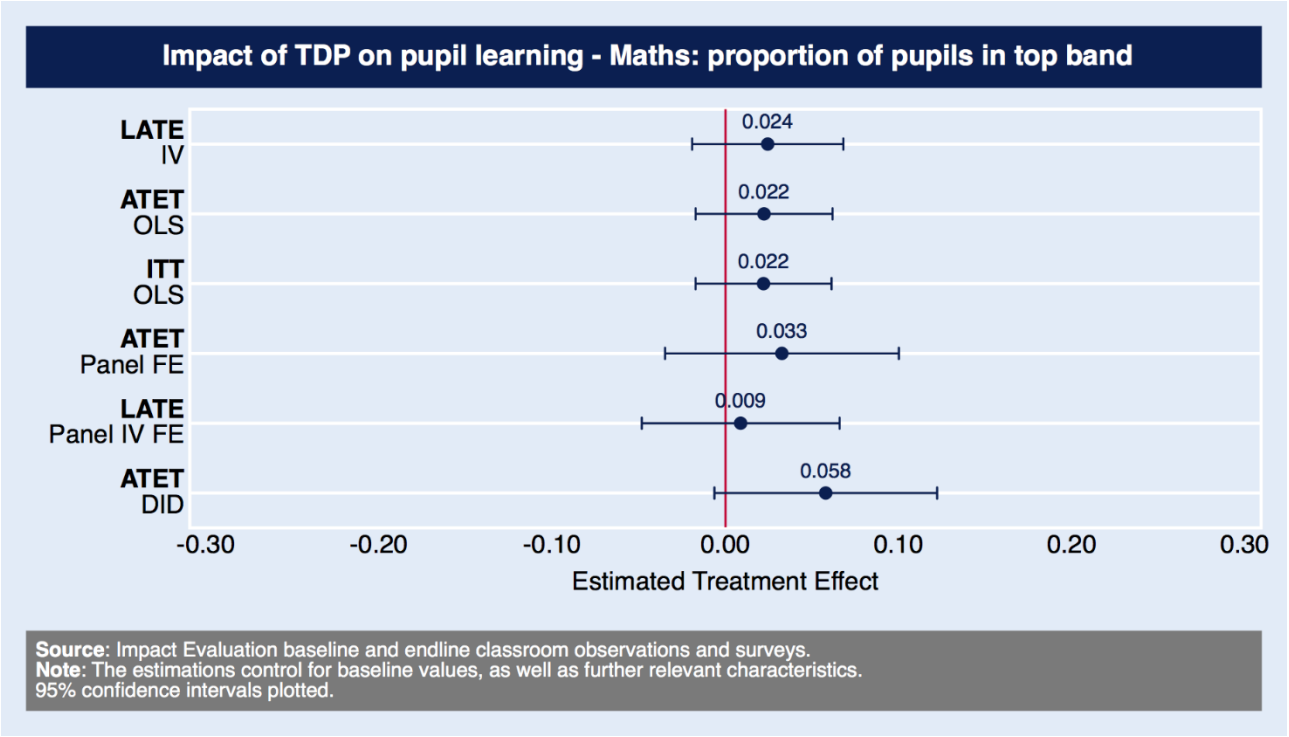


Table 3.14: Proportion of pupils in top band – maths: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Model	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel FE**	**Panel IV FE**	**DID**
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	0.024	-0.007	0.022	-0.021		0.033	0.009	-0.017
	(0.022)	(0.054)	(0.020)	(0.048)		(0.034)	(0.029)	(0.021)
DID	-	-	-	-	-	-	-	0.058
								(0.033)
Treatment assignment	-	-	-	-	0.022	-	-	-
					(0.020)			
Constant	0.023	-0.031*	-0.080	-0.031*	-0.079	-0.128	-0.031	-0.078
	(0.159)	(0.017)	(0.159)	(0.017)	(0.159)	(0.087)	(0.065)	(0.121)
Observations	1,377	1,377	1,377	1,377	1,377	3,106	3,106	2,861
R-squared	0.154	-	0.154	0.000	0.154	0.078		0.087

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex.

*** p<0.01, ** p<0.05, * p<0.1

Proportion of pupils in bottom and top rank – English

Figure 3.9: Proportion of pupils in bottom band – English

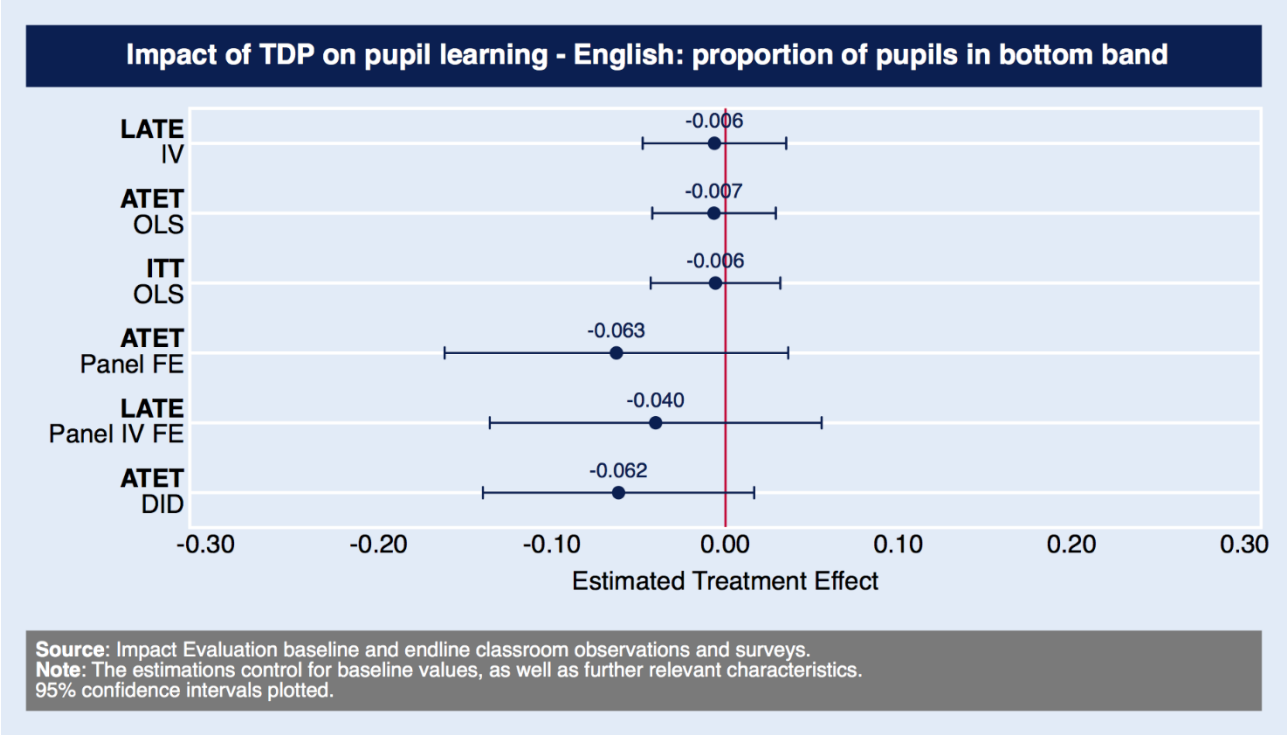


Table 3.15: Proportion of pupils in bottom band – English: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Model	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel FE**	**Panel IV FE**	**DID**
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	-0.006	-0.034	-0.007	-0.002		-0.063	-0.040	0.030
	(0.021)	(0.025)	(0.018)	(0.023)		(0.050)	(0.049)	(0.036)
DID	-	-	-	-	-	-	-	-0.062
								(0.040)
Treatment assignment	-	-	-	-	-0.006	-	-	-
					(0.019)			
Constant	0.649***	0.003	0.674***	0.003	0.673***	0.610***	0.656***	0.649**
	(0.159)	(0.009)	(0.161)	(0.008)	(0.161)	(0.012)	(0.118)	(0.226)
Observations	1,377	1,377	1,377	1,377	1,377	3,114	3,108	2,862
R-squared	0.133	-	0.133	0.000	0.133	0.510		0.371

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex.

*** p<0.01, ** p<0.05, * p<0.1

Figure 3.10: Proportion of pupils in top band – English

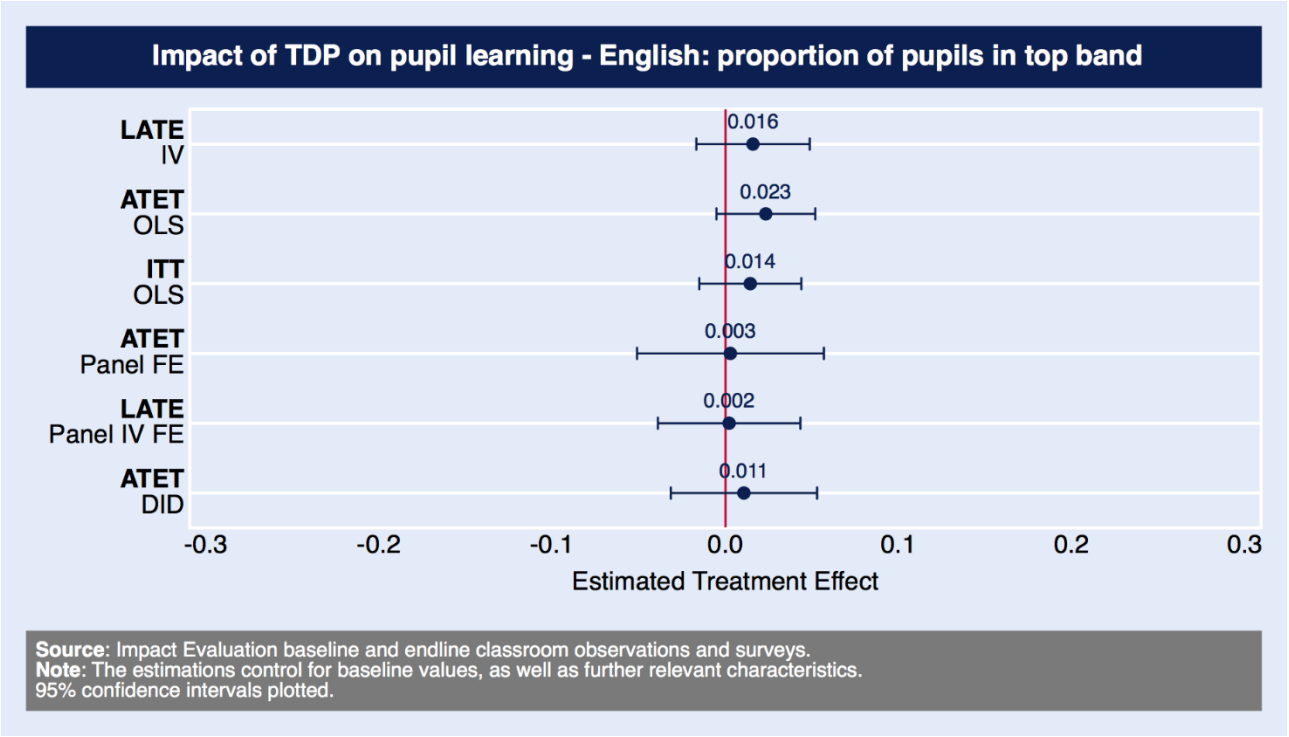


Table 3.16: Proportion of pupils in top band – English: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Model	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel FE**	**Panel IV FE**	**DID**
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	0.016	-0.031	0.023	-0.004	-	0.003	0.002	0.004
	(0.017)	(0.029)	(0.014)	(0.025)		(0.027)	(0.021)	(0.015)
DID	-	-	-	-	-	-	-	0.011
								(0.021)
Treatment assignment	-	-	-	-	0.014	-	-	-
					(0.015)			
Constant	-0.066	-0.011	-0.042	-0.011	-0.032	-0.043	0.053	-0.052
	(0.131)	(0.009)	(0.134)	(0.009)	(0.134)	(0.064)	(0.044)	(0.079)
Observations	1,377	1,377	1,377	1,377	1,377	3,108	3,108	2,862
R-squared	0.109	-	0.109	0.000	0.109	0.033		0.071

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex.

*** p<0.01, ** p<0.05, * p<0.1

As can be seen from Figure 3.9 and Figure 3.10, the main strategy suggests that there is no significant improvement in the performance bands in English. This is also confirmed by all robustness checks: no estimate is statistically significant.

Regression results including all covariates can be found in Annexes F.14, F.15, F.16, and F.17.

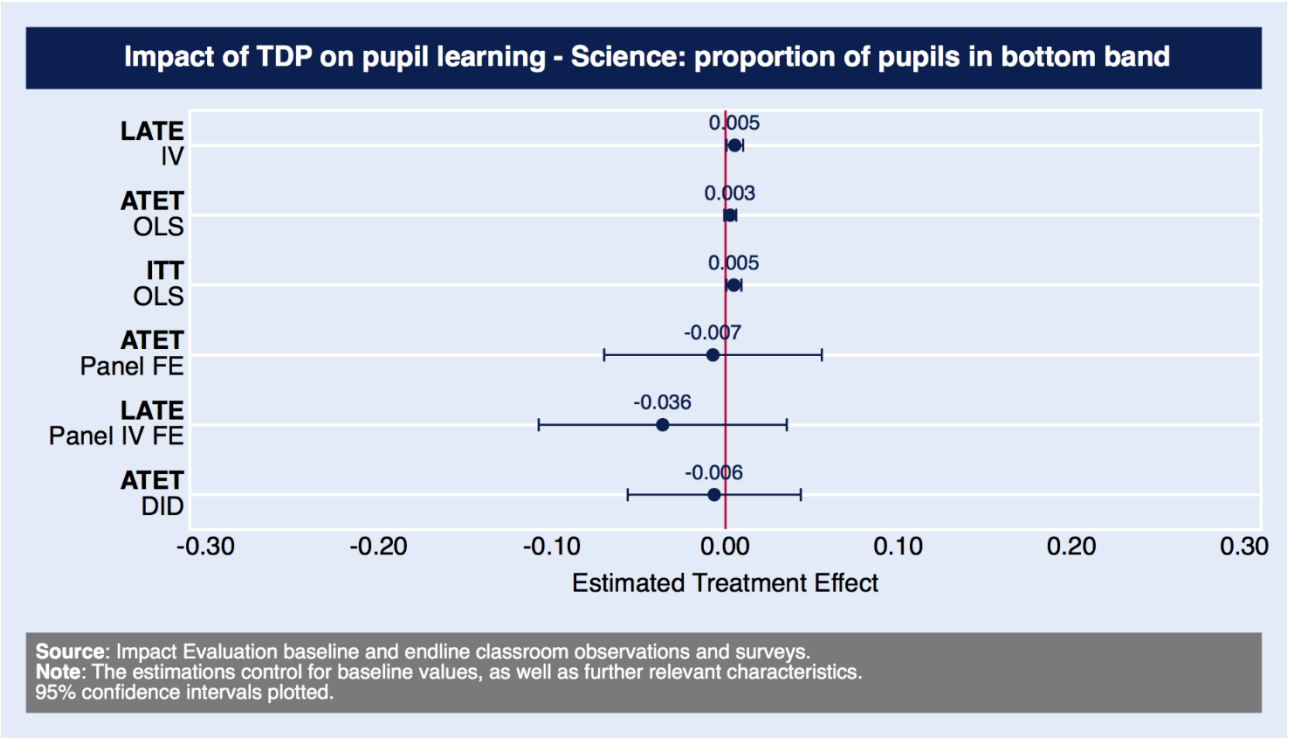
Proportion of pupils in bottom and top rank – science

Figure 3.11 shows a very small increase in the proportion of pupils in the bottom performance band in science . However, it is not possible to conclude that TDP has had an adverse effect, since there are only a few pupils in the bottom band of performance (less than 15).

As can be seen from Figure 3.12, the main strategy suggests that there is no significant improvement in the top performance band in science. This is also confirmed by all robustness checks: no estimate is statistically significant.

Regression results including all covariates can be found in Annexes F.18, F.19, F.20, and F.21.

Figure 3.11: Proportion of pupils in bottom band – science



Source: Impact Evaluation baseline and endline classroom observations and surveys.
Note: The estimations control for baseline values, as well as further relevant characteristics.
95% confidence intervals plotted.

Table 3.17: Proportion of pupils in bottom band – science: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Model	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel FE**	**Panel IV FE**	**DID**
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment receipt	0.005*	-	0.003	-	-	-0.007	-0.036	-0.000
	(0.002)		(0.002)			(0.032)	(0.037)	(0.022)
DID	-	-	-	-	-	-	-	-0.006
								(0.025)
Treatment assignment	-	-	-	-	0.005*	-	-	-
					(0.002)			
Constant	0.095		0.082		0.079	0.212*	0.354***	0.442
	(0.051)		(0.052)		(0.051)	(0.094)	(0.082)	(0.312)
Observations	1,382		1,382		1,382	3,122	3,122	2,875
R-squared	0.028		0.029		0.030	0.209		0.170

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex. Because of the small number of observations of pupils in the bottom performance band (13), running the LASSO regressions was not feasible for this indicator.

*** p<0.01, ** p<0.05, * p<0.1

Figure 3.12: Proportion of pupils in top band – science

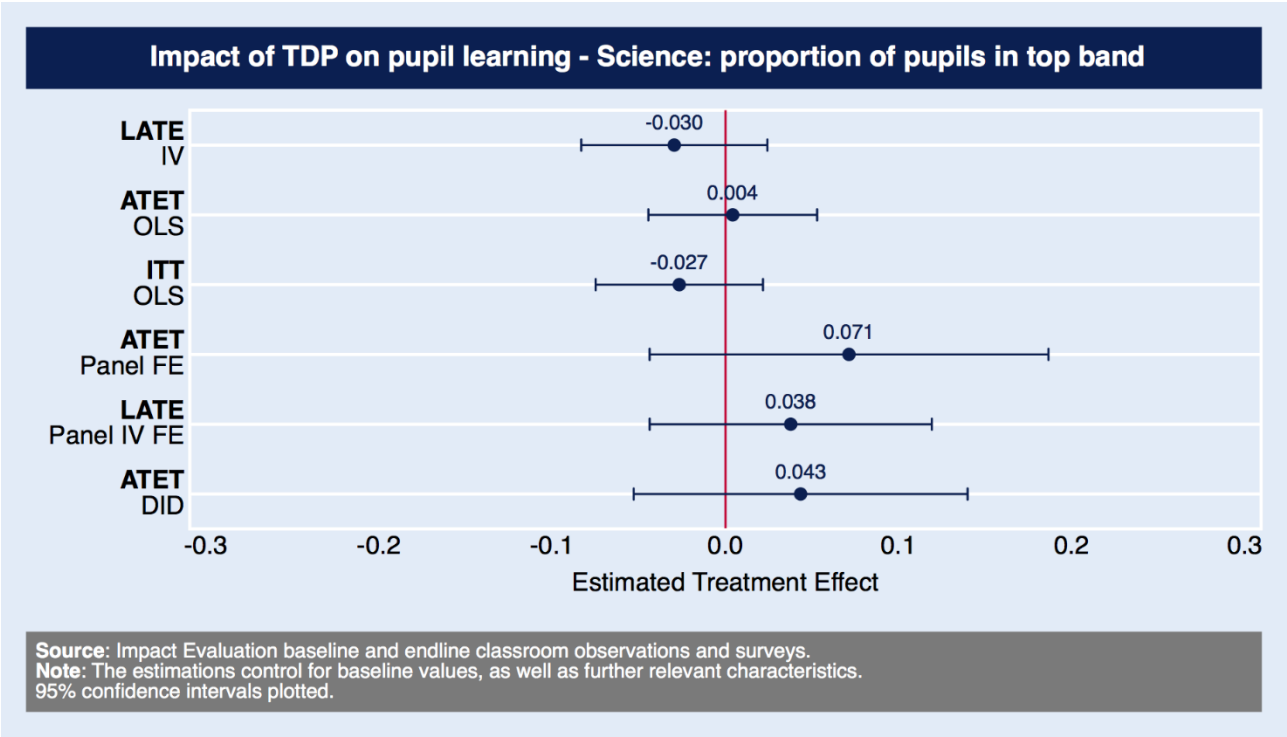


Table 3.18: Proportion of pupils in top band – science: Summary of results

	(1a)	(1b)	(2a)	(2b)	(3)	(4)	(5)	(6)
Model	**IV**	**IV LASSO**	**OLS**	**OLS LASSO**	**OLS**	**Panel FE**	**Panel IV FE**	**DID**
Parameter estimated	**LATE**	**LATE**	**ATET**	**ATET**	**ITT**	**ATET**	**LATE**	**ATET**
Treatment	-0.030	0.001	0.004	-0.000	-	0.071	0.038	-0.032
	(0.027)	(0.002)	(0.025)	(0.002)		(0.059)	(0.042)	(0.034)
DID	-	-	-	-	-	-	-	0.043
								(0.049)
Treatment assignment	-	-	-	-	-0.027	-	-	-
					(0.025)			
Constant	0.263	-0.002***	0.002	-0.002***	0.043	0.175	0.153	0.159
	(0.178)	(0.001)	(0.208)	(0.001)	(0.203)	(0.152)	(0.092)	(0.229)
Observations	1,382	1,382	1,382	1,382	1,382	3,122	3,122	2,875
R-squared	0.191	-	0.193	0.000	0.194	0.045		0.119

Note: Estimates include the full set of covariates, baseline values, state and LGA fixed effects, and survey weights, with standard errors clustered at school level. The complete tables with the results are available in the annex.

*** p<0.01, ** p<0.05, * p<0.1

3.2 Quantitative sampling strategy and weighting procedure

3.2.1 Sampling strategy at baseline

This section briefly summarises the sampling strategy used at baseline for the TDP quantitative survey. The sampling strategy has extensively been described in the TDP baseline report (De *et al*., 2016a), and hence this section provides a shortened description of this process. Please refer to Section 3.1 for a description of the study design that conditioned the sampling strategy for this survey.

At baseline, TDP set out to operate in 14 school clusters per state. There were also an additional 14 control clusters in which schools went through TDP teacher selection processes but where the programme was not intended to operate. Clusters consisted of 12 schools and from every cluster four schools were randomly sampled for the quantitative baseline survey. This yielded a total of 112 schools (56 treatment and 56 control) per state, and 336 schools (168 treatment and 168 control) in total for the three TDP Phase 1 states.

At each school, the (one) head teacher and (three) selected teachers were interviewed. Each teacher and head teacher who teach were also observed while they taught a class. Following the completion of the school survey, all teachers and head teachers (irrespective of whether they teach or not) were administered a TDNA at an examination centre. Note that teachers were not sampled randomly but were pre-selected before assignment of schools to treatment or control groups (see Section 3.1.1).

In order to assess pupil learning levels for the baseline survey, eight pupils among all those who started Primary 3 in September 2014, and who were being taught English, maths, or science by at least one

TDP/control teacher, were randomly selected for the combined English, maths, and scientific literacy learning assessment The pupils were drawn from a sampling frame consisting of all eligible Grade 3 pupils present in school on the day of the survey by data collectors, using a random-number generator programmed into their computer-assisted personal interviewing (CAPI) software.

It should be emphasised here that this sampling strategy allowed the evaluation to produce estimates of indicators that are representative of the schools – and selected teachers and pupils within those schools – across the treatment and control clusters in TDP Phase 1 states.

The TDP survey followed a longitudinal design. The same cohort of teachers, head teachers, and pupils who were surveyed at the baseline were to be surveyed again at endline. The following section describes how this follow-up process – and attrition – affected the sample and data collection for this evaluation.

3.2.2 Attrition and sample size at endline

As described above, for most of the relevant outcome indicators this evaluation aimed to collect longitudinal data, also called panel data, to be used in its quantitative component.[3] This means that the survey was implemented in the same set of schools at baseline and endline, and that within those schools the objective was to collect data from the same units of observation. When relying on longitudinal data, drop-out of units of observation between baseline and endline – attrition – can pose a risk to the impact evaluation strategy, either because it reduces the study's statistical power or because it affects the way in which impact can be identified (see Section 3.1). Each of these issues is explored in detail further below, whereas this section describes how attrition materialised in the present study generally, and how it affected the composition of the study sample.

Attrition can happen due to several reasons, such as, for example, pupils' families moving to a different location to live between baseline and endline surveys, or teachers being posted in new schools. Table 3.19 summarises the number of units from which data were collected at baseline and at endline for the different levels of analysis used in this impact evaluation, and the associated attrition rate. Note that for head teachers the data used in this report consist both of panel data points, where the same head teachers were found in the same schools between baseline and endline, and a non-panelled component, where new head teachers were interviewed at endline as well.

[3] Note that, strictly speaking, some descriptive analyses at the head teacher level are based on an analysis of cross-sectional data where new head teachers that were not interviewed at baseline were interviewed at endline. This section, however, limits its analysis to the longitudinal component of the quantitative data used in this evaluation, except in cases that are explicitly mentioned.

Table 3.19: TDP quantitative survey attrition analysis – panel data

Unit of observation	BL total actual	EL total actual	Attrition rate (%)
Primary schools	330	330	0%
Head teacher interviews[1]	330	134 (panel) 329 (non-panel)	59% (panel) 0% (non-panel)
Teacher interviews	908	447	51%
Classroom observations	1,077	460	57%
TDNA in English, maths, and science	1,158	556	52%
Pupil learning assessment in English, maths, and science	2,575	1,566	39%
Teacher roster and background (excl. teachers who only teach religious studies and are on long leave)	N.A.	329	N.A.
Classroom attendance	N.A.	330	N.A.
Notes: (1) Please note that the endline data on head teachers include both a panel and a non-panel component, where new head teachers were interviewed at endline. Source: TDP baseline and endline quantitative surveys.			

Table 3.19 shows that the achieved school sample size at baseline was 330, and that all of the schools from baseline were revisited at endline.

Head teachers who were present on the day of the survey were interviewed and were also meant to be tested in English, maths, and science. In a small number of cases the head teacher was absent on the day of the survey and instead the acting or the assistant head teacher was interviewed, and some modules only relevant to the head teacher were not administered. At baseline, one head teacher per school was interviewed. At endline, 329 head teachers were interviewed overall. Of those, 134 were the same head teachers as at baseline.

At baseline, the survey interviewed 908 **teachers**, observed 1,077 teachers (and head teachers) in classrooms, and tested 1,158 teachers in English, maths, and science. At endline, the survey re-interviewed 447 teachers (51% attrition), observed 460 teaching in classrooms (57% attrition), and tested 556 (52% attrition).

At baseline, in each school eight Grade 3 **pupils** were meant to be tested and interviewed for the purpose of this impact evaluation. Overall, the baseline survey interviewed and tested 2,575 pupils. At endline, these same pupils were supposed to be tested again, which means that they would be tested in Grade 6. The sample size achieved at endline was 1,566, yielding an attrition rate of 39%.

Two new instruments were developed for the endline survey: a teacher roster instrument and a classroom attendance instrument.

For the purposes of this impact evaluation, attrition for two units of observation were of particular relevance: pupil learning assessments and teacher interviews. To understand attrition dynamics for those units of observation in more detail, this evaluation assessed which characteristics of pupils and teachers at baseline could explain whether individuals dropped out of the sample or not. This means that baseline data were used to compare the group of individuals for which data were also collected at endline (the non-attriters) to the group of individuals for which data were not collected at endline (the attriters).

The purpose of this analysis is to understand whether estimates of characteristics of pupils or teachers at endline can generally be thought of as being produced on a sample that is comparable to the original group of individuals sampled at baseline.

The baseline variables used to examine attrition in such a way are of three types:

- pupil and teacher background characteristics (e.g. gender);
- pupil and teacher outcomes that TDP seeks to influence (e.g. pupil Rasch scores in each subject and teacher TDNA scores in each subject); and
- school characteristics that are likely to be correlated with pupil and teacher behaviour and outcomes (e.g. school size).

Table 3.20 below presents the results of this analysis for pupils, Table 3.22 for teachers. The tables produce descriptive statistics – using data at baseline – for the group of individuals who drop out of the sample (attriters in column 1) and the ones who do not (non-attriters in column 2), and compares them using a t-test (differences are shown in columns 3 and 4). All statistics are produced taking into account the full sampling structure of the data (weights, clustering, stratification) and stars indicate whether differences are statistically significantly different from zero or not. Test statistics are also corrected fur multiple hypotheses testing in column 4, given that these tables are comparing many different indicators at the same time.

The results for **pupils** indicate that pupils who drop out of the sample are mainly older and poorer than pupils who stay in the sample. When not correcting for multiple hypotheses testing, there is also some indication that pupils who drop out of the sample between baseline and endline perform slightly better on the science test at baseline and were in schools with slightly worse infrastructure.

As described above, there are many different reasons for why individual pupils might drop out of the sample for this survey between baseline and endline. One reason is that pupils might advance to junior secondary school (JSS). To examine this further, Table 3.21 plots the distribution of schools in the present sample by the number of pupils who were interviewed at baseline and who proceeded to JSS. This analysis shows that in the majority (67%) of schools no sample pupils advanced to JSS, in 10% of schools one pupil advanced to JSS, in 8% of schools two pupils advanced to JSS, and in 5% three pupils advance to JSS. Notably, in 8% of the sample schools five or more pupils out of the eight sample pupils reportedly advanced to JSS although they should have been in Grade 6 at endline.

The results for **teachers** (Table 3.22) indicate that teachers who dropped out of the sample since baseline were significantly more likely to have NCE qualification or higher and performed significantly better on TDNA tests than individuals whom could be found again at endline. When not adjusting for multiple hypothesis testing, there is also an indication that teachers who dropped out were more likely to be male, more likely to have worked in a school with fewer teachers, and more likely to have worked at a school with higher teacher absenteeism.

Table 3.20: Results – overall pupil attrition since baseline

Variables	Estimates					
	1		2		3	4
	Attriters		Non-attriters		Diff (1-2)	Diff (1-2)
	Mean	N	Mean	N		
Pupils' age in years	9.5	724	8.81	1184	0.69***	0.69***
Pupil is female (%)	40.59	1010	41.98	1565	-1.39	-1.39
State-wise pupil household asset index	0.37	1006	0.66	1559	-0.29***	-0.29**
Asset index – Quintile 1 (%)	17.42	1006	12.98	1559	4.44**	4.44
Asset index – Quintile 2 (%)	14.54	1006	12.45	1559	2.09	2.09
Asset index – Quintile 3 (%)	18.35	1006	18.82	1559	-0.47	-0.47
Asset index – Quintile 4 (%)	23.78	1006	25.45	1559	-1.67	-1.67
Asset index – Quintile 5 (%)	25.92	1006	30.3	1559	-4.38*	-4.38
Rasch score: literacy	501.26	1008	499.29	1563	1.97	1.97
% questions correct: literacy	17.86	1010	16.76	1565	1.1	1.1
Literacy Rasch level 0 (%)	62.19	1008	60.06	1563	2.13	2.13
Literacy Rasch level 1 (%)	32.7	1008	37.51	1563	-4.81	-4.81
Literacy Rasch level 2 (%)	5.11	1008	2.44	1563	2.67	2.67
Rasch score: science	504.5	1008	497.46	1563	7.04	7.04
% questions correct: science	47.39	1010	44.87	1565	2.52*	2.52
Science Rasch level 0 (%)	17.11	1008	18.28	1563	-1.17	-1.17
Science Rasch level 1 (%)	67.55	1008	66.86	1563	0.69	0.69
Science Rasch level 2 (%)	15.34	1008	14.85	1563	0.49	0.49
% questions correct: numeracy	28.07	1010	26.99	1565	1.08	1.08
Numeracy Rasch level 0 (%)	77.23	1008	79.94	1563	-2.71	-2.71
Numeracy Rasch level 1 (%)	15.36	1008	15.32	1563	0.04	0.04
Numeracy Rasch level 2 (%)	7.4	1008	4.74	1563	2.66	2.66
Num. of Primary 1–6 teachers currently employed	20.59	1010	20.74	1565	-0.15	-0.15
Num. of Primary 1–6 pupils currently enrolled	1219.58	997	1198.64	1563	20.94	20.94
Pupil–teacher ratio	68.49	997	65.22	1563	3.27	3.27
Average daily teacher absenteeism (% of teachers absent)	10.72	1010	10.88	1565	-0.16	-0.16
School has electricity supply (%)	21.06	1010	26.57	1565	-5.51*	-5.51
School needs major repairs (%)	90.53	1010	87.12	1565	3.41*	3.41

Source: TDP baseline survey. Notes: (1) Base population: all pupils. (2) Standard errors clustered at the school-level. (3) ***, ** and * correspond to 1%, 5% and 10% significance levels. (4) Column 3 uses unadjusted p-values and column 4 adjusts p-values for multiple hypothesis testing as described in (Sankoh, Huque, and Dubey, 1997)

Table 3.21: Distribution of pupils who advanced to JSS across impact evaluation sample schools

No. of pupils advanced to JSS	Freq.	Percent
0	222	67.27
1	34	10.3
2	26	7.88
3	16	4.85
4	6	1.82
5	12	3.64
6	6	1.82
7	3	0.91
8	5	1.52
	330	100

Implications

Overall, the above results indicate that the evidence for selective attrition is weak among pupils but slightly stronger among teachers. This means that, after attrition, the sample of pupils is comparable to the original sample the evaluation started with at baseline. However, this is less so for teachers.

Table 3.22: Results – overall teacher attrition since baseline

Variables	Weighted estimates					
	1		2		3	4
	Attriters		Non-attriters		Diff (1-2)	Diff (1-2)
	Mean	N	Mean	N		
Teachers' age	36.9	460	37.05	443	-0.15	-0.15
Teacher is female (%)	15.46	461	20.15	447	-4.69**	-4.69
Total teaching experience in ANY school in 2014 (years)	12.57	457	12.02	443	0.55	0.55
Teacher has NCE qualification of above (%)	72.95	461	64.86	447	8.09***	8.09***
Teacher attended teaching-related training in last two years (%)	46.87	461	48.95	446	-2.08	-2.08
Teacher owns a mobile phone (%)	98.58	461	97.62	446	0.96	0.96
Raw TDNA score: maths	45.7	418	42.97	426	2.73**	2.73*
Fully or near-sufficient maths subject knowledge (%)	41.52	418	38.09	426	3.43	3.43
Emerging maths subject knowledge (%)	42.84	418	41.17	426	1.67	1.67
Limited maths subject knowledge (%)	15.64	418	20.74	426	-5.10**	-5.1
Raw TDNA score: English	23.92	418	20.97	426	2.95***	2.95**
Fully or near-sufficient English subject knowledge (%)	6.63	418	3.32	426	3.31**	3.31
Emerging English subject knowledge (%)	40.36	418	37.46	426	2.9	2.9
Limited English subject knowledge (%)	53.01	418	59.22	426	-6.21**	-6.21
Raw TDNA score: science	22.14	418	20.38	426	1.76**	1.76
Fully or near-sufficient science and technology subject knowledge (%)	5.38	418	2.5	426	2.88***	2.88**
Emerging science and technology subject knowledge (%)	31.27	418	32.74	426	-1.47	-1.47
Limited science and technology subject knowledge (%)	63.35	418	64.76	426	-1.41	-1.41
Raw TDNA score: measuring pupil progress	13.72	418	13.08	426	0.64	0.64
Measuring pupil progress: fully or near-sufficient	5.26	418	3.13	426	2.13*	2.13
Measuring pupil progress: emerging	13.29	418	12.66	426	0.63	0.63

Measuring pupil progress: limited	81.46	418	84.21	426	-2.75	-2.75
Num. of Primary 1–6 teachers currently employed	11.92	461	12.86	447	-0.94*	-0.94
Num. of Primary 1–6 pupils currently enrolled	648.17	460	676.55	445	-28.38	-28.38
Pupil–teacher ratio	59.91	460	57.37	445	2.54	2.54
Average daily teacher absenteeism (% of teachers absent)	14.53	461	13.11	447	1.42*	1.42
School has electricity supply (%)	10.54	461	12.48	447	-1.94	-1.94
School needs major repairs (%)	88.28	461	87.12	447	1.16	1.16
Class size during lesson observation	42.52	455	43.98	435	-1.46	-1.46

Source: TDP baseline survey. Notes: (1) Base population: all teachers. (2) Standard errors clustered at the school-level. (3) ***, ** and * correspond to 1%, 5% and 10% significance levels. (4) Column 3 uses unadjusted p-values and column 4 adjusts p-values for multiple hypothesis testing as described in Sankoh *et al.* (1997).

Estimates and descriptive statistics on the sample of teachers presented at endline in this impact evaluation therefore need to be interpreted taking this attrition into account. It is not necessarily the case that they can be assumed to be representative of the population of teachers at baseline. Where necessary, results presented in Volume I and Volume II of this endline report are interpreted accordingly.

It is important to emphasise here that differential attrition between treatment and control groups, that is whether the group of drop-outs differed between the two treatment groups, is analysed in Section 3.1.2.

3.2.3 Minimum detectable effect (MDE) calculations

As described in Section 3.2.2, attrition affects quantitative impact evaluations because it can decrease the statistical power of a study via a reduction in the effective sample size that can be used to detect programme effects. For example, without attrition, the total sample of pupils that could be used in this impact evaluation would be about 2,500 pupils. With attrition, however, this impact evaluation uses data from a sample of about 1,500 pupils to assess whether TDP has had an impact or not.

To assess whether this level of attrition would be problematic in the present case, the evaluation team implemented MDE calculations prior to the implementation of the endline survey. These calculations provide an estimate of the smallest effect on key outcome indicators that could be identified to be statistically significantly different from zero given a certain sample size and sampling structure.

The following sections reproduce the calculations implemented for the TDP endline plan (Cameron *et al.*, 2017). They provide an assessment of how the MDE changes given two different attrition scenarios: a 'good' scenario (A) and a 'bad' scenario (B). Table 3.23 presents these two scenarios for pupils and teachers. Comparing this to the attrition analysis presented in Table 3.19, one can see that the level of attrition that materialised at endline is slightly higher than in scenario B.

Table 3.23: Two attrition scenarios for MDE calculations

	Scenario A	Scenario B
Teachers	37% attrition	50% attrition
Pupils	30% attrition	37% attrition

Calculating changes in the MDE

MDE calculations presented in this section assume a simple DID scenario for the identification of programme effects. Note that these calculations were preformed prior to the implementation of the endline survey.

This DID approach to impact evaluation involves looking at the difference between the difference between treatment and control schools at baseline and the difference between treatment and control schools at endline. That is, the estimate is:

$$\hat{\delta} = (\bar{y}_{TB} - \bar{y}_{CB}) - (\bar{y}_{TE} - \bar{y}_{CE}),$$

where $\bar{y}_{TB}$ is the mean of the indicator in the treatment group at baseline, $\bar{y}_{CB}$ the mean in the control group at baseline, $\bar{y}_{TE}$ the mean in the treatment group at endline and $\bar{y}_{CE}$ the mean in the control group at endline.

The MDE in this context is the minimum value of $\hat{\delta}$ that this evaluation would be able to detect as statistically significantly different from zero.

Calculation of the MDE requires certain assumptions to be made. When the evaluation team planned the baseline survey, an initial set of MDE estimates was made based on assumptions informed by other, earlier evaluations. The calculations presented in this section are based on TDP baseline data, which should make it possible to estimate MDEs more accurately with fewer assumptions.

Specifically, these calculations make assumptions about standard errors, the degree to which indicators are correlated within a school (the intra-class correlation coefficient) and the degree to which indicators are correlated over time within individuals (the intertemporal correlation coefficient (ITC)). The calculations make the conservative assumption that the ITC is 0.4 and use the baseline data to estimate standard errors and intra-class correlation coefficient under different attrition scenarios.

The results are shown in Table 3.24. To illustrate what these mean, consider the indicator '% of time spent in positive interaction'. At baseline, this indicator was about 24% in both treatment and control schools. The MDE is 2.26 in Scenario A and 2.33 in Scenario B. This means that if the proportion of time teachers spent in positive interaction increased to 26.4% in treatment schools at endline, while remaining the same in control schools, this evaluation would be able to attribute a statistically significant effect to TDP, even if attrition is high (50% of teachers have left their schools). This evaluation would not have been able to detect any smaller change.

Pupils' test scores in literacy, numeracy, and science are expressed using scaled item response theory (IRT) scales here. These are constructed in a way that ensures they have an average (mean) of 500 and a standard deviation of 100. Thus, an MDE of around 25 means that this evaluation would be able to detect a difference of 0.25 standard deviations away from the baseline mean.

The table below also presents MDEs for 'raw scores'. These are simply the percentage of questions that the pupil answered correctly. MDEs are around 4 percentage points for literacy, 5 percentage points for numeracy, and 6 percentage points for science.

The table below also includes one school-level indicator: teacher absenteeism according to school records. This indicator is not affected by attrition of teachers or pupils. The MDE shows that the impact evaluation can detect differences of around 3 percentage points between the treatment and control group.

Table 3.24: Estimated MDEs

	Average at baseline	MDE	
		Scenario A	Scenario B
Pupil indicators			
Literacy score (IRT)	500	23.28	23.34
Numeracy score (IRT)	500	25.75	25.83
Science score (IRT)	500	25.85	25.95
Literacy raw score (%)	17.16	4.21	4.23
Numeracy raw score (%)	27.38	4.92	4.94
Science raw score (%)	45.77	6.19	6.22
Teacher indicators			
English score (%)	23.07	2.24	2.31
Maths score (%)	45.08	2.85	2.92
Science score (%)	21.67	2.04	2.09
% of time spent in positive interaction	24.23	2.26	2.33
School indicators			
% of teachers absent on average, according to school records	13.805	2.88	2.88

This MDE analysis shows that MDEs are not very sensitive to the rate of attrition. There is little difference between the two scenarios. Even significantly increased attrition would do little to affect the MDE. For example, if pupil attrition were 70%, the MDE for the literacy IRT score would be 0.24 standard deviations, compared to 0.23 in Scenario A. The reason that the MDE is not very sensitive to higher attrition is that the decrease in the effective sample size is partially offset by the decrease in the design effect due to the smaller number of pupils and teachers per school.

The MDEs are more sensitive to different assumptions about the ITC. The assumption of an ITC of 0.4 is conservative. Higher levels of ITC could increase power, thereby allowing this impact evaluation to identify even smaller changes that could be due to the effect of TDP training. In fact, for the preparation of the TDP evaluation framework, an ITC of 0.8 was assumed. This would reduce the MDEs for pupil test scores to around 17–20 points using IRT scores, or 3–5 percentage points using raw scores. These estimates are in line with the estimates in the evaluation framework of 3.7 percentage points for English and 3.3 percentage points for mathematics.

Comparison to other programmes

A review of 21 'structured pedagogy' programmes – programmes based on changes to curricula or instructional approaches, along with lesson plans and training for teachers – finds, on average, that they improve language scores by 0.23 standard deviations, and maths scores by 0.14 standard deviations (Snilstveit *et al*., 2016). This average includes programmes across a wide range of contexts, and includes some that were not successful (zero or no effect). The largest effect sizes in this review are around 0.8–0.9 standard deviations for languages and 0.3–0.4 standard deviations for maths.

The ESSPIN impact evaluation (Cameron *et al*. 2016) estimated impacts on student test scores of two or more years of ESSPIN intervention at 0.1–0.4 standard deviations. ESSPIN was not found to have an impact

on teacher absenteeism but did result in an increase in the proportion of head teachers who were judged 'effective' of around 5 percentage points, and a similar magnitude of increase in the proportion of classes where teachers were present in the morning.

In light of this past research, it is ambitious but feasible to expect that TDP, if successful, could have an impact on pupil test scores of 0.20–0.25 standard deviations. Similarly, improvements in an indicator such as teacher absenteeism of around 3 percentage points appear to be a reasonable expectation. The present panel sample is sufficiently powered to detect effects of this magnitude, even with the levels of attrition observed.

Conclusion

The results presented above have two main implications. First, as can be seen in Table 3.24, even high levels of attrition affect the MDE in this analytical framework very little. The levels of attrition observed in Table 3.19 will therefore not be problematic in the sense that they have increased the MDE for this study significantly. Second, comparing these results with other studies and programmes, the MDEs for this study are in line with what could reasonably be expected from a three-year education programme. The implicit expectation is that a successful programme will have an impact on learning outcomes of around 0.23–0.27 standard deviations. This is somewhat ambitious compared to the results of past evaluations in Nigeria and elsewhere, but it should also be remembered that this is only equivalent to 3–6 percentage points if the test was scored on a simple percentage scale. This is in line with the MDEs estimated before the baseline, which suggests that original assumptions made at the design stage of this impact evaluation were not too far off. Hence, these seemed to be reasonable expectations for a programme like TDP. The MDEs for teacher and school indicators are around 2–3 percentage points, which appears well within the size of effect that one could expect from a successful intervention, based on previous research.

Note that the impact estimation results presented in Volume I and Volume II of this evaluation report seem to indicate that TDP did not affect pupil learning outcomes. In the context of these MDE calculations, it is important to clarify that this means that it is possible that TDP effects were smaller than what could be statistically identified to be significantly different from zero given this study's set up. The analysis presented in this chapter shows, however, that even with lower levels of attrition than presented in Table 3.19, it would have been unlikely to identify much smaller effects of TDP with much confidence.

3.3 Weighting

This section describes the process of how sampling weights were constructed for the purposes of this impact evaluation. It starts with a brief recap of the sampling design at baseline and endline, and then proceeds to describe how this sampling design and the longitudinal nature of the sample affected the construction of weights.

3.3.1 Background and sample design for TDP panel survey

As described above, the TDP survey was designed as a longitudinal survey with a sample of treatment and control schools to evaluate the impact of TDP in three states of Nigeria: Jigawa, Katsina, and Zamfara. The baseline survey was conducted in 2014, and the sampling design and weighting procedure for baseline purposes is described in Annex D. The endline survey was conducted as a panel survey following the same sample schools, teachers, and pupils selected for the baseline.

The weighting procedures for the TDP panel survey hence depend on the sample design for the TDP baseline survey, which is summarised here. In each of the 14 LGAs where TDP was operational a cluster of 12 control and 12 treatment schools was identified for the purposes of the programme. The primary sampling units (PSUs) within each cluster were the individual schools. The first sampling stage consisted of

randomly selecting a sample of four schools from each of the 14 treatment clusters and 14 control clusters in each state. All of the three (or fewer – in smaller schools) teachers who were supposed to receive TDP training in each sample treatment school, and a corresponding sample of up to four teachers in each control school, were selected to be tested, interviewed, and observed for the TDP baseline survey, as well as the head teacher from each of these sample schools.

For the pupil tests in each treatment school a sample of eight Primary 3 pupils was randomly selected from a list of all the eligible Primary 3 pupils who had a class taught by one of the eligible teachers who were supposed to receive TDP training. In each sample control school eight pupils were also selected from the list of pupils of the sample teachers. In the case of small schools with less than eight eligible Primary 3 pupils, all were selected for the TDP baseline survey.

The stratification of the sampling frame for the TDP baseline survey was by individual treatment or control cluster, since an independent sample of schools was selected from each cluster in the frame. In this case these are not 'clusters' based on the classic sampling terminology; actually, each PSU (school) is a cluster of teachers and pupils. All estimates presented in this volume and in Volume I of this endline report account for clustering at the PSU level – schools in this case.

Given the longitudinal nature of this study, all of the sample treatment and control schools successfully enumerated in the TDP baseline survey were included in the endline TDP panel survey. Within each of these sample schools, all of the baseline sample teachers and pupils who were still at the same school at the time of the endline survey were followed up with corresponding tests, interviews, and observations (in the case of sample teachers). In each sample school the head teacher was included in the TDP survey. See Section 3.2.2 for a description of how the samples at baseline and endline compared.

3.3.2 Weighting procedures for TDP panel (endline) survey

Given that all of the responding sample schools, pupils, and teachers from the TDP baseline survey were included in the endline survey, the basic probabilities and weights are the same between the two rounds of data collection. However, it was necessary to adjust each set of baseline weights for the sample pupils, teachers, and head teachers based on the attrition for the individual tests, interviews, and observations in the endline survey presented in Section 3.2.2.

The calculation of the final weights for the TDP baseline survey is described in Annex D. The different final sets of TDP baseline pupil, teacher, and head teacher weights within each school were compiled in a school-level weighting spreadsheet that was used for calculating all the weights.

In order to calculate the school-level adjustment factors for attrition related to each set of weights, two columns were added in the weighting spreadsheet for each set of weights. The first column was for the 'target' number of sample units (pupils, teachers, or head teacher) within the sample school, corresponding to the number of cases that were successfully completed in the TDP baseline survey for the school. The second column was for the 'actual' number of sample units that were successfully completed in the school for the endline survey.

In a case where the number of 'actual' sample units for a school is zero, there would be no weight since there are no data. However, in this case that school would not be represented in the weighted estimates. In order to compensate for this, the weights for the other sample schools in the same cluster are adjusted at the cluster level, in addition to any within-school weight adjustment for attrition.

The weight adjustment factor for each set of weights is calculated as follows:

$$A_{ci} = \frac{s_c}{s'_c} \times \frac{n_{ci}}{n'_{ci}}$$

where:

A_{ci} = weight adjustment factor for particular sample units (pupils, teachers, or head teachers) and activity (test, interview, or observation) for the i-th sample school in cluster c.

s_c = number of sample schools in cluster c from the TDP baseline survey;

s'_c = number of sample schools in cluster c with at least 1 'actual' (completed) sample unit in the endline data;

n_{ci} = number of 'target' sample units (pupils, teachers, or head teacher) in the i-th sample school in cluster c; and

n'_{ci} = number of 'actual'" sample units in the endline data for the i-th sample school in cluster c.

The basic weight for the head teacher is the same as the school weight since there is only one head teacher per school. However, some sample schools may have an 'actual' count of zero for the head teacher. In this case the corresponding head teacher weight for that school would be 0, but the corresponding weight for the other sample schools in that cluster would be adjusted by the cluster factor to take into account this school with no head teacher data.

This general approach was followed for calculating the adjustment factors for each set of baseline weights. The final weight was calculated as follows:

$$W_{Eci} = W_{Bci} \times A_{ci}$$

where:

W_{Eci} = final weight for particular sample units (pupils, teachers, or head teacher) and activity (test, interview, or observation) for the i-th sample school in cluster c; and

W_{Bci} = final weight from TDP baseline survey for particular sample units (pupils, teachers, or head teacher) and activity (test, interview, or observation) for the i-th sample school in cluster c.

In this case the weight adjustment factor for each set of weights is calculated for the corresponding sample unit and type of activity.

For some sample schools the head teacher for the endline survey was different from the head teacher for the baseline survey. For this reason, a separate longitudinal analysis was carried out for the matched panel head teachers. This involved a different number of 'target' and 'actual' sample units, so a separate set of panel weights was calculated for this longitudinal analysis of the data for matched head teachers.

In a final step, all weights were rescaled and truncated so as to prevent particularly large weights to have a very large influence on the estimates produced for this impact evaluation. Rescaling means that all weights were standardised so that they sum up to the total number of observations in the sample – hence producing sampling weights rather than population-level weights. Second, truncating means that any

weights that – after the scaling process – had values larger than 3 (or 5 in instances where the proportion of weights larger than 3 is more than 5%) were re-coded to having the value of 3 (or 5).

All analyses were conducted taking the resulting weights and the sampling structure into account. This means that standard errors were clustered at school level where necessary, stratification was taken into account, and finite population corrections were applied where necessary.

3.4 Limits of the quantitative approach

There are five key limits to the quantitative approach used in this evaluation:

1. The use of an RCT design meant focusing on groups of schools that were candidates for the first wave of TDP intervention. The survey results are not necessarily representative of all schools at the state level, and should not be treated as such. However, the survey does include schools with a wide range of characteristics, and statistics are likely to be broadly similar to those for the states as a whole. Secondary data sources, such as the National Education Data Survey (NEDS), a household survey conducted in 2015, are used to provide context where appropriate.

2. Effects smaller than the MDE sizes cannot be reliably detected. However, the use of the different impact estimation techniques described in Section 3.1 should help to achieve more precise estimates of impact.

3. The sample sizes are designed to assess overall impact, and allow limited scope for disaggregation. This is particularly an issue for teacher-level statistics, which may have wide confidence intervals due to sample attrition. The evaluation report includes disaggregation of pupil indicators (where sample sizes remain large) and of other key indicators, such as teacher subject knowledge.

4. Several measures depend on reporting by teachers and head teachers, and may be subject to desirability biases. This is the case for teacher motivation measures and responses about the usefulness of TDP training and materials. Teacher absence is measured both from school records and by asking teachers themselves when they were absent. Although this may affect the results, the evaluation team considers it unlikely that there would be systematic differences between treatment and control schools in the extent of this bias, and so do not think that this would bias statistics in favour of, or against, showing the effectiveness or impact of the programme.

5. Measures of teacher behaviour in the classroom depend on classroom observations (see Chapter 7) and are subject to the Hawthorne effect, where being observed may affect teachers' behaviour. The results of classroom observations are perhaps best seen as teachers demonstrating the behaviour that they think is expected of them, and is not necessarily an accurate guide to how teachers teach on a day-to-day basis when not observed.

4 Quantitative data collection

Oxford Policy Management's (OPM's) Nigeria office conducted the endline survey of the TDP impact evaluation. This chapter summarises key points regarding the fieldwork implementation.

4.1 Personnel

The fieldwork was led by the OPM Nigeria office, with support from OPM Oxford. The fieldwork management team comprised six members, including a project manager, fieldwork managers, data manager, and survey coordinators. The team also included several members with very strong computer programming skills in the relevant software (CSPro) in which the instruments were administered.

The overall project manager for the impact evaluation, who is responsible for the content of the instruments, worked closely with the fieldwork team during pre-testing, training, and piloting.

61 trainees were invited to the training, who at the completion of training were assigned into their respective roles of state coordinators, supervisors, and enumerators.

4.2 Fieldwork preparation

The early fieldwork preparation consisted of pre-testing and refining the instruments and protocols, developing the fieldwork manual, and training and piloting.

4.2.1 Pre-test

A full pre-test of all instruments and protocols took place from 18 September to 5 October 2017 in Kaduna State. Members of the OPM fieldwork management team, as well as six data collectors, who would later become the state coordinators during fieldwork, conducted the pre-test. The first seven days were dedicated to training the data collectors while in the latter days 16 schools in eight LGAs were visited to administer and test all the instruments.

The primary objectives of the pre-test were to test the changes to the baseline questionnaires that were made during the questionnaire development phase at endline, and to test the new pupil learning assessment instruments (at baseline Primary 3 pupils were tested, while at endline those same pupils, who were now in Primary 6 were tested and as a result new learning assessment instruments were needed). The pre-test resulted in the refinement of the instruments and data collection protocols, as well as the improvement of the instrument programming in CAPI.

4.2.2 Fieldwork manual

Using the baseline fieldwork manual as a basis, an extensive fieldworker manual was developed that covered an introduction to the project, a description of the fieldwork management and data collection teams, basic guidelines on behaviour and ethical attitudes, the use of CAPI, instructions on fieldwork plans and procedures, an overview of the instruments, as well as a dedicated part on the description of all instruments and protocols.

The manual was updated on an ongoing basis during the training and pilot phase, where updated conventions or additional clarifications were needed. The final version of the manual was printed at the end of the pilot phase and copies were provided to the field teams.

4.2.3 Training and pilot

Data collection training and a field pilot took place from 9 to 21 October 2017. In order to maximise training efficiency and minimise distractions to trainees, the training was conducted in-house at a hotel in Kaduna City, Nigeria. A total of 61 trainees participated in the training. The training was delivered by the fieldwork management team and other consultants from OPM.

The main objective of the training was to ensure that data collectors would be able to master the instruments, understand and correctly implement the fieldwork protocols, and comfortably use CAPI. Supervisors were furthermore trained on their extra responsibilities of data management, fieldwork and financial management, supervision of enumerators, and logistical tasks.

The training combined a variety of methods, including PowerPoint presentations, group sessions, mock interviews, role-plays, and in-class scenarios to ensure that the training was intensive and interactive. The performance of trainees was assessed on an ongoing basis. Participants were quizzed at the beginning of each day to assess their level of understanding of the information they received the previous day, and to inform the training facilitators on areas where participants had knowledge gaps. Furthermore, participants were given daily evaluation forms in order to obtain their feedback on the day's training, with the aim of learning how facilitators could improve their delivery of the training.

Over the course of the training, two pilot surveys were conducted which provided a full-team dress rehearsal. The trainees were closely observed by the training facilitators, who assessed their understanding of the instruments as well as their ability to interact with the respondents, code responses appropriately, and use CAPI and the show cards confidently.

At the end of the training and pilot phase, participants were assigned to their roles as supervisors and enumerators based on their language proficiency, level of understanding of the survey instruments, and its administration. Those who demonstrated desirable leadership and people management skills, in addition to mastery of the instruments and protocols, were appointed team supervisors.

A higher number of data collectors than needed for data collection were invited to and attended the training. This allowed for a selection of the best suited candidates at the end of the training and provided a pool of reserve additional trained staff that could be called upon in case of enumerator attrition during data collection.

4.3 Fieldwork implementation

Data collection commenced on 24 October and ended on 17 November 2017. The teams managed to complete the survey in all 330 schools that were visited at baseline in Katsina, Jigawa, and Zamfara.

4.3.1 Fieldwork model

For the first two days of fieldwork, the teams were collapsed into three teams per state. The state coordinators and fieldwork management team worked closely with the teams to make sure that data collectors were confident and were coding accurately. When it was confirmed that they could work independently, the data collectors were split into six data collection teams per state, with each team composed of one supervisor and two enumerators. Each team completed a school visit in one day.

There were two state coordinators in each of Katsina and Zamfara states, while there were three state coordinators for Jigawa. The state coordinators provided leadership in each state to ensure successful and high-quality fieldwork implementation. State coordinators were responsible for devising implementation plans for their assigned states, and for managing state teams and other survey resources. In addition to

these roles, they provided technical support to state teams, and supported the supervisors to perform their roles: for example, by working with their various teams to address data quality issues identified by the data management team on a day-to-day basis.

Additionally, members of the fieldwork management team were present in every state to provide administrative and technical support, supervision and mentoring, while the data management and IT team provided continuous back-end support to field teams.

4.3.2 Sample achievement

Table 3.19 in Chapter 3 shows the actual number of instruments completed across all schools, against the intended number. At endline, the survey was completed in all 330 schools that were covered at baseline.

The head teacher interview was administered and completed in 329 schools, with the exception of one school where the head teacher was critically ill. During fieldwork, in a number of schools the head teacher was absent on the day of the survey and as a result the acting or assistant head teacher was interviewed instead. At the end of fieldwork, the state coordinators revisited all these schools to re-administer the complete head teacher interview to the head teacher.

In all three states, there was a very high level of attrition for the teachers and pupils that were sampled at baseline:

- Only 1,566 pupils completed the pupil interview at endline out of the 2,575 pupils that were interviewed at baseline. The main reasons the other pupils could not be interviewed were due to: (i) pupils dropping out of school, (ii) pupils advancing to lower secondary, (iii) pupils transferring to another school, (iv) pupils not available on the day of the survey, and (v) pupils having passed away.

- Only 447 teachers completed the interview at endline, out of the 908 teachers that were interviewed at baseline. The main reasons the other teachers could not be interviewed were: (i) teachers transferring to another school, (ii) teachers not available on the day of the survey, (iii) teachers retiring from service, (iv) teachers being promoted to head teachers, (v) teachers having passed away, and (vi) teachers quitting the service.

Some of the steps taken to combat attrition included:

- revisits: state coordinators and teams revisited schools in order to conduct missing interviews for teachers and pupils that were unavailable on the first day of visit; and

- calling pupils from home: for the pupils that lived around the school community but were temporarily absent from school on the day of the visit, data collectors worked with the head teachers and teachers in order to ask pupils to come to the school to take the test if they were capable of doing so.

4.4 Quality control and data checking protocols

Several mechanisms were put in place in order to ensure high quality of the data collected during the survey. These are briefly summarised in turn below.

4.4.1 Selection and supervision of data collectors

Each enumerator was supervised by the training team during the training, piloting, and first week of data collection. This allowed a well-informed selection of enumerators and their allocation to roles matching individual strengths and weaknesses.

4.4.2 CAPI built-in routing and validations

One important quality control mechanism in CAPI surveys is the use of automatic routing and checking rules built into the CAPI questionnaires that flag simple errors during the interview, i.e. early enough for them be corrected during the interview. In addition to having automatic skip patterns built into the design in order to eliminate errors resulting from wrong skips, the CAPI validations also checked for missing fields, out-of-range values, and inconsistencies within instruments. The latter checks if any related information collected in different questions of the instrument are consistent. A warning or error message was given if an entry was out of range, inconsistent, or left empty. The enumerator would then try to understand why a warning or error message was showing up and reconfirm the information with the respondent.

4.4.3 Live observations

Live interviews were observed by state coordinators, the field manager, and members of the fieldwork management team. Any errors detected during observations were noted and discussed with the teams at the daily de-brief.

4.4.4 Daily and weekly reporting from the field

At the end of each working day, supervisors collected all interview files from their team members and transmitted the data to the data manager. The supervisors also sent their daily achievements to a WhatsApp group that was created for the survey. These reports were checked for consistency, completeness, and correctness by the field management team and they were cross-checked with the data received by the data manager. Any missing or inaccurate data identified were communicated to the data collection team.

Additionally, a Google tracking sheet was developed that was used by the teams at the end of each work day to fill in their achievements and comments for each school. The information provided in the Google sheet was cross-checked with the information provided on WhatsApp to ensure accuracy. Whenever there were discrepancies, the survey management team contacted the state coordinators to clarify.

At the end of each working week, the state coordinators collated all achievements and challenges recorded by their teams over the course of the week and shared those with the field management team. This allowed the field management team to keep track of weekly achievements and to ensure that there were no missing data.

4.4.5 Excel dashboard

An Excel dashboard was also created by the fieldwork management team to track the uploaded data. This information was cross-checked with the Google sheets. The dashboard was also used to check any inconsistent or missing data. In the event of missing data, the field team was informed, and revisits were conducted to ensure data completeness.

4.4.6 Secondary consistency checks and cleaning

OPM furthermore exploited a key advantage of CAPI surveys, the immediate availability of data, by running a range of secondary consistency checks across all data on a daily basis in Stata. Data received from the field were exported to Stata the following day, and a range of do-files were run to assess the consistency and completeness of the data, and to make corrections if necessary. The checks comprised the following:

- Completeness and ID uniqueness: during this process, the data manager ensured that all the data reported in the daily field update were consistent with the data captured and sent in by the teams. Unique identification in each dataset and sound linkage between the datasets were also paramount and had to be checked on a daily basis.

- Consistency and out-of-range checks: a range of consistency and out-of-range checks that had not been included in the CAPI instruments were programmed into a checking Stata do-file. The data manager ran the checking do-file on a daily basis on the latest cleaned data. This returned a list of potential issues which the data manager would then investigate, undertaking the necessary cleaning actions, if any. On a daily basis, all errors flagged were collated and shared with the survey management team in the field, as well as with the state coordinators and supervisors, so that the errors could be discussed with the data collectors. The purpose of these errors was to monitor the performance of data collectors and provide them with feedback to help them improve.

4.5 Fieldwork challenges

In addition to the high attrition of pupils and teachers, the TDP endline survey experienced a few other challenges:

- **Pupil identification**: There were minor challenges with identifying pupils whose photos were not taken at baseline, especially in classes with pupils that had similar names. In order to resolve this challenge, data collectors probed pupils to find out if they were sampled at baseline, and confirmed with teachers when necessary.

- **Head teachers attempting to replace sampled pupils**: Some head teachers across the three states attempted to replace sampled pupils with other pupils in the school whom they anticipated would do better on the test, possibly in an effort to improve the school ratings. In a few cases, there were also attempts to invite pupils who had graduated and were currently in JSSs to take the pupil test in place of the sampled pupil. It was perceived that the head teachers misunderstood the purpose of the pupil test, and felt that their school's performance was being assessed. In order to prevent this occurrence, data collectors were diligent in verifying the identities of pupils by comparing the presented pupil with the photos of the pupils from baseline that they had access to on the tablets.

- **Security challenges**: Security challenges were minimal across the three states. However, there was one case of a security threat in a school in Zamfara State. Only the head teacher and teachers were present on the day of the visit due to an attack on the community by cattle rustlers two days before the visit. The research team visited the school based on an assurance of safety given by the head teacher. The parents of some pupils were killed in the attack. As a result, the school was void of pupils on the day of the visit. The head teacher and teacher interviews were conducted; however, no classroom observations or pupil tests were conducted.

- **Difficult terrains**: About 5% of schools were located in hard-to-reach communities across the states. Alternative means of transportation, such as motorcycles, were used to access these difficult-to-reach schools. As a result, this challenge did not affect overall fieldwork.

4.6 Preparation of TDP teachers for TDNA assessments

The evaluation team used the same TDNA at both baseline and endline. While this would possibly have given teachers an advantage in terms of familiarity with the tests, the advantage was expected to be similar in control and treatment schools, and so it was expected that it would not create bias in the test results. Using the same TDNA simplifies analysis by ensuring comparability between baseline and endline in terms of the skills being examined, but it does involve some risk in relation to leakage of test papers.

TDP's state government partners in the three states where this evaluation was conducted (Jigawa, Katsina, and Zamfara) conducted readiness sessions in preparation for the TDNA among teachers in TDP treatment schools who were part of the evaluation. These sessions used actual copies of the TDNA, and in Katsina the head teacher of each school was able to take away a copy of the test paper.

The evaluation team were unaware of this preparation in advance. As the preparation was given only to teachers in treatment schools, it is likely to bias TDNA results in favour of the treatment group. After discovering teachers were using filled-in copies of the test paper in Katsina during fieldwork, the team took two steps to avoid a recurrence:

- **Stricter test environments:** Teachers were required to take the test simultaneously, with all data collectors present to closely monitor the process. Cell phones, books, and other materials were disallowed in the testing venue. Additionally, other teachers were not allowed to gain access to the testing venue.

- **Test revision:** In the third week of fieldwork, the maths section of the TDNA was revised and the new version was administered to sampled schools.

At the pre-analysis stage, three steps were taken to examine whether there was a bias due to the TDNA preparation:

- The analysis looked at whether TDNA scores were higher among teachers in the treatment schools during the first week of data collection – before the stricter test environment was introduced – than during the rest of the data collection. No statistically significant effect was found.

- The analysis looked at whether teachers in Katsina had a particular advantage during the first week, given that Katsina head teachers were able to take away copies of the TDNA paper from the preparation sessions. No such effect was found.

- The analysis looked at whether teachers in treatment schools had significantly lower scores when they took the revised version of the test, introduced during the third week of fieldwork. The analysis did find such an effect. Teachers in treatment schools had significantly lower maths scores if they took the revised version of the test, while teachers in control schools did not. This suggests a positive effect of the preparation (and a bias in the TDNA results) of 4–6 percentage points for maths (depending on the regression model specification used to examine this).

Thus, there was some ambiguity in the analysis of the possible bias due to test preparation, with no effect found of tightened supervision but some effect found of revising the TDNA maths section. The latter suggests that maths scores may be biased upwards among treatment school teachers by 4–6 percentage points, but does not yield any estimate of bias in the English or science sections.

The potential effect of test preparation should be taken into account when reading the TDNA results in Volume I, Chapter 6. The outcome of the analysis there is that there was no significant change over time in TDNA scores in either treatment or control schools. It is possible that there would have been a *worsening* over time in treatment schools, and a negative effect of the intervention on test scores, had it not been for

the test preparation. However, given the data available, the team is only able to conclude that there was no improvement in TDNA scores and no positive impact.

5 Qualitative research design and data collection

5.1 Sampling

5.1.1 Selection of schools

In each state, three treatment schools in different LGAs were sampled; no control schools were sampled for the qualitative research data collection. The research team used a stratified purposive sampling approach, based on composite indicators, informed by the baseline survey. The theory-based assumptions underlying this approach are that there is: a) a positive correlation between pupil performance (literacy and numeracy) and TDP impact, b) an inverse relationship between class size (at baseline) and TDP impact, and c) a negative correlation between teacher absenteeism (at baseline).

A composite score was used to list all treatment schools in each of the three states in descending order. In a second stage, schools with head teacher transfers over the last three years were excluded from the sample, based on the information collected in the validation survey. A caveat here is that in some cases head teachers were transferred after the completion of the validation survey. In all but one school the validation survey provided reliable data for this sampling stage.

In a final stage, due to security policies for the research team, schools in very remote areas, as well as those that are difficult to access, were excluded from the sample. All sampled schools are within a radius of three hours of travel time to large urban centres.

Subsequently, the best performing schools, the worst performing schools, as well as medium-performing schools (at baseline) were grouped together. In each of the three strata, three schools were randomly selected, the other two were labelled as replacement schools. The research team relied on one replacement school in Zamfara, due to poor road conditions faced during data collection.

5.1.2 Selection of respondents

Pupils for a school transect walk and most significant change exercise

A group of Primary 6 pupils were selected for the participatory research activities (no more than six pupils). Where possible, the research team included an equal number of female and male pupils. Pupils self-selected into taking part in the research, often with encouragement by the teachers or head teachers. Where possible, a new group of pupils was selected for the second research activity.

Teachers for interview, for observation, for Proportional Piling proportional piling exercise, for Timeline exercise

Only teachers of the TDP-relevant subjects – maths, English and science – were selected for classroom observations and subsequent discussions. Participatory activities were led with a group of no more than six teachers across all subjects. The timeline activity was based on TDP teachers only, while non-TDP teachers and those recently transferred could be included in the proportional piling activity. For these activities, head teachers either selected teachers or teachers volunteered to participate.

Teacher facilitators for timeline exercise

Teacher facilitators (TFs) self-selected into participating in the timeline activity; their participation was dependent on the distance of their home to the place of facilitation.

School-Based Management Committee members for most significant change exercise

School-Based Management Committee (SBMC) members self-selected into participating in the discussion and most significant change (MSC) activity, encouraged by the head teacher, or the head of the SBMC. The head teacher was excluded from participating in both the discussion and the MSC activity, to avoid bias.

5.2 Data collection tools

The qualitative research was conducted using a range of research tools, such as interviews and participatory tools (MSC activity, proportional piling activity, timeline tool, force field analysis, and school transect walks). In addition, the research team used observations (school-level and classroom-level) and follow-up discussions.

Research guides were developed for all research tools to guide data collection and allow for inter-school and inter-state comparison. All interviews were conducted in a semi-structured manner to accommodate context-specific probing and the flexibility to explore unanticipated or new themes. Observations were guided by a framework of pre-determined factors and aspects to guide the assessment of teaching practice and behaviour, as well as school environment, but provided room for interpretation and explanation.

These research techniques are described in more detail in the text below. Table 5.1 indicates how the tools were designed to answer the evaluation questions.

Table 5.1: Qualitative research respondents and techniques used at endline

Research techniques/tools	Respondents	Evaluation questions (references are to the evaluation matrix in Annex A)
In each school, the qualitative study employed the following techniques:		
School transect walk[4], MSC	Pupils	• Has TDP improved teacher effectiveness in the classroom? Confounding and contributing contextual factors. • What factors facilitated or inhibited TDP's achievement of its outcomes? (Effe-1, Effe-13, Effe-16, Effe-20, Effe-21)
Head teacher interview	Head teachers	• How does TDP's organisational and management setup facilitate delivery? • What factors facilitated or hindered TDP's achievement of its outcomes and impacts? (To examine, in particular, support for teachers and teacher motivation). • Are there any unanticipated (positive or negative) TDP impact? (Effi-17 to Effi-22, Effe-12 to Effe-15, Effe-19 to Effe-22, Im-3, Im-8, Im-14, Su-24, Su- 29, Su-31, Su-34)
Teacher interview, timeline tool[5], proportional piling	Teachers	• How does TDP's organisational and management setup facilitate delivery? • Has TDP improved teacher effectiveness in the classroom? Confounding and contributing contextual factors. • What factors facilitated or inhibited TDP's achievement of its outcomes and impacts? (To examine, in particular,

[4] https://siteresources.worldbank.org/EXTTOPPSISOU/.../1_Transect_walk.pdf
[5] siteresources.worldbank.org/EXTTOPPSISOU/Resources/.../6_Time_line.pdf

		teacher motivation and the role of materials provided to teachers.) • Are there any unanticipated (positive or negative) TDP impacts? (Effe-12, Effe-13 to Effe-15, Im-5, Im-8, Im-11, Im-12, Im-13, Im-14)
MSC	SBMCs	• Has TDP improved teacher effectiveness in the classroom? • What factors facilitated or inhibited TDP's achievement of its outcomes and impacts? • Are there any unanticipated (positive or negative) TDP impacts? • To examine confounding and contributing contextual factors, including those relating to the home and community, class size, and peer support. • (Effe-13, Effe-27, Im-3, Im-5, Im-8, Im-11, Im-12, Im-13)
Lesson observation and school observation	Schools	• Has TDP improved teacher effectiveness in the classroom? • What factors facilitated or inhibited TDP's achievement of its outcomes and impacts? • Are there any unanticipated (positive or negative) TDP impacts? • To examine confounding and contributing contextual factors, including those relating to the home and community, class size, and peer support. (Effe-13, Effe-27, Im-3, Im-5, Im-8, Im-11, Im-12, Im-13)
In addition to school-level data collection, the qualitative study employed the following techniques and tools with key representatives from government and TDP:		
Timeline tool	TFs	• Were TDP outputs achieved on time and in full? • How does TDP's organisational and management setup facilitate delivery? (This includes a focus on accountability mechanisms and teacher support.) • Has TDP improved teacher effectiveness in the classroom? • What factors facilitated or inhibited TDP's achievement of its outcomes and impacts? • Are there any unanticipated (positive or negative) TDP impacts? • Is the TDP model applied sustainably in TDP schools, in other schools in TDP states, and in Nigeria? • Are TDP's partner institutions the appropriate institutional homes for a teacher INSET model? Are these partners open to these partnerships? Do these partners have the capacity to engage in the INSET model on a sustained basis? • Do the SUBEBs and Local Government Education Authorities (LGEAs) have the incentives and capacity to maintain, support, and renew the teacher educator teams without support from TDP?

		(Effe-12, Effe-13, Effe-14, Effe-15, Effe-16, Effe-24, Im-3, Im-15, Im-16, Su-23, Su-24, Su-26, Su-31, Su-32)
Teacher Development Team (TDT) interview	TDTs	• Were TDP outputs achieved on time and in full? • How does TDP's organisational and management setup facilitate delivery? (This includes a focus on accountability mechanisms and teacher support.) • Has TDP improved teacher effectiveness in the classroom? • What factors facilitated or inhibited TDP's achievement of its outcomes and impacts? • Are there any unanticipated (positive or negative) TDP impacts? • Is the TDP model applied sustainably in TDP schools, in other schools in TDP states, and in Nigeria? • Are TDP's partner institutions the appropriate institutional homes for a teacher INSET model? Are these partners open to these partnerships? Do these partners have the capacity to engage in the INSET model on a sustained basis? • Do the SUBEBs and LGEAs have the incentives and capacity to maintain, support and renew the teacher educator teams without support from TDP? (Effe-12, Effe-13, Effe-14, Effe-15, Effe-16, Effe-24, Im-3, Im-15, Im-16, Su-23, Su-24, Su-26, Su-31, Su-32)
Interviews and force field analysis	Programme staff	• How does TDP's organisational and management setup facilitate delivery? • Has TDP improved teacher effectiveness in the classroom? • What factors facilitated or inhibited TDP's achievement of its outcomes and impacts? • Is the TDP model applied sustainably in TDP schools, in other schools in TDP states, and in Nigeria? • Are TDP's partner institutions the appropriate institutional homes for a teacher INSET model? Are these partners open to these partnerships? Do these partners have the capacity to engage in the INSET model on a sustained basis? • Do the SUBEBs and LGEAs have the incentives and capacity to maintain, support, and renew the teacher educator teams without support from TDP? (Su-23 to Su-26, Su-30, Su-31)
Key informant interviews (KIIs)/in-depth interviews	SUBEBs, other government staff, civil society organisation[6] representatives*	• How does TDP's organisational and management setup facilitate delivery? (This will focus on state and local government level.) • Has TDP improved teacher effectiveness in the classroom? • What factors facilitated or inhibited TDP's achievement of its outcomes and impacts?

[6] In Katsina only.

		• Are TDP's partner institutions the appropriate institutional homes for a teacher INSET model? Are these partners open to these partnerships? Do these partners have the capacity to engage in the INSET model on a sustained basis? • Do the SUBEBs and LGEAs have the incentives and capacity to maintain, support, and renew the teacher educator teams without support from TDP? (Effi-17, Effe-12, Effe-27, Im-3, Im-6, Su-27 to Su-29)

5.2.1 Description of the tools

MSC tool

The MSC tool is a collective participatory exercise where individuals record in writing or verbally the most significant change that they perceive to have occurred during a given period, while also identifying the most likely cause of the change. The purpose of this exercise is to generate insights into people's perception of the most important change in the school and what they see the cause of that change to be. A well-facilitated discussion will generate insights for respondents themselves, in this case SBMC members in particular, who take the time to reflect on their situation and learn from each other. Furthermore, the story-telling exercise is designed to reveal what respondents in a homogenous group of people regard as significant, and why.

In each school, the MSC exercise was conducted with a group of pupils and with a group of SBMC members. At the beginning of the exercise, groups of respondents were encouraged individually to reflect upon and discuss their story of the MSC in the school over the past three years. One by one, respondents were asked to recount their story, while the rest of the group listened and then discussed the story. Once each respondent's story had been told to the group, the group was then asked to select a single MSC story to put forward with the most likely cause. The group was asked to reach consensus in order to facilitate a discussion around why they felt the chosen story represented the MSC.

Transect walk

The school transect walk tool was designed to use the physical environment of the school and classroom to prompt pupils to discuss their perceptions and experiences of being at school and learning. Specifically, the purpose of the tool was to understand how elements of the school environment facilitate or inhibit learning based on pupils' experiences. The tool was also designed to provide information about the pupils' priorities and preferred spaces. The tool kept pupils engaged by walking with them around the school, which also provided researchers with the opportunity to observe pupils' non-verbal reactions to different places in the school, as well as their verbal explanations.

A group of six P6 pupils were asked to participate in the transect walk at each school. The exercise began with pupils drawing their school with the help of the researcher, to produce a map which was used to guide the walk. Once pupils had agreed on the drawing of the map, they were asked to slowly walk around the school, guiding the researcher to each of the places identified on the map. Stopping at each point, the researcher asked pupils about their experiences of each space, with a particular focus on what helps them to learn or makes it difficult to learn.

Head teacher interview

In each school, semi-structured discussions were conducted with head teachers to understand their experiences of TDP, including processes of delivery of the programme, any leadership and management

training received, their engagement in supporting and mentoring teachers, and their experiences and perceptions of TFs' support visits.

These were used to understand the head teachers' perceptions of the main drivers of or challenges for improved learning in the school, and their perceptions of how teachers in the school have been engaging with the training and materials provided to them through TDP. Discussions also covered head teachers' perspectives of the MSC in the school. In schools where the head teacher was new, they would be asked about their experience with the TDP training, but there would also be a focus on the challenges of being transferred to a new school and if they faced problems in applying their skills after the transfer.

Teacher interview

One teacher in each school was interviewed individually. This interview was designed to help the research team to understand the teacher's perceptions of the main drivers of improved learning in the school or challenges faced to improve learning, as well as their engagement with and perception of TDP's training and the materials provided. The tool was designed as a semi-structured discussion to explore the challenges teachers face and to better understand their motivation. This last part of the interview was closely linked to the proportional piling exercise.

Timeline tool

One timeline activity was conducted with a group of teachers in each school, as well as an additional timeline in each state, which was conducted with TFs. With teachers, the group timeline exercise was used to understand key activities and changes to TDP from the teachers' point of view. The group of teachers included teachers who had become part of the programme at different points in time and therefore had different experiences of the programme. The purpose of the activity was to understand revisions to TDP from the perspective of the teachers – to understand what the impact of these changes have been on teachers (both positive and negative) and how they teach.

With the group of TFs, the timeline exercise was used in order to understand key activities and unanticipated changes to the TDP model from the revisions made after the first year of the pilot to the endline, from the perspectives of teachers and TFs, and to facilitate discussions about the rationale for these changes, as well as their impact (both positive and negative) on the efficiency, effectiveness, impact, and sustainability of the programme. Additionally, the timeline was able to explore perceived changes to teaching and learning that were seen in the schools as a result of the TDP intervention.

With each group, the exercise began with the researcher presenting a large sheet of paper with a long line on it representing time from 2013 to 2017 and beyond. Respondents were asked to take some time to think about changes they had observed and subsequently feed these back to the group, placing positive changes above the line and negative changes below. This sheet was then used to facilitate discussions about these events, whether they represent a fulfilment of the original plan or a deviation from it, and what the respondents perceived to be the result of these events or changes for efficiency, effectiveness, impact, and sustainability.

Proportional piling exercise

The proportional piling exercise was designed as a group activity for teachers to explore potential factors that motivate or demotivate them from teaching. Teachers were asked to list all the factors that they consider motivating and demotivating. Once listed, teachers were asked to score each factor by allocating 50 counters against the motivating or demotivating factors identified by the group. Throughout the exercise, the researchers asked questions and allowed a discussion to take place in order to understand what was meant by each motivating or demotivating factor, as well as to understand the scoring and

ranking of factors. The proportion of counters assigned to each factor was recorded by the research team and this has been used as part of the analysis of what motivates teachers.

Lesson observation

Researchers used this tool to conduct observations during P6 maths and/or English lessons at the sampled schools to add further depth and insight to the qualitative findings, and to inform the case studies developed for each school sampled. The tool was designed in two parts and began with a lesson observation, followed by a discussion with the teacher to understand how the lesson was planned and why the lesson was conducted in the way it was. The lesson observation tool comprised a semi-structured observation form, which led researchers to focus on different aspects of the lesson, including the classroom environment (e.g. seating arrangement), teacher practice, teacher pedagogy, use of materials, and classroom management. The discussion, which followed the observation, was designed to understand why teachers used certain teaching practices and to understand the way in which TDP had affected the way in which they conduct their lessons.

School observation

In order to gather data about school infrastructure and the school environment, researchers conducted a school observation while at the school. The observation form directed researchers' attention to the surroundings of the school in terms of accessibility, safety, proximity to the community, and amenities such as toilets, water sources, playground etc. Secondly, researchers observed the inside of the school – particularly inside classrooms – to better understand the infrastructure, and conducted a final observation in the head teacher's office. With explicit permission from the head teacher, researchers took photographs of the school, which were used to understand the school context.

TDT interview

The TDT interview was an open-ended interview conducted with one TDT in each state to understand their role, its evolution during the programme, and their impression of the programme. Interviews probed further into the challenges faced in, and successes of, the programme.

SUBEB, TDP state team, and civil society organisation interviews

In order to understand the perspectives of programme staff regarding the delivery, evolution, and perceived impacts of TDP, as well as its longer-term sustainability, the research team conducted KIIs with staff at TDP in each of the states engaged in strategic decision making and in programme delivery.

Force field analysis

In order to capture the perspectives of TDP's central programme management, we used a force field analysis to map their understanding of the sustainability of the programme. The activity began with the team describing what, in an ideal world, a sustainable version of the TDP would look like for them. They were then asked to list what steps they would need to put in place to ensure that the TDP is sustainable, and list gaps or weaknesses which would work against sustainability. They were then asked to think about the strengths and weaknesses of TDP as currently implemented and map these onto a chart with concentric circles drawn on it, with TDP at the centre. Respondents were asked to place strengths or weaknesses which were in close proximity to themselves in circles closer to the centre to indicate that these are issues that either TDP or a close partner work on. Similarly, they were asked to place strengths and weaknesses which fall outside of the direct control of TDP further away from the centre. The force field analysis was conducted ahead of some of the KIIs with other stakeholders to allow priorities for sustainability to be followed up with the identified external stakeholders in the states.

5.3 Fieldwork

5.3.1 Interviewer training

Six local researchers and one national team leader with previous experience in qualitative data collection were recruited for the research. The qualitative research training took place in Abuja during 22–25 November 2017. The research teams stayed at the training venue and training took place between 9.00 and 18.00 over the four-day period. OPM staff conducted the training with inputs from EDOREN staff based in Abuja. The training included an introduction to TDP and the endline research, as well as a presentation of the results from the baseline and midline studies, and the implementation review. The training also covered sessions on qualitative research and ethical practices and protocols. In particular, time was taken to ensure that researchers were familiar with sensitivities related to working with children. In-depth workshops were used to introduce research tools, allowing for application and discussion. Based on the input of local researchers, the research guides for all tools, as well as the fieldwork manual, were finalised. The training concluded with a session on safety and security in the three states, and on protocols for anonymising and labelling transcripts, photos, and voice recordings.

5.3.2 The fieldwork teams

The six researchers were divided into three teams, each comprising one female and one male researcher. The teams were overseen by a national team leader who travelled with the teams to each of the three states. In each team, one researcher conducted the interview while the other took notes; roles were switched throughout the research. Additionally, where possible[7], interviews were recorded and later transcribed. International researchers joined the fieldwork teams at the start of the project in Jigawa State, to provide support and ensure the quality and direction of the fieldwork.

During the research, the composition of the teams was changed, to counter researcher fatigue and promote exchange and discussion among the teams.

5.3.1 Sequencing of research activity

The following table provides an indicative sequence of research activities at the schools. The timelines made provision for each school to be visited for three consecutive days. These activities were deliberately left flexible for research teams to accommodate the availability of teachers, head teachers, and community members.

Table 5.2: Sequencing of qualitative research

Day	Research activity
DAY 1	Introductions at state SUBEB and LGEA – Education Secretary, SUBEB interviews
	Introductions at school, head teacher interview
	Classroom observation
	Teacher interview
DAY 2	Transect walk with students
	Proportional piling with teachers

[7] During the transect walk, for example, it was not possible to record the discussion as the group of pupils were moving around the school throughout the discussion. In this case, detailed notes were used in the analysis.

	MSC with SBMC members
Day 3	MSC with students
	Timeline exercise with teachers
	TDT/TF timeline exercise and interview

5.3.2 Debriefs and team checks in the field

The researchers, led by the team leader, conducted debriefs every evening, attended by the entire research team and one international researcher to compare notes and to attempt an initial analysis of the day's findings. Emerging trends were identified and research gaps or areas of interest were highlighted by the group, which facilitated learning across teams. This also allowed international researchers to provide support or viewpoints, and to monitor the progress of the fieldwork remotely.

The fieldwork was concluded by a three-day long debrief workshop with two international researchers and the local research team. This facilitated a first extraction of themes and topics, as well as comparative analysis between schools and states. The workshop enabled researchers to discuss and capture emerging trends observed in the schools and across the states. It also allowed international researchers to clarify any questions and ensure a balanced review and initial analysis.

5.3.3 Transcription and translating

Audio research files were transcribed and translated by the EDOREN team (from Hausa into English). Transcriptions were literal and entire conversations were captured between researchers and respondents. In certain instances, Hausa words were left in the text, where literal translation was not possible. In one case the transcription was unclear and was sent back for corrections and revisions.

5.4 Analysis

5.4.1 Approach to analysis

The approach to analysing the qualitative data was based on thematic analysis, an inductive approach that requires a higher degree of involvement and interpretation by the researcher. Thematic analysis rejects a quantitative approach to analysing qualitative data (such as frequency or cluster analysis) and instead focuses on interpretation of the accounts shared by respondents in order to identify and examine themes.

Two of the three researchers involved in the analysis and coding had accompanied research teams to schools in Jigawa. Using the evaluation questions, as well as initial themes that emerged during the daily debriefs and final analysis workshop, a coding framework was developed to guide the initial stages of analysis. The three international researchers developed individual codebooks, which were then compared and integrated for final coding stages.

This process enabled the researchers to triangulate their interpretation of the text and codes, and discuss any differences in understanding. The coding framework, or 'node tree', comprised descriptive codes known as nodes and sub-nodes, against which data from the KIIs, participatory discussions, and observations could be organised according to emergent themes. The initial node tree is illustrated in Table 5.3. Parent nodes tend to categorise the data according to high-level categories, while child nodes were named to capture the essence of the data to be coded against them. Specifically, parent nodes comprise the broad thematic areas that the evaluation was to cover, including teacher motivation and effectiveness, pupil motivation and

performance, the head teacher and SBMC and TDP training and materials, among others. Child nodes allowed for more specific analysis and for researchers to analyse emergent trends.

Before starting to code, the research team discussed the nodes to ensure there was a common understanding and to try to ensure consistency between each state. However, as Saldana (2013) notes, 'all coding is a judgement call' and it is important to be cognisant of the fact that we all bring our 'subjectives, our personalities, our predispositions' to the coding process. Therefore, throughout the coding and analysis process, the researchers kept their own biases in mind. Since transcripts from the three states were coded in parallel, each by one member of the research team, new nodes were added to the initial coding framework as themes emerged from a close reading of the transcripts.

Table 5.3: Node tree

Parent node	Child node
Teacher motivation and performance	Salary
	Student absenteeism
	Distance
	Resources and infrastructure
	Support
	TDP or TDP related
	Teaching
	Other
Teacher quality	Qualifications and knowledge
	Training and support
Pupil motivation and performance	Resources and infrastructure
	Teacher absenteeism
	Work and labour
	TDP or TDP related
	Learning
	Peer support
Head teacher	Motivation
	Training
SBMC	Role of parents
	Support to head teacher
	Support to school infrastructure
	Support to teachers
Support and monitoring of teachers	TDP
	Other
School environment	Infrastructure
	Classrooms
	Location
	Safety
	Other
TDP meta	Sustainability
	Unintended consequences
TDP training	Positive feedback / used

	Negative feedback / not used
TDP material	Positive feedback / used
	Negative feedback / not used
Language	Language of instruction
	Feasibility of English for P4–P6

Coding was carried out using NVivo version 11 and was undertaken concurrently by international researchers, each coding transcripts from a particular state. The use of NVivo facilitated analysis by providing a useful way of storing and organising the data but analysis was carried out by the research team, and manually coded, rather than using tools in the programme, such as word frequency functions. Once the data were coded, the team had an initial meeting to discuss emergent themes and findings from each of the states. This began the third stage of analysis, where each team member took responsibility for individual evaluation questions and areas and began to delve into the data. Each researcher then revisited the data by looking at the coded files from each state in order to assess how findings compared across the states and to bring them together.

In order to analyse the data, the researchers were required to interpret the data collected by identifying and describing the explicit and implicit ideas that emerged from the coding process. Researchers considered the strength of the data in light of the particular respondent – including their knowledge of the subject and their incentives to answer questions in a certain way – as well as the context. In order to ensure rigour, the researchers sought to triangulate the findings with other qualitative data and the quantitative findings. Following the analysis of the qualitative data, the quantitative and qualitative research teams met for a full-day workshop to discuss and compare the findings and piece together the emerging story. This gave the team a chance to validate the findings and discuss areas of disagreement. This process of triangulating and bringing together the quantitative and qualitative findings was iterative throughout the analysis and writing process.

5.5 Limitations of the qualitative research component

5.5.1 Qualitative sampling and generalisability

The qualitative research component is based on a relatively small sample of schools. As explained in the evaluation framework (EDOREN, 2014), it is a major challenge to generate results that have wider application beyond the schools visited by the research team. The need to specify a sample in advance meant that the team was not able to guarantee that all possibilities will have been exhausted and all processes understood by the end of the study.

Sampling for the qualitative component was purposive: its aim was to include schools with particular characteristics, rather than being representative of all schools in the three states. Representativeness was further affected by the inability of the research teams to access all or any LGA as they wished, due to security concerns. Thus, more remote schools further from the state capital were less likely to ultimately feature in the final sample of schools. Although the evaluation framework suggested qualitative research in control schools, it was decided given limited resources and a need to explain what has helped or hindered TDP from working, to focus on the treatment schools.

The qualitative research is not designed to produce results that are generalisable in the same sense as quantitative data. Generalisability derives from linking qualitative findings to the TOC and to findings from the quantitative research. In some cases, it may be more appropriate to talk of whether findings are transferable rather than generalisable: in order to construct an argument that a finding in one setting is likely

to apply in another, an understanding is needed of what aspects of context were important for that finding. The risk of visiting atypical schools and gaining an incorrect or incomplete understanding of the relevant processes remained, but was mitigated by visiting several schools in different states and by paying close attention to ways in which the context of each school may be atypical.

5.5.2 Structured and unstructured research instruments

Qualitative research uses instruments (interviews or discussion guides, observational tools, etc.) which are purposefully designed to have less structure, to allow for more open dialogue and to discover themes and responses that structured quantitative tools may find harder to uncover. This leaves scope for the interviewer and respondent to shape the research. For example, the interviewer can ask further questions that occur to him or her, in response to an interesting or unexpected response from the research participants. This can help capture impacts or explanations that were not anticipated, but makes qualitative research hard to reproduce and subject to researcher bias. While the dialogue may be unstructured, the researchers applied structured methods in recording and analysing the discussion: for example, through the application of structured templates organised by thematic categories for note-taking, and use of the evaluation matrix (Annex A) to provide a framework for analysing the research. A reflective approach, with a mixed team of international and local researchers, and discussion about findings at the end of each day was intended to reduce bias from individual researchers. However, qualitative research inevitably involves greater implication of the researchers' (and participants') own perspectives. This more embodied, personal approach compared to quantitative research should be seen as a strength as well as a limitation.

5.5.3 Sensitive issues

Some issues may have been sensitive for participants to discuss in the schools. For example, head teachers and teachers were likely to be nervous at first about revealing potentially negative, critical, or self-damaging views or information. A longer-term engagement with the participants would have helped to gain their trust, but this was not possible given resource constraints and the need to visit a range of schools in the three states within a limited period of time. Nevertheless, their continued presence in the school for three days gave researchers time to gain some trust from the participants, and also allowed time for informal talk and observation, as well as more structured discussions. The answers to some questions posed to teachers may have involved overt or tacit criticism of the head teacher or other teachers, if they were frank, and it has to be acknowledged therefore that researchers may not have received comprehensive and honest answers on these questions. Wherever the environment allowed, researchers ensured that discussions took place in private so that only the participants in each part of the research were present. Researchers also used strict codes of data confidentiality and reassured participants that their responses would not be shared more widely. Although the qualitative research remains limited in its ability to explore sensitive issues, it is likely to be stronger in this respect than quantitative research, where there is very little time for researchers to gain the trust of the respondents or to probe evasive or incomplete answers.

5.5.4 Language issues

Interviews were conducted in Hausa and both interviewers and note-takers were Hausa speakers. The research team performed preliminary analysis of the findings each day, based on their notes and recollection, but the main analysis was conducted by the international researchers upon their return to the UK. The conversations were recorded and the transcripts translated for full analysis of the data. However, there is some risk in this process of inaccurate or incomplete translation. The inclusion in the team of a majority of researchers who were fluent in both Hausa and English was essential in managing this risk. Hausa-speaking researchers checked transcripts in English and Hausa to ensure accurate translation. Researchers were mindful of the need for precision in interpretation in the field, and carried out discussion among each other to ensure a clear shared understanding. An added complication was variation and dialects in the Hausa

language found in different states (particularly Zamfara). In order to mitigate this risk, to the extent possible, national researchers were selected on the basis of having extensive experience of working in the study areas.

5.5.5 Lack of parental input

The qualitative study did not speak with parents of students, and so lacks their perspective on improvements in teaching and learning in their children's school. It also results in a gap in information about challenges to schooling from the parent's perspective. To a small degree, this was mitigated by speaking with SBMC members who also had children in the school they represented. However, the research addressed them and asked their opinion primarily as their position as SBMC members.

6 Pupil test design and analysis

6.1 Pupil test design

At baseline, pupils in Primary 3 (towards the beginning of the school year) were sampled and tested in English, maths and science, using a test that targeted the skills and knowledge expected to be acquired during Primary 1 and Primary 2. At endline, the same cohort of pupils were tested again, where they could be found in the same school. If the pupils had progressed without any grade repetition, they would be in Primary 6. Again, they were tested towards the beginning of the school year.

It was therefore necessary to develop a test that would be appropriate for the pupils, capturing knowledge and skills gained during Primary 3, 4, and 5 as well retaining some items more appropriate to lower grades.

Pupils needed to be tested in all three subjects – English, maths, and science. The test length was limited to around one hour maximum by both practical fieldwork considerations and the need to avoid distressing or exhausting pupils with a very long test.

The test was developed using the Nigerian primary syllabus and primary school textbooks to guide content appropriate for each grade. A test with approximately one-third too many items was piloted in order to examine the functioning of different test items and to select those for the final tests. The test was designed to be administered largely in Hausa, except for parts of the English test that tested understanding of spoken English, maths questions involving word problems, and some science questions that could not be administered without using written language (e.g. labelling a flower).

The final scale consisted of 17 items for English, 19 items for maths, and 18 items for science.

Table 6.1: TDP test design parameters

Purpose	Measure change in pupil learning levels over three years since start of P3
Length	Maximum one hour / approximately 45 questions in total (15 questions per subject). Each question can have multiple parts
Administration mode	One-on-one
Item type / style	Oral administration, oral response (including gesture) Oral administration, written response Written administration, oral response Written administration, written response

6.2 Pupil test analysis

The psychometric properties of the English, maths, and science tests were analysed within the Rasch measurement model. The more restrictive Rasch model was also applied for the baseline analysis.

Items found to be less effective in assessing the traits are discussed. Items are flagged if they do not fit the Rasch model. The following sections describe the reliability and targeting of items in the English, maths, and science sections.

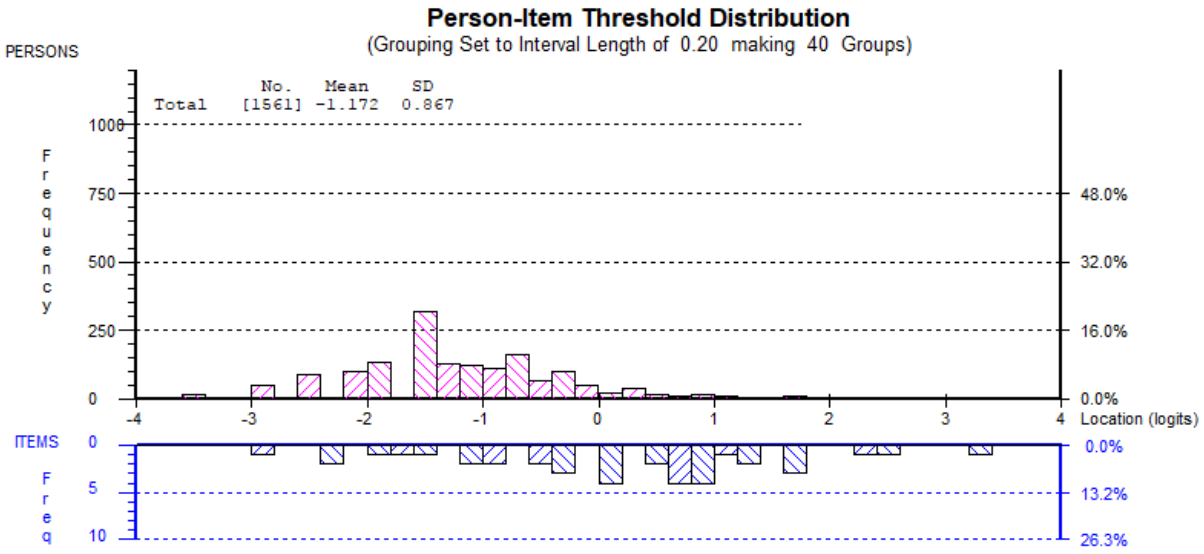
6.2.1 English

Overall fit of the items to the Rasch model: The Person Separation Index (r_β = 0.72) and Cronbach's alpha (α = 0.76) indicate that, overall, the test distinguishes well between high and low performers.

Item targeting. Figure 6.1 plots the overall distribution of person ability and item difficulty estimates, on the same continuum. More difficult items and more proficient students are located to the right of the continuum; easier items and lower ability students are located towards the left. The distribution shows whether there are sufficient items for low and high performing learners. A test is well targeted if the average of item difficulties is about the same as the average of the students' abilities and the item difficulties are evenly spread across the ability distribution (Organisation for Economic Co-operation and Development (OECD), 2012: 222).

The English test is very hard for the students. The distribution is substantially skewed to the left, with a majority of students at the lower end of the scale, below -1.0 logits. The mean person location is far from the mean item location of zero.

Figure 6.1: Distribution of person ability in relation to item difficulty, literacy test

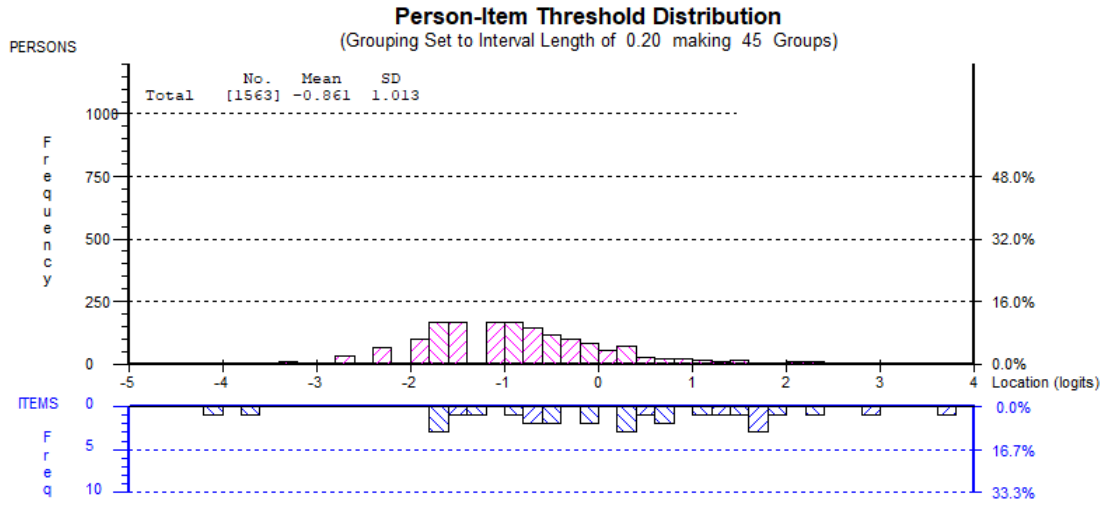


6.2.2 Maths

Overall fit of the items to the Rasch model: The Person Separation Index (r_β = 0.76) and Cronbach's alpha (α = 0.77) show that the test distinguishes well between high and low performers.

Item targeting: Figure 6.2 shows a wide spread of items relative to persons. However, the person ability distribution is skewed to the left, and there is a gap in the continuum of items at the lower end of the scale.

Figure 6.2: Distribution of person ability in relation to item difficulty, maths test

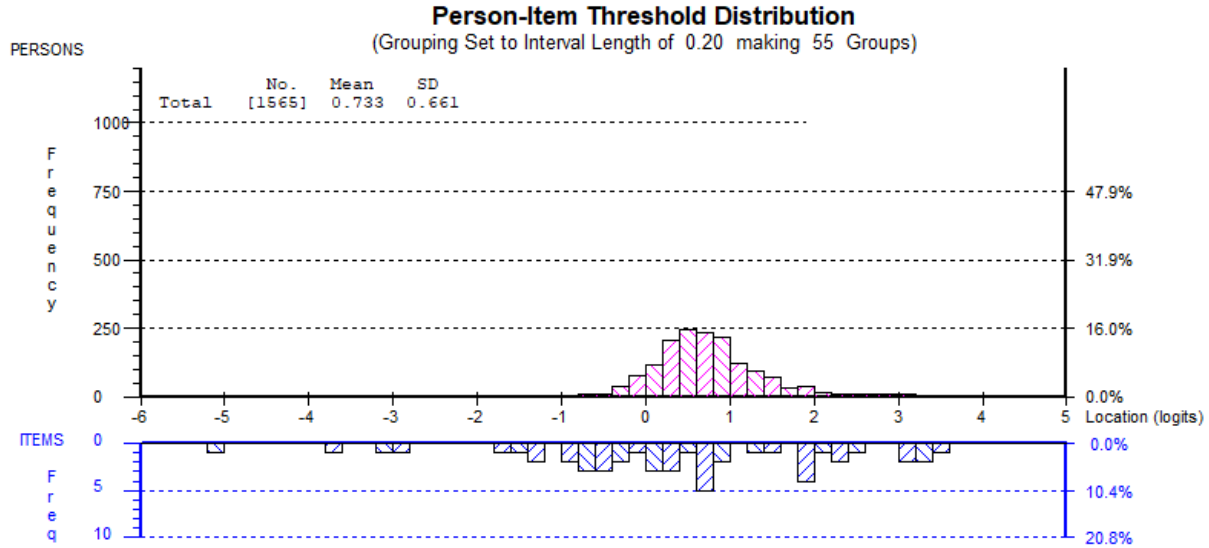


6.2.3 Science

Overall fit of the items to the Rasch model: The Person Separation Index (r_β = 0.75) and Cronbach's alpha (α = 0.68) show that the test distinguishes well between high and low performers.

Item targeting: The students spread well along the continuum of items. As shown in Figure 6.3, the distribution is skewed to the right and the mean person location is above the mean item location.

Figure 6.3: Distribution of person ability in relation to item difficulty, science test



6.2.4 Creation of bands

Based on the test results, pupils were placed into the same three bands as those used at baseline: 'pre' (pre-literacy, pre-numeracy, or pre-science), 'emerging', and 'functional'. As the English test contained a number of shared items between baseline and endline, the tests were first 'equated' so that baseline and endline scores could be placed on a single scale. Exactly the same cut-offs could then be used as at

baseline. For maths, there were no identical items between baseline and endline, so cut-offs were assigned based on comparisons of the difficulty level of items. For science, there were some similar items between baseline and endline, but they had been modified at endline and appeared to behave differently, so cut-offs were again assigned based on comparisons of the difficulty level of items.

7 Classroom observation behaviour descriptors and scoring scheme

TDP provides teacher in-service training on pedagogical skills and subject knowledge, with the aim of increasing the use of effective teaching practices in the classroom, which in turn is expected to raise pupil learning levels (see Chapter 3 in Volume I for more on TDP). To allow measurement of any changes over time in teaching practices that are attributable to TDP, the impact evaluation quantitative surveys examined teaching practices at baseline and again at endline, using timed classroom observations.[8]

To assess the amount of total lesson time teachers engaged in positive interaction with the pupils in the classroom, the enumerators recorded what the teacher and pupils said and did at three-minute intervals for the duration of the observed lesson for up to 45 minutes (see Box 2). One lesson was observed for each selected teacher and head teachers who taught Primary 1 to Primary 6 classes, and this could cover any grade and subject the teacher or head teacher teaches.

At the end of the classroom observation, the enumerators also recorded the availability and use of materials, including those provided by TDP; whether multiple grades were being taught in the same class; if there was co-teaching; and the use of praise and reprimands (for findings on this see Chapter 6 in Volume I).

Box 2: Classroom observation excerpt from the enumerator manual

To determine if the lesson has begun, you can use the following cues:

- New teacher enters the classroom and is greeted by the pupils.
- New or existing teacher erases writing on the blackboard from the previous lesson and writes the new subject, date, etc. on the blackboard.
- This could be accompanied by the teacher saying, 'Today we will learn about…'

Once you have selected 'Ready to start' on the CAPI screen you can begin observing the lesson by answering the questions about teacher talk, teacher language, teacher action, and pupil activity for each three-minute interval.

The teacher selected to be observed may teach a lesson while being assisted by a co-teacher(s), who may perform a number of functions, such as classroom management, distributing notebooks / handouts / teaching aids etc. We are only interested in the teacher talk, teacher language, and teacher action of the selected teacher who is being observed.

In the CAPI software, the screen on the tablet for each three-minute interval will look as shown below. For each three-minute interval you will select the observed teacher talk, teacher language, teacher activity, and pupil activity. You have to select one option under each heading that best describes what is happening.

------TEACHER TALK------		What is the teacher saying?
Instructs / presents dictates	☐	
Chants	☐	
Explains	☐	
Closed question / response	☐	
Open question / response	☐	
Assists /group discussion	☐	
None of the above	☐	

[8] In addition, qualitative lesson observations were conducted in selected treatment schools, where observers watched a lesson and made notes on the ways in which teachers and pupils interacted, and later interviewed the teacher about her or his methods of teaching (see Chapter 6 in Volume I).

------TEACHER LANGUAGE-----		What language is the teacher speaking?
Hausa	☐	
English	☐	
Other	☐	
Teacher was silent / teacher was not present in class	☐	
------TEACHER ACTION-----		What is the teacher doing?
Writes on / reads from blackboard	☐	
Demonstrates / displays work	☐	
Moves around among pupils	☐	
Uses textbook	☐	
Uses materials (printed / improvised)	☐	
None of the above	☐	
-----PUPIL TALK AND ACTIVITY-----		What are the pupils doing and saying?
Group discussion / presentation	☐	
Group or pair work	☐	
Respond to open question	☐	
Respond to closed question	☐	
Individual work	☐	
Listen to teacher	☐	
Chant	☐	
Wait or copy while teacher writes on board	☐	
None of the above	☐	

The types of observed behaviours and their classification draws on the 2010 ESSPIN Teaching and Learning Baseline Survey (Davison, 2010), and were adapted based on a review by classroom observations specialists and TDP staff.

The descriptors for the different teacher talk and action and pupil activity recorded during the classroom observations can be found in Table 7.1 .

A small number of new categories was added at endline to reduce the large sizes of the 'none of the above' categories for teacher language and pupil talk and activity at baseline (see categories in blue font in Box 2). For teacher language, one new category, 'Teacher was silent / teacher was not present in class', was added, while for pupil talk and activity, three new categories were added 'Listen to the teacher', 'Chant', and 'Wait or copy while the teacher writes on the board'. These new categories are not included for the analysis that compares teachers' positive interaction with pupils in the classroom over time.

Table 7.1: Classroom observation: teacher talk and action and pupil activity descriptors

Code	Talk / action / activity	Practice descriptor
Teacher talk		
a	Instructs / presents / dictates to the whole class	Teacher talks to the whole class but does not question or give feedback. He/she might be giving instructions, 'presenting' some text directly from the textbook or the blackboard, or dictating a text for pupils to write. Examples: ➢ **Instructs**: Teacher saying 'today we will learn about shapes. Open your notebooks and copy these drawings of shapes from the blackboard'. **Note that classroom management instructions**, such as 'sit down', 'stand up and raise your hands', etc. **do not count as instructions** here. ➢ **Presents**: Teacher reading directly from a textbook or the blackboard without any additional own explanation. This could include reading out a story, poem, or a passage. ➢ **Dictates**: Teacher says 'open your notebooks and start writing as I say, "There are seven, s-e-v-e-n, days in a week""; or "Today I will give you a spelling test. Write in your notebooks. First word is umbrella…"um-bre-lla". Second word is machine…"ma-sheen"'.
b	Leads whole class chants	The teacher asks pupils to repeat what he/she has said, leading a whole class chant. This could include pupils repeating what the teacher has said; a poem; chorus song; or the whole-class reading out aloud from a text. Chants are usually preceded by the teacher saying 'Say/repeat after me'. Example: ➢ Teacher says, 'Repeat after me…Today is Monday…', and the pupils repeat together, 'Today is Monday'.
c	Asks a closed question or gives a closed response to the whole class	The teacher asks a closed question, which has only one right answer or only a short answer, usually for the pupils to remember facts; or answers pupils' questions in a way that closes the conversation, even if the pupils' question was an open one. Examples: ➢ Teacher asks, 'Children, what is the capital of Nigeria?'; Pupils: 'Abuja' or 'the capital of Nigeria is Abuja'. Here there is only one right answer. ➢ Teacher asks, 'Children, tell me…do you enjoy coming to school?'; Pupils: 'Yes teacher, we/I do' or 'No teacher, I/we don't'. Though both answers are correct, they are brief and it closes the conversation. ➢ Pupil asks, 'Teacher, how can I grow up to become a doctor?' Teacher: 'Study hard'.
d	Asks an open question or gives an open response to the whole class	The teacher asks a question that has many possible answers so that pupils imagine or analyse; or answers pupils' questions in a way that invites further discussion or thought, even if the pupil's question was a closed one. Examples: ➢ Teacher asks, 'Children, why do you like coming to school?'; Pupil 1: 'Because I like meeting my friends'; Pupil 2, '…because I like to read books'; Pupil 3, '…because I want to study and be a doctor when I grow up.' ➢ Pupil asks, 'Teacher, how many states does Nigeria have?'. Teacher: 'Ok, that is a good question…let us try to answer it together. Each of you will name a state and I will write it on the blackboard and we will then count. Aminu, tell me the name of a state'. Aminu: 'Kaduna'. Teacher: 'Good. Aisha, tell me the name of another state…'

e	Assists individuals or groups / joins group discussion	The teacher helps groups of pupils or individual pupils, or joins pupils' discussions. This may typically involve the teacher moving around individual pupils or groups, stopping to check on them and assist them.
f	Explains how something works / how to do a task[9]	Teacher explains how something works or how to do a certain task, often using a teaching aid but not necessarily so. This is different from instructions in that it does not involve telling pupils what to do but rather how to do it and typically involves breaking down single activities or concepts into smaller, easier sub-activities. Examples: ➢ The teacher may say, 'Do this addition sum: 3+2=?'. This is an instruction, while an explanation would be the teacher saying 'Here are three apples and here are two more apples. When we put them all together, they add up to 1-2-3-4-5, five apples. So 3+2 is equal to 5'. ➢ Teaching drawing or explaining scientific processes (with or without the use of models or equipment) will usually be included here. The teacher may say, 'Today you will all draw a duck. Look at my drawing on the board and try it yourself. First draw the beak, then the head. Draw an eye and then draw the neck…' Explanation is different from presenting as defined in Behaviour 1 above. Presenting involves reading directly from the textbook or blackboard, but explanation would mean the teacher adds content to the text from her/his own head, to make the text/concept simpler for the pupils. Teachers sometimes ask short questions, such as 'Okay? Understood?' to ensure the children are following while they explain things. If the teacher is doing this at the three-minute interval, select Explains rather than Closed question / response.
Teacher action		
a	Writes on / reads from blackboard	The teacher writes on or reads out aloud from the blackboard. Examples: ➢ The teacher writes mathematical exercises from the textbook or homework assignments on the blackboard. ➢ The teacher reads out aloud what has been written on the blackboard. If the teacher is using the blackboard to demonstrate a concept and not just writing out a phrase or copying from the textbook, then select Demonstrates / displays work.
b	Demonstrates / displays work using the blackboard	Teacher uses the blackboard to explain a concept or problem, shows how to solve a mathematics problem, illustrates a grammar or spelling point, or holds up a pupil's exercise book to explain something. If the teacher uses a textbook or materials to demonstrate something, select Uses textbook or Uses materials (below).
c	Moves around among pupils	Teacher moves away from the front of the class and may look at pupils' work or join group discussions. Generally, this is not accompanied by any of the other behaviours. If the teacher demonstrates something or displays work while moving around, select Demonstrates / displays work. If the teacher uses the textbook while moving around, select Uses textbook. If the teacher uses materials while moving around, select Uses materials.
d	Uses materials (printed/impro	Teacher uses printed materials (other than textbooks) or improvised materials, or observes as pupils using these materials under her/his guidance. Improvised materials includes things that the teacher or pupils have made.

[9] This teacher action was added after the TDP impact evaluation classroom observation design note was written.

	vised, that teacher has made)	
e	Uses textbook to explain something / reads from it	Teacher explains something from the textbook; explains a task in the textbook; or reads from the textbook.
Pupil talk and activity		
a	Group or pair discussion / presentation	Pupils are organised in groups or pairs and discuss a topic, or are reporting back on the results of a group discussion or group work.
b	Group or pair work to complete a task	Pupils are organised into groups or pairs to complete some task. They have either started the task, or are organised in groups and are waiting to be told how to start the task or continue.
c	Respond to open question	One or several pupils respond to a question that has many possible answers and that invites discussion.
d	Respond to closed question	One or several pupils respond to a question that only has one right answer or that can be answered with a short response.
e	Individual work	Pupils work on their own tasks individually, using independent thought in the process: for example, pupils completing exercises set by the teacher.
New pupil talk and activity categories at endline. These are not included in the trend analysis of teachers' positive interaction with pupils in the classroom.		
n.a.	Listen to teacher	Pupils listen to the teacher while he/she speaks.
n.a.	Chant	Pupils repeat what the teacher has said in a whole class chant. This could include pupils repeating what the teacher has said; a poem; chorus song; or the whole-class reading out aloud from a text. Chants are usually preceded by the teacher saying 'Say/repeat after me'. Example: ➢ Teacher says, 'Repeat after me…Today is Monday…', and the pupils repeat together, 'Today is Monday'.
n.a.	Wait or copy while the teacher writes on the board	Pupils wait while the teacher writes on the blackboard without talking, or they copy what the teacher has written on the board. Example: ➢ The teacher writes mathematical exercises from the textbook or homework assignments on the blackboard. ➢ The teacher has written an English passage on the board and the pupils copy it. If the pupils are individually working on a problem or question that the teacher has written on the board, then select Individual work. If the teacher is talking while writing on the board and the pupils are listening, select Listen to teacher.

Each observed teacher and pupil practice is classified as 'neutral', 'effective', or 'very effective', and is assigned a corresponding score of 0, 0.5, or 1 (Table 7.2). Among the 16 types of classroom practices identified by the baseline quantitative survey, five were classified as effective and five as very effective. These classifications are not immutable but include practices generally considered part of effective classroom practice (Siraj *et al*., 2014; Westbrook, 2013).

Table 7.2: Classroom observation: teacher talk and action, and pupil activity scoring scheme

Practice	Classification	Score	Practice	Classification	Score	Practice	Classification	Score
Teacher talk			**Teacher action**			**Pupil activity**		
Instructs / presents / dictates to the whole class	Neutral	0	Writes on / reads from blackboard	Neutral	0	Group or pair discussion / presentation	Effective	0.5
Leads whole class chants	Neutral	0	Demonstrates / displays work using the blackboard	Very effective	1	Group or pair work to complete a task	Effective	0.5
Asks a closed question or gives a closed response to the whole class	Neutral	0	Moves around among pupils	Effective	0.5	Respond to open question	Very effective	1
Asks an open question or gives an open response to the whole class	Very effective	1	Uses materials (printed or improvised, that teacher has made)	Effective	0.5	Respond to closed question	Neutral	0
Assists individuals or groups / joins group discussion	Very effective	1	Uses textbook to explain something / reads from it	Effective	0.5	Individual work	Neutral	0
Explains how something works / how to do a task	Very effective	1						
None of the above	Neutral	0	None of the above	Neutral	0	None of the above	Neutral	0

Note: (1) The practices in each of the three categories teacher talk, teacher action, and pupil activity are mututally exclusive and exhaustive. (2) The minimum score for each three-minute interval is 0 and the maximum score is 3. For the analysis, the overall score for each teacher is rescaled to obtain a total score between 0 and 1 for each three-minute interval.

8 The teacher motivation scale

For the TDP evaluation, teacher motivation was defined as the propensity of teachers to start and maintain behaviours that are directed towards fulfilling their professional goals, and in particular towards achieving better learning outcomes for the school's pupils (Cameron, 2015). The TDP theory of change argues that TDP materials, training and support can improve teacher motivation by raising teacher self-esteem (see McCormick, 2013; EDOREN, 2014). The evaluation framework elaborates on this mechanism, suggesting that teachers' motivation is increased as teachers feel more effective and see their pupils' learning outcomes improve.

In an attempt to measure teacher motivation systematically, the literature was reviewed on teacher motivation in developing countries and existing attempts to measure teacher motivation in developed and developing countries (Cameron, 2015). A simple theory was developed based on a distinction between two main aspects of motivation. The first aspect was referred to as 'can-do' – the extent to which teachers see themselves as able to do their jobs and do not see themselves as being overly constrained by factors beyond their control. This is similar to the concept of self-efficacy (Bandura, 1977, cited in Fernet et al., 2008) which is central to many motivation studies. The second aspect was referred to as 'will-do' – the extent to which teachers express the will, enthusiasm, and commitment, to do their job well.

A set of items was devised, taking influence from a range of existing scales, in order to measure these aspects of motivation and related concepts (Table 8.1). The scale consists of a list of items which are read to teachers by the data collector. The participant responds by saying that they strongly agree, agree, disagree, or strongly disagree. A simple graphic flashcard was used to remind participants of the four options and as an attempt to reduce the cognitive load of answering the questions on teachers whose reading skills may be limited. As well as motivation items, the scale includes several items related to the quality of interaction between teachers. Interaction between teachers is taken to be closely related to motivation, but not part of it *per se*.

This framework and the associated scale was applied in the TDP baseline, and has since been used and improved in other evaluations and research (Cameron et al., 2016; Pellens and Binci, 2018).

Table 8.1: Items in the teacher motivation scale

No.	Item
1	I enjoy teaching very much
2	As a teacher, I perform an important role in society
3	There is no point trying to teach pupils whose parents cannot read or write (R)
4	It is difficult to teach in this school because the building is in poor condition (R)
5	It is difficult to manage pupils in my classrooms (R)
6	If I could choose another job today, I would still choose teaching
7	In the past two years, my job has become more satisfying
8	I teach too many classes (R)
9	Teachers at my school have the knowledge and skills to do our jobs well
10	Teachers at this school are highly committed to their job
11	I have teachers that I consider my friends at my school
12	Teachers at my school work well together
13	Teaching my class today/yesterday was boring (R)
14	It is worth working harder to make sure the pupils do well
15	Most of the pupils in this school are not intelligent enough to do well (R)

16	Teaching is very tiring (R)
17	Teaching well is important to me
18	Teachers at this school trust each other
19	Pupils work hard in this school and want to succeed
20	I try my best to teach my pupils but their parents do not help (R)
21	Teachers at this school feel responsible to help each other out
22	There are too many pupils in my classroom (R)
23	I don't always have the materials I need to do my job (R)
24	I have all the support I need to teach my pupils well
25	There is no point in spending a lot of time preparing for a class (R)

(R) indicates that the question is reverse-coded (stronger agreement is associated with lower motivation or worse interaction).

Exploratory factor analysis was used to examine the extent to which teachers' patterns of responses in the endline survey fit the theory (Beavis, 2018). The sub-scales, particularly those relating to the 'can-do' factor, had poor reliability, and could not be improved by dropping individual items. It may be that items such as 'There are too many pupils in my classroom' (item 22) are picking up changes in external circumstances rather than in teachers' perceptions of their ability to teach well despite challenging circumstances – although similar items worked relatively well in the TDP baseline and in other evaluations.

Therefore a different approach was taken. The analysis sought to identify what structures were identified in the data and the extent to which such structures were interpretable in terms of teacher motivation. Two factors emerged which had both good enough reliability and interpretability. The first, labelled commitment, consisted of items 2, 7, 9, 10, 14, 17, and 19. It had good reliability (alpha = 0.75) relates to teachers' commitment to their job, their perception of their importance, but also their sense of satisfaction and their perception that 'pupils work hard and want to succeed'. The second factor, labelled enthusiasm, consisted of items 5, 13, 15, 16, and 20. This factor had marginally acceptable reliability (alpha = 0.57). The items in this factor are all reverse-coded and seem to relate to a sense of the job being tiring, difficult or boring. (The reverse-coding means that teachers with a high score in this factor are those with *less* sense of the job being tiring, difficult or boring.) These factors were calculated in both baseline and endline data so that comparisons could be made across time.

Thus, the motivation scale provides reliable and intuitively meaningful measures of two aspects of motivation which can be used in the TDP evaluation, but these do not have a strong connection to the theory that originally drove the development of the scale. In particular, it is not able to measure teachers' perceptions of their ability to do their job. This would have been useful because it is a potentially important causal factor, relating to the idea of self-esteem in the programme's theory of change: teachers who are trained should become more confident in their ability to do their job, particularly if there are visible improvements in their pupils' learning. Teachers who do have this confidence have more reason to expend effort than those who perceive it not to be possible to make any difference to their pupils' learning.

The scale has been improved in more recent work (Cameron et al., 2016), but the original baseline scale was used for the TDP endline in order to ensure comparability between baseline and endline. Future research on this topic will need to explore why the scale may not always work well in measuring 'can-do', whether items can be adjusted or added to improve reliability of this sub-scale, or whether the underlying theory needs to be adjusted.

9 Permits, consent, confidentiality, and datasets

Conducting fieldwork requires high ethical standards, to ensure that expectations are not raised, confidentiality is maintained, respondents are never forced to participate, and respondents (particularly children) are not encouraged to speak about subjects that may be traumatising. Both quantitative and qualitative data collection research proposals for this impact evaluation were passed through OPM's ethical review board. The application for ethical approval of the quantitative and qualitative research of the endline TDP evaluation was then submitted, ahead of the fieldwork, to the following authorities in Nigeria. Ethical approval was received from all four authorities before fieldwork commenced.

1. National Health Research Ethics Committee of Nigeria (protocol number: NHREC/01/01/2007-10/08/2017);

2. Ministry of Education, Science and Technology, Jigawa State (reference number: MOEST/ADM/263 VOL.I);

3. Katsina State Teacher's Service Board, Katsina State (reference number: KTS/TSB/GEN/VOL.I/1); and

4. Ministry of Education, Zamfara State (reference number: MOE/PLAN/GEN/350/VOL.I).

The key areas for ethical consideration for research involving human subjects are: (i) informed consent; (ii) harms and benefits; (iii) payment and compensation; and (iv) privacy and confidentiality. Adherence of the endline research to the ethical standards in each of these areas is outlined below.

Ahead of the visits to schools, the state coordinators visited the various LGA education board secretaries and TDP state focal persons for introductory purposes and to inform them of the presence of the survey team in their LGAs. Upon arrival at the school, the team supervisors introduced themselves and their teams to the head teacher, explaining the purpose of the visit and the time that would be required to complete the survey. Verbal consent to carry out the interview/test/observation was obtained from all respondents. Respondents were read a statement that informed them of the nature of the study and what would be required of them as study participants; given the option to refuse to participate in the study or to withdraw consent at any point during the interviews; and assured of the confidentiality of their responses (and pupils and teachers were particularly assured that their responses would not affect their grades or jobs). Consent was obtained from the head teachers (in their role as the responsible persons, in place of the pupils' parents) to test the pupils, and assent to participate was also obtained from the pupils themselves.

The research had no risk of physical harm to any of the respondents. The research ensured that other forms of harm were minimised by safeguarding the privacy and confidentiality of the respondents who participated. Participants were interviewed in an environment in which they were comfortable, and which secured their privacy. Particular care was given to the pupils, who were generally between the ages of 10 and 12. The data collectors were trained on the ethics of working with children – ensuring a safe and private space for their participation, letting them ask questions, making it clear it was fine for them to leave a question or leave the interview entirely, keeping responses confidential and anonymous. All personal data collected as part of this survey are only available to authorised individuals for analytical purposes and are handled in accordance with data protection best practices. Each respondent/unit of analysis (school, head teacher, teacher, and pupil) was assigned a unique identifier that was used to analyse the data. All data related to this survey will be anonymised before they are made publicly available (subject to DFID approval). This means that no responses are attributable or can be traced to any individual or school.

The quantitative fieldwork was carried out by field teams made up of national enumerators and field supervisors, supported by staff from the OPM Nigeria office. The interviews with head teachers and teachers, and pupil testing, were conducted in Hausa. After discussions with TDP state staff the OPM

Nigeria office team arranged for the delivery of letters of permission to visit schools to the SUBEB officials and Education Secretaries concerned in sampled LGAs. Sending the permit letters was not considered sufficient to ensure the Education Secretaries had read and agreed to the school visits. Therefore, follow-up phone calls were carried out to confirm that they had received the letters seeking permission to visit schools from the SUBEBs, and that they understood the purpose of the research and allowed the field teams to visit schools in their LGA.

Informed written consent was sought from all participants for the quantitative research. Given that the baseline surveys were school-based (and not home-based) it was not possible to seek consent from pupils' parents, and hence consent from the head teachers (as the 'guardians' of the pupils while they are in school) and from the pupils themselves was sought. Verbal assent was sought from children, and the head teacher signed a written consent form on their behalf. When they arrived at schools, the team supervisors started by introducing themselves and their teams to the head teacher, explaining the purpose of the visit and the time that would be required to complete the survey. The enumerators introduced the study and interviews/texts to the head teacher and to all the respondents (pupils and teachers), and were given the option to refuse to participate in the study. If a respondent was reluctant and/or further explanation was requested, the enumerators were trained to be as exhaustive as possible in explaining the study and its purpose. No head teacher or pupil declined to participate in the survey.

The qualitative fieldwork was carried out by a team of national researchers. Interviews were conducted in Hausa. The field teams undertook all possible measures to keep disruptions of the school day to a minimum by ensuring that head teachers were informed in advance of the dates of the school visits and regarding which types of research activity would take place. The interviews were recorded after the informed written consent of participants was granted. The sequencing of interviews and other qualitative research activities was also – as far as possible – organised in cooperation with school members, in order to minimise disruption to school life and to ensure smooth running of the research. Interviews frequently took place outside the school building in order to minimise disruption within teaching spaces.

Informed written consent was sought from all participants at the state, LGA, and school levels for the qualitative research. The aims of the research and their ability to withdraw consent at any point during the interviews or discussions were explained to participants. In order to ensure that the participants were comfortable with the procedure the researchers would read out the explanation and ask the participants whether the information provided was clear. Participants were invited to either end or temporally interrupt the interview or discussion if additional questions or concerns arose.

The fieldwork included discussions with children. The children participating in the research were boys and girls from Primary 6. There is some debate in the development community regarding who is in a position to provide consent for research conducted with participants who are young children. For ethical reasons, the team decided to gain both the consent of head teachers and children. As a first step, head teachers were asked to provide written consent that they were willing to allow the children to participate in the research. If permission was granted, the children were asked to also provide their assent to participating in the research. At both stages, the nature of the research was explained and it was made clear that children were under no obligation to participate.

The evaluation upholds several aspects of DFID's human rights approach (especially participation and inclusion) (Piron and Watkins, 2004) through rigorous training on, and practice of, ethical standards during data collection. This includes seeking consent from respondents, facilitating participation of respondents irrespective of disability status, and training gender-balanced data collection teams, among other considerations.

Though not totally avoidable, the interviews were scheduled by data collectors in order to minimise any interruption to the normal flow of activities in the school and the need for teachers and pupils to stay beyond school hours.

No monetary incentives were given to respondents for participation in the study. Each school that was part of the qualitative study received a gift as a token of thanks for their time and participation. The total value of these gifts was well under £3. Participants in the qualitative fieldwork received refreshments and, additionally, children received a pencil, two to three crayons, and an eraser. Children who participated in the qualitative survey also received similar items and refreshments. The possibility of adverse effects of these gifts on respondents is considered to be minimal.

This independent impact evaluation is being carried out by EDOREN and is intended for primary consumption by TDP and DFID Nigeria. As such, the final ownership and copyright of the data, analysis, and reports rests with EDOREN, which is managed by OPM. However, all outputs (especially reports) produced under this evaluation – by joint agreement – are being co-branded to bear EDOREN, UK Aid, and TDP logos.

Data ownership is defined by DFID's contracts with OPM for EDOREN, and with Mott McDonald for TDP. It is EDOREN's understanding that the data collected are co-owned by OPM and DFID. As stated in the TDP evaluation framework, the clean, anonymised evaluation datasets and metadata will be made publicly available, probably on the EDOREN website and in the World Bank micro-databank (subject to DFID approval), so that researchers can replicate and extend the evaluation analysis, in line with DFID's Open Access policy.

Intellectual property rights in respect of any materials produced by EDOREN (such as evaluation reports, policy briefs etc.) are the property of OPM. However, OPM has granted DFID a worldwide, non-exclusive, irrevocable, royalty-free licence to use all of these data and materials.

All personal data collected as part of this survey are available only to authorised individuals for analytical purposes and are handled using data protection best practices. Each respondent has been assigned a unique identifier that is used to analyse the data. All cleaned and documented datasets, anonymised by removing personal information that could be used to identify respondents, related to the study will be made public through the EDOREN website and World Bank micro-databank (subject to DFID approval) to enable national researchers, research students, and other education stakeholders to access and use the impact evaluation data to conduct additional analysis and research. Baseline data are already available on the World Bank micro-databank (http://microdata.worldbank.org/index.php/catalog/2672). All data have been backed up and are stored in an 'OPM Stats archive'. OPM will store all original data and transcripts for three years, after which time they will be destroyed. Qualitative data will not be released outside of the research team, as it is difficult to assure the confidentiality of participants given the small sample of the qualitative study.

10 Stakeholder engagement and impact evaluation governance

The EDOREN evaluation team created an endline plan (Cameron *et al*., 2017), in consultation with staff of DFID and TDP. The endline plan also drew on the results of initial research, including an implementation review (Durrani *et al.*, 2018) and a validation telephone survey (Cameron and Pettersson, 2017) in which head teachers were asked about the training they had received in their schools, and the extent of teacher and pupil attrition. The plan was reviewed by EQUALS in November 2017, and was revised in response to the EQUALS reviewer's comments in January 2018. The endline plan effectively constitutes the agreed terms of reference for the endline evaluation.

The implementation review (Durrani *et al*., 2018) was important in engaging TDP staff in Abuja, TDP state teams, and SUBEBs. Key informants were interviewed, and a meeting was held in November 2017, to verify the results of the implementation review. Further comments were then provided by TDP in order to finalise and agree the review report. This report has acted as a detailed guide to the TDP intervention, its TOC, and its evolution over time, for the EDOREN evaluation team.

A steering committee for the evaluation was set up in November 2017 and met in January 2018. It included representatives from TDP, DFID, the Federal Ministry of Education (FMOE), the Universal Basic Education Council (UBEC), and EDOREN. The steering committee was initiated too late to advise on the design of the evaluation. Its purpose, instead, is to advise on the lessons and recommendations emerging from the evaluation, and on ways of maximising the impact of the evaluation through the dissemination of the results.

As well as TDP and the head teachers and teachers in the schools, other implementers, such as Jolly Phonics and RANA, were consulted during the evaluation. Evaluators of other programmes, including GEP3/RANA and ESSPIN, were also consulted during the evaluation design. The evaluation fieldwork was conducted through EDOREN and OPM Nigeria, helping to build evaluation capacity in Nigeria. State and local government officers were employed as data collectors, building their capacity for measuring teaching and learning outcomes in schools, and helping to ensure buy-in at the local level for the evaluation's findings.

The initial plan for dissemination of the report's findings, and finalisation of the report, is shown in Table 10.1.

Table 10.1: Plan for report dissemination

Date	Activity
End of March 2018	Complete draft of the endline impact evaluation report and submit to DFID, EQUALS, TDP, and the steering committee for comments
Early April 2018	Present the report to the steering committee
Mid-April 2018	Present the report to DFID, TDP, and other DFID-funded programmes in a learning day
Late April 2018	Present the report in further meetings at the state level
End of April 2018	Finalise the report taking into account comments from all stakeholders, particularly on lessons and policy recommendations

May 2018	Develop three policy briefs on key issues raised by the evaluation report. Share policy briefs and presentation slides online, and among key stakeholders in Nigeria, for wider dissemination of results and recommendations.

References

Please see Volume I of this report, which includes references for both volumes.

Annex A Impact evaluation matrix with quantitative and qualitative indicator definitions

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
TOC LEVEL: FINAL IMPACT									
Re-1	Relevance	Final impact	• Improved learning levels in English, maths, and science for pupils in TDP schools 2014–2017	• Does this objective address the needs, priorities, and constraints of pupils in northern Nigeria? • Has TDP improved learning outcomes for pupils in TDP schools? • Are TDP impacts on pupil learning heterogeneous for teachers with different qualifications and years of experience?	• Pupil learning levels in English, maths, and science • Change between P3 (2014) and P6 (2017) in proportions of pupils in different performance bands in English, maths, and science, in TDP and control schools	• IRT-based scale scores in English, maths, and science • Proportion of pupils in the bottom and top performance bands in each of English, maths, and science and technology, respectively (pupils tested in P3 at baseline and in P6 at endline)	Tabulation from survey data; Impact estimation	TDP evaluation quantitative surveys – pupil learning assessment	State, pupil gender, teacher qualifications and years of experience
Im-1	Impact								
Im-2									
Im-3	Impact	Final impact		• Are there any other positive or negative unanticipated TDP impacts?			Inference from qualitative data	TDP evaluation qualitative research	
TOC LEVEL: INTERMEDIATE IMPACT									
Re-2	Relevance	Intermediate impact	• Improved teacher effectiveness in classroom	• Does this objective address the needs, priorities, and constraints of primary teachers in northern Nigeria? • Has TDP improved teacher effectiveness in the classroom?	• Time teachers involve pupils in positive interaction during lesson (% of total lesson time)	• (Numerator/denominator)*100 *Numerator*: For each category of teacher talk and action and pupil activity in a three-minute lesson observation interval, teachers/pupils are assigned a score of 0, 0.5 or 1 depending on whether they demonstrated a 'neutral', 'effective', or 'very effective' practice. For each three-minute interval, the scores for the three categories are summed. This gives each teacher a score of from 0 to 3 for each observed interval. For each teacher, the score for each observed interval is rescaled by dividing by 3 and the scores for all observed intervals are summed to obtain a total	Tabulation from survey data; Impact estimation	TDP evaluation quantitative surveys – classroom observations	
Effe-1	Effectiveness				• Percentage change in time teachers involve pupils in positive interaction during lesson (% of total lesson time)				

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
						score. *Denominator*: The total number of three-minute intervals observed for each teacher. *Note*: To construct the pedagogy indicator the actual lesson time for each teacher is used as lesson length varies. Lessons of nine minutes or shorter are excluded from the analysis			
Effe-2	Effectiveness	Intermediate impact	• Improved teacher effectiveness in the classroom	• Has TDP improved teacher effectiveness in the classroom?	• Percentage change in average daily teacher absence from school (% of teachers) • Reasons for teacher absenteeism	• (Numerator/denominator)*100 *Numerator*: Total number of teachers absent over the previous five school days *Denominator*: Total number of teachers employed multiplied by 5 (from school record checks) • Number of interviewed teachers who reported being absent from school in the last five school days and reported reason X divided by the number of interviewed teachers who reported being absent from school in the last five school days multiplied by 100	Tabulation from survey data; Impact estimation	TDP evaluation quantitative surveys – school record checks and teacher interviews	
Effe-3	Effectiveness	Intermediate impact	• Improved teacher effectiveness in outside classroom support	• Has TDP improved teacher effectiveness in outside classroom support?	• Percentage change in teacher scores on ability to assess and monitor pupil academic progress (*TBC*)	• Number of correct answers as a proportion of the maximum score on the TDNA assessing and monitoring pupil academic progress component multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – TDNA	
TOC LEVEL: INTERMEDIATE IMPACT TO FINAL IMPACT ASSUMPTIONS									
Re-13	Relevance	Intermediate impact to final impact assumption	• Children attending school regularly	• Is this assumption correct in the Nigerian context?	• Proportion of pupils missing more than one day of school in past four weeks • Average number of school days missed in past four weeks • Proportion of pupils spending five or more hours per day in school	• Of children in Jigawa, Katsina, and Zamfara and who attended primary school at all in the current school year, the proportion whose parents said they missed more than one day of school in the past four weeks • For children in Jigawa, Katsina, and Zamfara who attended primary school at all in the current school year, the average number of days missed in the past four weeks	Review of survey estimates	NEDS 2015	State; pupil gender

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
						• Of children in Jigawa, Katsina, and Zamfara who attended primary school at all in the current school year, the proportion whose parents said they left the school five or more hours after arriving on the last day that they went to school			
Im-8	Impact			• What factors facilitated or inhibited TDP's achievement of its impacts?	• Perceptions of regular attendance by pupils and teachers		Inference from qualitative data	TDP evaluation qualitative research – teachers and pupils	
Re-15	Relevance	Intermediate impact to final impact assumption	• Children receiving adequate support for learning at home	• Is this assumption correct in the Nigerian context?	• Proportion of children doing homework • Proportion of children receiving assistance with homework	• Of children in Jigawa, Katsina, and Zamfara and who attended primary school at all in the past school year, the proportion reported by their parents ever to do homework • Of the same group, the proportion of parents who reported that the child ever had help at home with homework	Review of survey estimates	NEDS 2015	State; pupil gender
Re-16	Relevance	Intermediate impact to final impact assumption	• A class size small enough to allow improved teacher effectiveness to have an impact	• Is this assumption correct in the Nigerian context? • What factors facilitated or inhibited TDP's achievement of its impacts?	• Average class size in observed lessons	• Number of pupils present during all classroom observations divided by the total number of classroom observations	Tabulation from survey data; Inference from qualitative data	TDP evaluation quantitative surveys – classroom observations and school record checks; TDP qualitative research – teachers and pupils	State
					• Average pupil–teacher ratio	• Total number of P1–P6 pupils registered at the school divided by the total number of P1–P6 teachers employed at the school			
Im-12	Impact				• Perceptions of appropriateness of class size				
Re-17	Relevance	Intermediate impact to final impact assumption	• Adequate classroom materials (blackboards, books, desks, etc.) being available	• Is this assumption correct in the Nigerian context? • What factors facilitated or inhibited TDP's achievement of its impacts?	• Perceptions of availability of classroom materials		Inference from qualitative data	TDP evaluation qualitative research – teachers, classroom observations	
Im-13	Impact								

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
Re-18	Relevance	Intermediate impact to final impact assumption	• Curriculum and materials that are appropriate for the language and ability of pupils	• Is this assumption correct in the Nigerian context? • What factors facilitated or inhibited TDP's achievement of its impacts?	• Perceptions of pupils' understanding of materials and curriculum		Inference from qualitative data	TDP evaluation qualitative research – teachers and pupils, classroom observations	
Im-14	Impact								
Im-9	Impact	Intermediate impact to final impact assumption	• Children having the capacity to learn from improved teaching in the language of instruction (they are school-ready)	• What factors facilitated or inhibited TDP's achievement of its impacts?	• Average pupil competencies in English	• IRT-based score in English tests • Proportion of pupils in the bottom and top performance bands in English	Tabulation from survey data	TDP evaluation quantitative surveys – pupil learning assessment	State; gender of pupils
Im-11	Impact	Intermediate impact to final impact assumption	• Children supporting their peers to learn	• What factors facilitated or inhibited TDP's achievement of its impacts?	• Perceptions of peer support		Inference from qualitative data	TDP evaluation qualitative research – teacher and pupils, classroom observations	
TOC LEVEL: OUTCOME									
Re-3	Relevance	Outcome	• Improved teacher subject knowledge	• Does this objective address the needs, priorities, and constraints of primary teachers in northern Nigeria? • Has TDP improved teacher subject knowledge?	• Average teacher raw scores on TDNA for English, maths, and science and technology • Percentage change in average teacher raw scores on TDNA for English, maths, and science and technology	• Number of correct answers as a proportion of the maximum score on each of the TDNA English, maths, and science and technology components multiplied by 100	Tabulation from survey data; Impact estimation	TDP evaluation quantitative surveys - TDNA	State; gender of teachers (if adequate number of female teachers in the sample)
Effe-4	Effectiveness								
Effe-5	Effectiveness	Outcome	• Improved head teacher leadership and management	• Has TDP improved head teacher leadership and management?	• Percentage change in number of head teachers holding formal meetings with teachers at least once per month	• Number of head teachers who report holding formal meetings with all or a group of teachers at least once per month as a proportion of all head teachers multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – head teacher interview	

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
					• Percentage change in number of head teachers observing lessons	• Number of head teachers who reported carrying out lesson observations during the last 10 working days as a proportion of all head teachers multiplied by 100			
					• Percentage change in number of head teachers taking action to reduce pupil absenteeism	• Number of head teachers who reported taking action to reduce pupil absenteeism during the last school year as a proportion of all head teachers multiplied by 100			
					• Percentage change in number of head teachers taking action to reduce teacher absenteeism	• Number of head teachers who reported taking action to reduce teacher absenteeism during the last school year as a proportion of all head teachers multiplied by 100			
Effe-6	Effectiveness	Outcome	• Improved teacher pedagogical knowledge	• Has TDP improved teacher pedagogical knowledge?	• Percentage change in teacher scores on ability to assess and monitor pupil academic progress	• Number of correct answers as a proportion of the maximum score on the TDNA assessing and monitoring pupil academic progress component multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – TDNA	State; teacher gender
TOC LEVEL: OUTCOME TO INTERMEDIATE IMPACT ASSUMPTIONS									
Re-19	Relevance	Outcome to intermediate impact assumption	• TDP materials being appropriate and available	• Is this assumption correct in the Nigerian context? • What factors facilitated or inhibited TDP's achievement of its outcomes?	• Teachers and government officials reporting TDP materials appropriate and available		Inference from qualitative data; tabulation from survey data	TDP evaluation qualitative research – teachers, LGEA interviews, SUBEB interviews; TDP evaluation quantitative surveys – teacher interviews and classroom observations	State; teacher gender
Effe-12	Effectiveness				• Proportion of teachers with access to TDP materials	• The number of TDP-trained sample teachers reporting they have access to each TDP material (teacher's guide, lesson plans, flash cards, Reading Assessment Guide) as a proportion of the total number of TDP-trained sample teachers multiplied by 100 • The number of observed classrooms that have each type of TDP material (posters, teacher's guide, lesson plans, flash cards) as a proportion of the total number of observed classrooms multiplied by 100			

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
					• Proportion of teachers using TDP materials in the classroom	• The number of TDP-trained sample teachers reporting they use each TDP material in the classroom (teacher's guide, lesson plans, flash cards, Reading Assessment Guide) as a proportion of the total number of TDP-trained sample teachers who reported having access to the TDP material multiplied by 100 • The number of sample teachers observed in the classroom who used each type of TDP material (poster, teacher's guide, lesson plans, flash cards) as a proportion of the total number of observed classrooms where the TDP material was available multiplied by 100			
					• Proportion of teachers reporting that TDP lesson plans are appropriate given lesson length • Proportion of teachers reporting that TDP lesson plans are appropriate given curriculum	• The number of TDP-trained sample teachers reporting the length of TDP lesson plans is appropriate given the duration of the lessons they teach as a proportion of the total number of TDP-trained sample teachers who reported having access to the lesson plans multiplied by 100 • The number of TDP-trained sample teachers reporting the contents of TDP lesson plans are appropriate given the curriculum they teach as a proportion of the total number of TDP-trained sample teachers who reported having access to the lesson plans multiplied by 100			
Re-20	Relevance	Outcome to intermediate impact assumption	• Selected teachers being retained in schools where TDP is operating	• Is this assumption correct in the Nigerian context? • What factors facilitated or inhibited TDP's achievement of its outcomes?	• Proportion of teachers in TDP schools who left the school by endline • Proportion of teachers in TDP schools who transferred to another	• Number of teachers who were interviewed at baseline in the TDP schools and who were no longer employed at the school by endline divided by the total number of teachers who were interviewed at baseline in the TDP schools multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – teacher interviews, head teacher interviews, and school record checks	State
Effe-15	Effectiveness								

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
					school since the last school year	• Number of teachers who transferred to another school since the last school year as a proportion of the total number of teachers employed at the school in the last school year multiplied by 100			
Effe-13	Effectiveness	Outcome to intermediate impact assumption	• Selected teachers being sufficiently intrinsically motivated to turn improved knowledge into improved effectiveness	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Levels of (intrinsic and extrinsic) motivation	• Teachers' average levels of *commitment* and *enthusiasm*, as measured through factor analysis of a self-reported attitudinal survey instrument	Tabulation from survey data; Inference from qualitative data	TDP evaluation quantitative surveys – teacher interviews; TDP evaluation qualitative research – teacher interviews, classroom observations	
					• Teacher perceptions of own motivation				
Effe-14	Effectiveness	Outcome to intermediate impact assumption	• Selected teachers being sufficiently extrinsically motivated to apply their new knowledge	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Proportion of teachers reporting salaries paid on time and in full	• Number of sample teachers who report receiving their salary always on time, usually on time, usually delayed, always delayed, or who did not receive any salary in the last school year as a proportion of all sample teachers multiplied by 100 • Number of sample teachers who report receiving the correct salary amount for all of their last three payments, or some of the payments, or receiving no payments as a proportion of all sample teachers multiplied by 100	Tabulation from survey data; inference from qualitative data	TDP evaluation quantitative survey – teacher interviews; TDP evaluation qualitative research – teacher interviews	
					• Levels of (intrinsic and extrinsic) motivation	• Teachers' average levels of *commitment* and *enthusiasm*, as measured through factor analysis of a self-reported attitudinal survey instrument			
					• Teacher perceptions of own motivation				

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
Effe-16	Effectiveness	Outcome to intermediate impact assumption	• Selected teachers being class-ready: in other words, have the capacity to apply their new knowledge	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Proportion of teachers with appropriate qualifications	• The number of sample teachers who hold an NCE qualification as a proportion of all sample teachers multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – teacher interviews	State
					• Proportion of teachers receiving training in last three years	• The number of sample teachers who received INSET during the last three school years as a proportion of all sample teachers multiplied by 100			
Effe-17	Effectiveness	Outcome to intermediate impact assumption	• Selected teachers being supported to apply their new knowledge	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Proportion of head teachers holding formal meetings with teachers at least once per month	• Number of head teachers who report holding formal meetings with all or a group of teachers at least once per month as a proportion of all head teachers multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – head teacher interviews	
Effe-18					• Proportion of head teachers observing lessons	• Number of head teachers who reported carrying out lesson observations during the last 10 working days as a proportion of all head teachers multiplied by 100			
					• Proportion of head teachers taking action to reduce pupil/teacher absenteeism	• Number of head teachers who reported taking action to reduce pupil/teacher absenteeism during the last school year as a proportion of all head teachers multiplied by 100			
TOC LEVEL: OUTPUT									
Re-5	Relevance	Output	• Develop strategic partnerships with the FMOE, UBEC, and National Education Resource Development Council (NERDC) to ensure the national roll-out and long-term sustainability of the INSET model	• Is this approach coherent with the broader policy environment at the state and federal levels in Nigeria, and with DFID policy	• Are FMOE, UBEC, and NERDC appropriate custodians of in-service teacher training in Nigeria?		Inference from qualitative data	TDP evaluation qualitative research – FMOE, UBEC, NERDC interviews; policy documents	
Re-6	Relevance	Output	• Partner with SUBEBs that will become key strategic homes for TDP	• Is this approach coherent with the broader policy environment at the state and federal levels in	• Are SUBEBs appropriate institutional homes for INSET in each state?		Inference from qualitative data	TDP evaluation qualitative research – SUBEB, FMOE, TDP staff interviews;	

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
				Nigeria, and with DFID policy?				policy documents	
Re-7	Relevance	Output	• Provide continuous support to teachers for a prolonged period of time and embed this mechanism in both schools and the TDP states' teacher education systems	• Is this approach coherent with the broader policy environment at the state and federal levels in Nigeria, and with DFID policy?	• Is a programme of continuous support to teachers realistic and aligned with the policy context in Nigeria?		Inference from qualitative data	TDP evaluation qualitative research – FMOE, UBEC, NERDC interviews; policy documents	
Re-8 (*TBC*)	Relevance	Output	• Engage with NERDC at state level for teacher training material development and strengthen its capacity to become the custodian of all audio-visual resources within and beyond the project period	• Is this approach coherent with the broader policy environment at the state and federal levels in Nigeria, and with DFID policy?	• NERDC is the appropriate institution for material and technology		Inference from qualitative data	TDP evaluation qualitative research – NERDC, SUBEB interviews	
Re-9	Relevance	Output	• Produce Hausa-based instructional materials for teachers on maths P1–3	• Does this approach address the needs, priorities, and constraints of primary teachers in northern Nigeria?	• Proportion of observed P1–3 lessons where teacher was using Hausa only	• The number of observed P1–3 classrooms where teacher was using Hausa only as a proportion of the total number of observed P1–3 classrooms multiplied by 100	Tabulation from survey data; inference from qualitative data	TDP evaluation quantitative survey – classroom observations; TDP evaluation qualitative research – teacher interviews, classroom observations	
					• Teachers using Hausa to teach P1–3				
Re-10	Relevance	Output	• Use English as the language of instruction for all materials for P4–6	• Does this approach address the needs, priorities, and constraints of primary teachers and	• Proportion of observed P4–6 lessons where teacher was using English only	• The number of observed P4–6 classrooms where the teacher was using English only as a proportion of the total number of observed P4–6 classrooms multiplied by 100	Tabulation from survey data; inference	TDP evaluation quantitative survey – classroom observations;	

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
				children in northern Nigeria?	• Teachers using English to teach P4–6		from qualitative data	TDP evaluation quantitative surveys – teacher interviews and classroom observations	
Re-11	Relevance	Output	• Select appropriate technology based on an assessment against a range of pre-defined criteria	• Does this approach address the needs, priorities, and constraints of primary teachers in northern Nigeria? • How does TDP's organisational management and setup facilitate delivery?	• Teachers being able to use TDP technology • Efficiency of TDP's organisational management and setup		Inference from qualitative data	TDP evaluation qualitative research – teacher interviews, classroom observation; TDP staff interviews	
Effi-20	Efficiency								
Re-12 (*TBC*)	Relevance	Output	• Support states to expand teacher support to other schools in the state from 2015 onwards	• Is this approach coherent with the broader policy environment at the state and federal levels in Nigeria, and with DFID policy?	• Is the application of the TDP model across TDP states appropriate and consistent with the state policy context?		Inference from qualitative data	TDP evaluation qualitative research – SUBEB interviews; policy documents	
Effi-5	Efficiency	Output	• Implement a two-year pilot with 6,000 teachers from 1,500 schools across all LGEAs in three initial states before scaling up to six states • Recruit a pool of teacher educators as the TDT • Recruit TFs from the current school supervisor cadre • Design and implement separate professional development	• Were TDP results achieved on time and in full? • How does TDP's organisational management and setup facilitate delivery?	• Quantity undertaken against plan • Date activity undertaken against plan • Reason for deviation • Efficiency of TDP's organisational management and setup		Implementation review	TDP programme monitoring data; TDP evaluation qualitative research – TDP interviews, programme documents	
Effi-6									
Effi-7									
Effi-9									
Effi-10									
Effi-11									
Effi-12									
Effi-13									
Effi-14									

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
Effi-15 Effi-16 Effi-17 Effi-18 Effi-19 Effi-21 Effi-22			programmes and activities for teachers, head teachers, TFs, and TDTs • Develop new or adapt existing materials for teachers, TFs, and head teachers • Select appropriate technology based on an assessment against a range of pre-defined criteria • Produce audio-visual resources on SD cards • Engage head teachers actively in pedagogy and leadership and management training						
Effi-8	Efficiency	Output	• Provide continuous support to teachers for a prolonged period of time and embed this mechanism in both schools and the TDP states' teacher education systems	• Were the TDP results achieved on time and in full?	• Frequency of inspection visits to schools	• Number of head teachers reporting that a SUBEB or LGA supervisor visited the school during the last school year (more than three times, two to three times, or once a month or less) as a proportion of all head teachers multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys, head teacher interviews	State
Effe-7	Effectiveness	Output	• Implement a two-year pilot with 6,000 teachers from 1,500 schools across all LGEAs in three initial states before scaling up to six states	• Has TDP trained more teachers?	• Percentage change in number of teachers receiving TDP INSET • Training attendance	• The number of sample TDP teachers who received TDP INSET during the last three school years as a proportion of all sample TDP teachers multiplied by 100	Tabulation from survey data; review of programme implementation data	TDP evaluation quantitative surveys – teacher interviews; implementation review	State

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
							and documents		
Effe-8	Effectiveness	Output	• Provide continuous support to teachers for a prolonged period of time and embed this mechanism in both schools and the TDP states' teacher education systems	• Has TDP improved support to schools?	• Percentage change in the frequency of inspection visits to schools	• Number of head teachers reporting that a SUBEB or LGA supervisor visited the school during the last school year (more than three times, two to three times, or once a month or less) as a proportion of all head teachers multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – head teacher interviews	State
Effe-9	Effectiveness	Output	• Develop new or adapt existing materials for teachers, TFs, and head teachers	• Has TDP improved teacher materials?	• Perception of usefulness of materials		Inference from qualitative data	TDP evaluation qualitative research – pupil interviews, teacher interviews, classroom observations	State
Effe-10	Effectiveness	Output	• Develop new or adapt existing materials for teachers, TFs, and head teachers	• Has TDP improved teacher materials?	• Proportion of teachers using TDP materials	• The number of TDP-trained sample teachers reporting they use each TDP material in the classroom (teacher's guide, lesson plans, flash cards, Reading Assessment Guide) as a proportion of the total number of TDP-trained sample teachers who reported having access to the TDP material multiplied by 100 • The number of teachers observed in the classroom who used each type of TDP material (poster, teacher's guide, lesson plans, flash cards) as a proportion of the total number of observed classrooms where the TDP material was available multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – teacher interviews and classroom observations	State
Effe-11	Effectiveness	Output	• Engage head teachers actively in pedagogy and leadership and	• Has TDP trained head teachers?	• Percentage change in number of head teachers receiving INSET on school	• The number of sample TDP head teachers who received TDP SLM in-service training during the last three school years as a proportion of all	Tabulation from survey data;	TDP evaluation quantitative surveys – head teacher	State

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
			management training		leadership and management (SLM) • Head teacher training attendance	sample TDP head teachers multiplied by 100	review of programme documents	interviews; implementation review	
Effe-28	Effectiveness	Output		• What were the synergies between the TDP outputs?				All	State
TOC LEVEL: OUTPUT TO OUTCOME ASSUMPTIONS									
Effe-19	Effectiveness	Output to outcome assumption	• Teacher educators and TFs are able to absorb and then transfer the skills offered in the TDP training	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Teacher educator and TF perception of capacity to absorb training		Inference from qualitative data	TDP evaluation qualitative research – TF interviews, training observations, teacher interviews	State
Effe-20	Effectiveness	Output to outcome assumption	• Teacher educators and facilitators have sufficient time and material support to provide support to schools (i.e. they are not directed to other tasks, transferred, or made redundant)	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Teacher educator and TF perception of adequate time and support		Inference from qualitative data	TDP evaluation qualitative research – TF interviews, LGEA interviews, SUBEB interviews	State
Effe-21	Effectiveness	Output to outcome assumption	• Head teachers have appropriate incentives/motivation to apply their new knowledge in support of teachers	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Head teacher perceptions of incentives/motivation		Inference from qualitative data	TDP evaluation qualitative research – head teacher interviews, teacher interviews	State
Effe-22	Effectiveness	Output to outcome assumption	• Teachers and head teachers attend training regularly (i.e. they have the time and are managed and supported to so,	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Proportion of teachers and head teachers receiving TDP training • Average number of days teachers and	• The number of sample TDP teachers and head teachers who received TDP in-service training during the last three school years as a proportion of all sample TDP teachers multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – head teacher and teacher interviews	State

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
			and not transferred during the training process)		head teachers attended the TDP training per year	• Number of days each sample TDP teacher/head teacher attended the TDP training per school year in the last three years			
Effe-23	Effectiveness	Output to outcome assumption	• Teachers can access and use the TDP audio-visual materials (i.e. the technology works, can be charged, is not lost, stolen or broken, is upgraded or fixed where appropriate, can be understood, etc.)	• What factors facilitated or inhibited TDP's achievement of its outcomes?		•	Tabulation from survey data	TDP evaluation quantitative surveys – head teacher interviews, teacher interviews	State
					• Proportion of TDP teachers who received a mobile phone from TDP (%)	• Number of sample TDP teachers who received a mobile phone from TDP as a proportion of all sample TDP teachers multiplied by 100			
					• Proportion of TDP teachers who received an SD card from TDP (%)	• Number of sample TDP teachers who received an SD card from TDP as a proportion of all sample TDP teachers multiplied by 100			
					• Proportion of TDP teachers for whom the TDP SD card was compatible with their mobile phone (%)	• Number of sample TDP teachers for whom the TDP SD card was compatible with the TDP mobile phone as a proportion of sample TDP teachers who received a TDP mobile phone and SD card multiplied by 100			
					• Proportion of TDP teachers with TDP teacher's videos on teaching methods (%)	• Number of sample TDP teachers who have TDP teacher's videos on teaching methods as a proportion of sample TDP teachers multiplied by 100			
					• Reasons for teachers not having access to TDP teacher's videos on teaching methods (%)	• The number of sample TDP teachers reporting each reason they do not have access to TDP teacher's videos on teaching methods (not given the TDP videos, lost the TDP videos, not possible to play the TDP videos on teacher's mobile phone, other) as a proportion of the total number of sample TDP teachers who reported not having access to the videos multiplied by 100			
					• Proportion of TDP teachers with access TDP audio materials to use in the classroom (%)	• Number of sample TDP teachers who have access to TDP audio clips to use with pupils in the classroom as a proportion of sample TDP teachers multiplied by 100			

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
					• Reasons for teachers not having access to TDP audio materials to use in the classroom (%)	• The number of sample TDP teachers reporting each reason they do not have access to TDP audio materials to use in the classroom (not given the TDP audio clips, lost the TDP audio clips, not possible to play the TDP audio clips on teacher's mobile phone, other) as a proportion of the total number of sample TDP teachers who reported not having access to the audio materials multiplied by 100			
					• Proportion of schools with working amplifier (%)	• The number of sample TDP schools that have at least one amplifier provided by TDP as a proportion of all sample TDP schools multiplied by 100 • The number of sample TDP schools that have at least one amplifier that <u>works</u> and was provided by TDP as a proportion of all sample TDP schools multiplied by 100			
					• Reasons for TDP amplifier not working (%)	• The number of sample TDP head teachers reporting each reason the TDP amplifier(s) are <u>not</u> working (faulty, physical broken, cannot charge it, other) as a proportion of the total number of sample TDP head teachers reporting the school has a TDP amplifier that is <u>not</u> working multiplied by 100			
					• Proportion of TDP schools with regular electricity supply (%)	• Number of sample TDP schools that have regular electricity supply as a proportion of all sample TDP schools multiplied by 100			
Effe-24	Effectiveness	Output to outcome assumption	• 2.5 days' training, monthly cluster meetings, and the support visits and	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• TF, head teacher, and teacher perceptions of adequacy of training		Inference from qualitative data;	TDP evaluation qualitative research – TF interviews,	

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
			mentoring are sufficient to instil new pedagogical knowledge in teachers		• Proportions of teachers reporting the TDP training very useful, somewhat useful or not useful (% TDP-trained teachers)	• The number of TDP-trained sample teachers reporting they found the TDP teaching training very useful/somewhat useful/not useful as a proportion of all TDP-trained sample teachers multiplied by 100	tabulation from survey data	teacher interviews, head teacher interviews; TDP evaluation quantitative surveys – teacher interviews	
					• Gains from TDP teaching training (% TDP-trained teachers)	• The number of TDP-trained sample teachers reporting each type of gain (curriculum knowledge, subject knowledge, sound groups in English, teaching methods, inclusive teaching methods, classroom management skills, lesson planning skills, development of instructional materials, assessment and monitoring of pupil learning, use of ICT during lessons, confidence in their teaching, support network, other, nothing) from the TDP teaching training as a proportion of all TDP-trained sample teachers multiplied by 100			
					• Difficulties with TDP teaching training (% TDP-trained teachers)	• The number of TDP-trained sample teachers reporting each type of difficulty (not relevant to my job, materials difficult to understand, too much content, too theoretical, ignored the reality of the teaching environment, took up too much time, other, none) from the TDP teaching training as a proportion of all TDP-trained sample teachers multiplied by 100			
Effe-25	Effectiveness	Output to outcome assumption	• Teachers have the basic language, subject, and pedagogical skills to absorb the new knowledge and skills available from TDP	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Average teacher raw scores on TDNA for English, maths, and science and technology	• Number of correct answers as a proportion of the maximum score on each of the TDNA English, maths, and science and technology components multiplied by 100	Tabulation from survey data	TDP evaluation quantitative surveys – TDNA	State

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
Effe-26	Effectiveness	Output to outcome assumption	• The TDP materials are appropriate for pupils of different abilities, particularly around language	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Teacher and pupil perceptions of adequacy of materials		Inference from qualitative data	TDP evaluation qualitative research – teacher interviews, classroom observations, pupil interviews	
Effe-27	Effectiveness	Output to outcome assumption	• There are no changes to the curriculum or other features of the education system that render the TDP materials redundant	• What factors facilitated or inhibited TDP's achievement of its outcomes?	• Exogenous changes		Inference from qualitative data	TDP evaluation qualitative research – SUBEB interviews; policy documents	
TOC LEVEL: SCALE IMPACT									
Im-6	Impact	Scale impact	• TDP model applied in other schools in TDP states	• Are there any indications that the TDP model is applied in other schools in TDP states?	• Perceptions of application of model in other schools in TDP states		Inference from qualitative data	TDP evaluation qualitative research – SUBEB, LGAs, FMOE, TDP staff interviews; policy documents	
TOC LEVEL: SCALE ASSUMPTIONS									
Re-21	Relevance	Scale assumption	• SUBEBs remain the appropriate institutional home for INSET	• Is this assumption correct in the Nigerian context?	• Perception of SUBEBs as appropriate institutional homes for INSET in each state		Inference from qualitative data	TDP evaluation qualitative research – SUBEB, FMOE, TDP staff interviews; policy documents	
Im-20	Impact			• What factors facilitated or inhibited TDP's achievement of its impacts?					
Re-22	Relevance	Scale assumption	• State governments/SUBEBs consider in-service teacher training a priority, and the TDP model	• Is this assumption correct in the Nigerian context? • What factors facilitated or inhibited TDP's	• Do state governments/SUBEBs consider improving in-service teacher training through the TDP model a priority?		Inference from qualitative data	TDP evaluation qualitative research – SUBEB, FMOE interviews;	
Im-21	Impact								

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
			appropriate, on a sustained basis	achievement of its impacts?				policy documents	
Re-23	Relevance	Scale assumption	• SUBEBs have the capacity (managerial, technical and financial) to apply TDP model in other schools	• Is this assumption correct in the Nigerian context? • What factors facilitated or inhibited TDP's achievement of its impacts?	• Has TDP adequately designed an approach to account for variations in states' capacity to adopt TDP? • Perceptions of SUBEB capacity to apply TDP model		Inference from qualitative data	TDP evaluation qualitative research – SUBEB, FMOE interviews; policy documents	
Im-22	Impact								
TOC LEVEL: SUSTAINABILITY IMPACT									
Su-23	Sustainability	Sustainability impact	• TDP model applied sustainably in TDP schools	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Perceptions of sustainability of TDP approach and practice		Inference from qualitative data	TDP evaluation qualitative research – interviews with FMOE, SUBEBs, LGEAs, TFs, head teachers and teachers	
TOC LEVEL: SUSTAINABILITY ASSUMPTIONS									
Re-24	Relevance	Sustainability assumption	• TDP's partner institutions are the appropriate institutional homes for an in-service teacher training model	• Is this assumption correct in the Nigerian context? • Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Are FMOE, UBEC, and NERDC appropriate custodians of in-service teacher training in Nigeria?		Inference from qualitative data	TDP evaluation qualitative research – FMOE, SUBEB, UBEC, NERDC, TDP staff, interviews; policy documents	
Su-24	Sustainability								
Re-25	Relevance	Sustainability assumption	• Reform to teacher effectiveness is an appropriate policy priority to improve learning in Nigeria	• Is this assumption correct in the Nigerian context? • Are the TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Teacher effectiveness reform is a policy priority in Nigeria		Inference from qualitative data	TDP evaluation qualitative research – FMOE, SUBEB, UBEC, NERDC, TDP staff, interviews; policy documents	
Su-27	Sustainability								

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
Su-25	Sustainability	Sustainability assumption	• TDP's partner institutions are open to these partnerships being developed, on a continuous basis (e.g. in a way that is resilient to changes in their leadership and political standing)	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Perceptions of institutional capacity to apply TDP model		Inference from qualitative data	TDP evaluation qualitative research – TDP staff, FMOE, UBEC, NERDC, SUBEB interviews, policy documents	
Su-26	Sustainability	Sustainability assumption	• TDP's partner institutions have the capacity (broadly defined to include e.g. funding, human resources, management capacity, skills, etc.) to engage in the INSET model	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Perception of institutional capacity to apply TDP model		Inference from qualitative data	TDP evaluation qualitative research – TDP staff, FMOE, UBEC, NERDC, SUBEB interviews, policy documents	
Su-28	Sustainability	Sustainability assumption	• Reform to in-service teacher training is an appropriate policy priority to improve teacher effectiveness in Nigeria	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• In-service teacher training reform is a priority for teacher effectiveness reform in Nigeria		Inference from qualitative data	TDP evaluation qualitative research – TDP staff, FMOE, UBEC, NERDC, SUBEB interviews, policy documents	
Su-29 (*TBC*)	Sustainability	Sustainability assumption	• A single in-service teacher training model is appropriate for all different states in Nigeria (rather than different models in different states)	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Perceptions of appropriateness of TDP model in different states		Inference from qualitative data	TDP evaluation qualitative research – TDP staff, FMOE, UBEC, NERDC, SUBEB interviews, policy documents	
Su-30	Sustainability	Sustainability assumption	• The TDP in-service teacher training model is	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Pupil learning levels in English, maths, and science	• IRT-based scale scores in English, maths, and science	Impact estimation	TDP evaluation quantitative	

Ref.	Evaluation criterion	TOC level	TOC detail	Evaluation question	Indicator	Quantitative indicator definition (levels)	Analytical approach	Information collection methods	Disaggregation
			appropriate in Nigeria, and improves teacher effectiveness and learning		• Change between P3 (2014) and P6 (2017) in proportions of pupils in different performance bands in English, maths, and science, in TDP and control schools	• Proportion of pupils in the bottom and top performance bands in each of English, maths, and science and technology, respectively (pupils tested in P3 at baseline and in P6 at endline)		surveys – pupil tests	
Su-31	Sustainability	Sustainability assumption	• TDP training materials can be updated to reflect changes to the curriculum, language of instruction, etc.	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Flexibility of TDP training materials		Review of TDP materials	TDP documents	
Su-32	Sustainability	Sustainability assumption	• The technology remains robust to technological and infrastructural change and can be maintained in working order	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Longevity of technology		Inference from qualitative data	TDP evaluation qualitative research – head teacher interviews, teacher interviews	
Su-33	Sustainability	Sustainability assumption	• The SUBEBs and LGEAs have the incentives and capacity (managerial, financial, technical) to maintain, support and renew the teacher educator teams without TDP support	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Perceptions of SUBEB and LGEA capacity to maintain TDP approach and practice		Inference from qualitative data	TDP evaluation qualitative research – FMOE, SUBEB, LGEA, UBEC interviews	State
Su-34	Sustainability	Sustainability assumption	• Head teachers and teachers have the interest and capacity to continue to support the TDP INSET model	• Are TDP's impacts likely to be sustainable when DFID ends funding in 2019?	• Perceptions of head teachers' and teachers' interest and capacity to maintain TDP approach and practice		Inference from qualitative data	TDP evaluation qualitative research – head teacher interviews, teacher interviews	State

Annex B Detailed statistical annexes

B.1 The TDP intervention

Annex Table 1: In-service TDP teaching and TDP Reading Programme training coverage, length, content, and perceptions

	Treatment						
	Endline						
Indicator[1]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N
TDP teaching training							
Attended TDP teaching training in last three years (% of all P1–P6 teachers excluding head teachers)	39.5			2.7	34.3	44.8	1,887
Attended TDP teaching training in last three years (% of sample teachers and head teachers)	92.3			1.4	89.6	95.0	318
Attended TDP teaching training by school year (% TDP-trained teachers and head teachers)							
2016/17	46.1			2.8	40.5	51.6	296
2015/16	54.8			2.7	49.4	60.1	296
2014/15	67.2			2.4	62.4	71.9	296
Average number of days of TDP teaching training attended by sample teachers and head teachers							
2016/17	9.7	3.0	20.0	0.6	8.4	10.9	115
2015/16	11.5	3.0	24.0	0.7	10.1	13.0	131
2014/15	12.0	3.0	24.0	0.6	10.8	13.3	155
Contents of TDP teaching training (% TDP-trained teachers and head teachers)							
Teaching methods	75.4			2.3	70.8	80.0	296
English literacy	54.3			2.7	49.0	59.6	296
Hausa literacy	10.0			1.8	6.4	13.5	296
Numeracy/maths	51.7			2.7	46.2	57.1	296
Science	28.9			2.5	24.0	33.8	296
Other curriculum subject	7.7			1.5	4.7	10.6	296
Inclusive teaching	32.6			2.4	28.0	37.3	296

Different sounds groups (in English)	18.9			2.1	14.7	23.0	296
Phonics methods to teach letter sounds, letter formation, and blending	13.9			1.9	10.1	17.6	296
Development of instructional materials	42.0			2.9	36.3	47.7	296
Assessment and monitoring of pupil learning	25.1			2.5	20.1	30.2	296
Use of ICT during lessons (e.g. use of audio-visual materials)	13.4			2.1	9.4	17.5	296
Other	8.2			1.7	4.8	11.5	296
Considers TDP teaching training (% TDP-trained teachers and head teachers)							
Very useful	99.5			0.3	98.9	100.1	296
Somewhat useful	0.5			0.3	-0.1	1.1	296
Not useful	0.0			0.0	0.0	0.0	296
Gain from TDP teaching training (% TDP-trained teachers and head teachers)							
Curriculum knowledge	24.6			2.4	19.9	29.4	296
Subject knowledge	44.8			2.9	39.0	50.5	296
Different sounds groups (in English)	23.8			2.1	19.6	28.0	296
Teaching methods	82.0			2.1	77.8	86.3	296
Inclusive teaching methods	38.4			2.4	33.7	43.0	296
Classroom management	55.0			2.7	49.7	60.4	296
Lesson planning skills	45.7			2.6	40.5	50.8	296
Development of instructional materials	43.8			2.7	38.4	49.1	296
Assessment and monitoring of pupil learning	24.7			2.6	19.6	29.8	296
Use of ICT during lessons (e.g. use of audio-visual materials)	12.1			2.1	8.0	16.2	296
Confidence in my teaching	20.5			2.0	16.5	24.5	296
Support network	2.2			0.8	0.5	3.9	296
Other	6.0			1.5	3.1	9.0	296
Nothing	0.0			0.0	0.0	0.0	296
Difficulties with TDP teaching training (% TDP-trained teachers and head teachers)							
Not relevant to my job	2.5			0.9	0.8	4.2	296
Materials difficult to understand	5.7			1.1	3.5	7.9	296
Too much content	4.0			1.0	2.1	5.9	296
Too theoretical	2.4			0.8	0.8	3.9	296

Ignored the reality of the teaching environment	1.1			0.5	0.1	2.2	296
Took up too much time	12.1			2.0	8.1	16.0	296
Other	27.0			2.5	22.1	31.8	296
No difficulties	52.7			2.5	47.8	57.6	296
TDP Reading Programme training							
Attended TDP Reading Programme training in 2016/17 (% of all P1–P6 teachers excluding head teachers)	47.1			3.0	41.2	52.9	1,887
Attended TDP Reading Programme training in 2016/17 (% of sample teachers and head teachers)	55.9			2.9	50.3	61.5	318
Average number of days of TDP Reading Programme training attended in 2016/17 by sample teachers and head teachers	8.7	3.0	14.0	0.4	7.9	9.5	173
Source: TDP impact evaluation endline survey, teacher roster, head teacher interview, and teacher interview.							
Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator.							

Annex Table 2: Access to TDP materials and technology

	Treatment						
	Endline						
Indicator[1]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N
Have access to (% TDP-trained teachers and head teachers)							
Teacher's guide in any subject	93.7			1.5	90.7	96.7	296
Teacher's guide in English	85.0			2.1	80.9	89.1	296
Teacher's guide in mathematics	80.4			2.2	76.1	84.7	296
Teacher's guide in science and technology	59.1			2.5	54.0	64.1	296
Teacher's guide in Pedagogy	55.0			2.5	50.0	60.0	296
Lesson plan in English	78.1			2.5	73.1	83.1	296
Lesson plan in maths/numeracy	75.5			2.4	70.7	80.4	296
Lesson plan in science and technology	37.6			2.4	32.8	42.4	296

Flash cards English/literacy	84.1			2.1	80.0	88.2	296
Flash cards maths/numeracy	78.3			2.2	74.0	82.7	296
Flash cards Hausa	25.2			2.4	20.5	29.9	296
Reading Assessment Guide	55.2			3.0	49.4	61.1	293
Considers TDP teacher's guide (% TDP-trained teachers and head teachers with guide and have used it)							
Useful	99.5			0.3	98.8	100.1	279
Somewhat useful	0.3			0.2	-0.2	0.7	279
Not useful	0.3			0.2	-0.2	0.7	279
TDP teacher's guide used to (% TDP-trained teachers and head teachers with guide)							
Prepare lessons	81.4			2.3	76.8	86.0	280
Teach during my lessons	83.2			2.2	78.8	87.6	280
Not used yet	0.2			0.2	-0.1	0.6	280
Other use	1.5			0.8	-0.1	3.0	280
Considers TDP lesson plans (% TDP-trained teachers and head teachers with lesson plan and have used it)							
Useful	98.3			0.3	97.7	99.0	269
Somewhat useful	1.7			0.3	1.0	2.3	269
Considers TDP lesson plan length appropriate given lesson length (% TDP-trained teachers and head teachers with lesson plan and have used it)	75.7			2.3	71.2	80.2	269
Considers TDP lesson plan contents appropriate given curriculum (% TDP-trained teachers and head teachers with lesson plan and have used it)	85.4			1.8	81.8	89.1	269
TDP lesson plans used to (% TDP-trained teachers and head teachers with lesson plan)							
Prepare lessons	85.9			2.1	81.7	90.0	270
Teach during lessons	85.4			2.1	81.2	89.6	270
Not used yet	0.2			0.2	-0.1	0.6	270
Other use	0.8			0.7	-0.5	2.2	270
Considers TDP reading assessment guide (% TDP-trained teachers and head teachers with guide and have used it)							
Useful	99.3			0.7	98.0	100.6	162
Somewhat useful	0.7			0.7	-0.6	2.0	162
TDP reading assessment guide used to (% TDP-trained teachers and head teachers with guide)							
Prepare lessons	60.6			3.5	53.5	67.6	166

Teach during my lessons	72.2			3.8	64.5	79.9	166
Assess pupils' reading skills	53.6			4.9	43.8	63.4	166
Not used yet	2.8			1.4	-0.1	5.7	166
Other use	3.0			1.7	-0.4	6.4	166
Considers TDP flash cards (% TDP-trained teachers and head teachers with flash cards)							
Useful	100.0			0.0	100.0	100.0	274
Uses TDP flash cards for teaching (% TDP-trained teachers and head teachers with flash cards)	98.4			0.6	97.1	99.6	274
Have access to teacher's videos on teaching methods (% TDP-trained teachers and head teachers)	92.6			1.5	89.7	95.5	296
No access to TDP videos because (% TDP-trained teachers and head teachers with no access to videos)							
Not given TDP teacher's videos	60.7			21.7	-8.5	129.8	21
Lost TDP teacher's videos	5.4			0.9	2.6	8.1	21
Cannot play TDP teacher's videos on mobile phone	9.0			17.5	-46.6	64.6	21
Other	25.0			13.6	-18.3	68.3	21
Have access to audio materials for classroom use (% TDP-trained teachers and head teachers)	92.4			1.5	89.4	95.4	296
No access to TDP audio materials because (% TDP-trained teachers and head teachers with no access to audio material)							
Not given TDP audio materials	64.9			20.5	-0.5	130.2	23
Lost TDP audio materials	2.0			0.3	1.0	3.0	23
Cannot play TDP audio materials on mobile phone	8.8			16.5	-43.6	61.2	23
Other	24.4			12.8	-16.5	65.2	23
TDP-trained teachers and head teachers (%)							
Own working smartphone	97.9			0.8	96.4	99.4	296
Given TDP mobile phone	91.1			1.5	88.2	94.0	296
Given TDP SD card	81.3			2.2	76.9	85.6	296
SD card compatible with TDP mobile phone (% TDP-trained teachers and head teachers with phone and SD card)	97.9			1.0	95.9	99.9	233
Has one or more TDP amplifiers, working and non-working (% schools)	92.8			1.6	89.6	96.0	165
Has one or more working TDP amplifiers (% schools)	47.7			3.0	41.8	53.6	165
Location of TDP amplifier(s) (% schools with TDP amplifier(s), working and non-working)							
Head teacher's office	83.4			2.4	78.6	88.2	153
Staff room	2.6			1.1	0.5	4.7	153

Classroom	0.7			0.5	-0.4	1.8	153
Storage away from the school	12.8			2.3	8.4	17.3	153
Other	3.1			1.0	1.1	5.2	153
Reason TDP amplifier not working (% schools with non-working TDP amplifier)							
Physically broken	1.4			1.4	-1.4	4.2	73
Cannot charge the amplifier	71.0			4.5	62.0	80.0	73
Faulty	25.1			4.5	16.0	34.1	73
Other	2.6			1.8	-1.0	6.1	73
Has electricity (% schools)	9.5			1.9	5.9	13.2	165
Source: TDP impact evaluation endline survey, head teacher interview, and teacher interview.							
Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator.							

Annex Table 3: TDP teacher support through SSVs

	Treatment						
	Endline						
Indicator[1]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N
Attended TDP English Reading Club last school year (% TDP-trained teachers and head teachers)	12.3			1.6	9.1	15.5	295
Head teacher reporting on SSVs							
Received SSV last term from TFs (% schools)	91.2			1.7	87.9	94.6	162
Average number of SSVs last term (% schools that received SSV)							
One	6.8			1.6	3.6	10.0	147
Two	25.9			3.0	20.0	31.8	147
Three	26.3			2.9	20.5	32.1	147
Four	12.2			2.2	7.7	16.6	147
More than four	28.9			2.9	23.1	34.7	147
TF's activities during last SSV (% schools that received SSV)							
Conducted lesson observations	87.7			2.3	83.1	92.3	148
Feedback to teachers after lesson observations	72.4			3.0	66.5	78.3	148
Held interview and feedback session with head teacher	41.2			3.4	34.6	47.8	148
Provided in-school support to head teacher	27.4			3.0	21.5	33.3	148
Provided in-school support to teachers	31.5			3.0	25.5	37.4	148
Met with SBMC	8.5			1.9	4.8	12.1	148
Other	4.6			1.4	1.8	7.4	148
Teachers reporting on SSVs							
Received SSV last term from TFs (% teachers)	89.9			1.8	86.3	93.4	220
Average number of SSVs last term (% teachers that reported school receiving SSV)							
One	10.7			2.3	6.2	15.3	192
Two	20.3			3.4	13.6	27.0	192
Three	32.1			3.2	25.7	38.5	192

Four	13.0			2.4	8.3	17.7	192
More than four	23.9			2.4	19.1	28.7	192
TF activities during last SSV (% teachers that reported school receiving SSV)							
Conducted lesson observations	86.2			2.7	80.8	91.7	201
Feedback to teachers after lesson observations	77.5			3.0	71.6	83.4	201
Held interview and feedback session with head teacher	33.5			3.1	27.2	39.7	201
Provided in-school support to head teacher	20.6			2.8	15.0	26.2	201
Provided in-school support to teachers	32.7			3.5	25.7	39.7	201
Met with SBMC	6.1			1.4	3.3	8.9	201
Other	5.7			1.9	2.0	9.5	201
Source: TDP impact evaluation endline survey, head teacher interview, and teacher interview.							
Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator.							

Annex Table 4: In-service TDP SLM training coverage, contents, and perceptions

	Treatment						
	Endline						
Indicator[1]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N
Received TDP SLM training (% head teachers)	90.0			1.8	86.4	93.6	164
Contents of TDP SLM training (% TDP-trained head teachers)							
Academic leadership	62.4			3.0	56.3	68.4	147
School leadership, role of the head teacher	88.5			2.1	84.3	92.7	147
School development planning	47.7			3.4	41.0	54.4	147
Pupil assessment	40.4			2.9	34.5	46.2	147
Record-keeping, reporting, information systems	54.0			3.1	47.8	60.1	147
School budgeting, financial management	28.7			2.9	22.9	34.4	147

School self-evaluation	33.3			3.0	27.2	39.3	147
Staff development	24.8			3.0	18.9	30.7	147
Teacher management	52.1			3.4	45.3	58.8	147
Teacher mentoring	40.4			3.2	34.1	46.7	147
Other	10.3			2.1	6.2	14.4	147
Considers TDP SLM training (% TDP-trained head teachers)							
Very useful	99.3			0.6	98.2	100.4	147
Somewhat useful	0.7			0.6	-0.4	1.8	147
Gain from TDP SLM training (% TDP-trained head teachers)							
Academic leadership	68.8			3.0	62.8	74.7	147
School leadership	81.2			2.5	76.4	86.1	147
Teacher management	73.3			3.0	67.4	79.1	147
Staff development	40.8			3.4	34.1	47.6	147
Confidence in my role as the head teacher	58.6			3.3	52.0	65.1	147
Support network	14.5			2.2	10.1	18.9	147
Nothing	0.0			0.0	0.0	0.0	147
Other	13.6			2.1	9.4	17.8	147
Difficulties with TDP SLM training (% TDP-trained head teachers)							
Not relevant to my job	2.7			1.1	0.5	4.9	147
Materials difficult to understand	3.4			1.2	1.0	5.9	147
Too much content	2.1			1.0	0.1	4.0	147
Too theoretical	3.9			1.2	1.5	6.3	147
Ignored reality of the teaching environment	0.7			0.6	-0.4	1.8	147
Took up too much time	3.5			1.3	1.0	6.0	147
No difficulties	70.2			3.0	64.3	76.1	147
Other	17.5			2.6	12.5	22.6	147
Have TDP head teacher handbook (% TDP-trained head teachers)	91.2			1.6	88.0	94.4	147
Considers head teacher handbook useful (% TDP-trained head teachers with handbook)	100.0			0.0	100.0	100.0	134

Source: TDP impact evaluation endline survey, head teacher interview.

Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator.

B.2 Context for TDP

Annex Table 5: Pupils, characteristics, and home environment

	Treatment							Control						
	Endline							Endline						
Indicator[1,2]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N
Pupil demographics														
Female (% pupils)	41.5			2.4	36.9	46.1	783	40.2			2.5	35.3	45.1	782
Average age (years)	11.6	10.0	13.0	0.1	11.4	11.8	729	11.7	10.0	13.0	0.1	11.6	11.9	724
Appropriate age for Grade 6 (% pupils)	59.5			2.6	54.3	64.7	729	57.7			2.6	52.6	62.7	724
Under-age for Grade 6 (% pupils)	21.2			2.2	16.9	25.5	729	17.9			2.1	13.8	21.9	724
Over-age for Grade 6 (% pupils)	19.3			2.5	14.3	24.3	729	24.5			2.5	19.6	29.3	724
Socio-economic background														
Type of toilet/latrine facilities household has access to (% pupils)														
Flush toilet	20.7			2.1	16.5	24.9	780	21.2			2.4	16.4	26.0	779
Pit latrine	77.0			2.1	72.8	81.1	780	75.3			2.4	70.6	80.0	779
Bush/field	2.3			0.6	1.2	3.5	780	3.5			0.7	2.0	4.9	779
Average family size	11.8	6.0	19.0	0.3	11.2	12.3	766	12.5	6.0	21.0	0.4	11.8	13.2	765
Average number of family members less than 15 years	4.9	2.0	8.0	0.1	4.7	5.2	736	5.2	2.0	8.0	0.2	4.9	5.5	737
Father lives in same house all or some of the time (% pupils)	89.1			1.7	85.8	92.4	782	90.9			1.6	87.7	94.0	781
Father attended school (% pupils whose father lives in same house)	86.0			2.3	81.5	90.5	678	87.7			1.7	84.3	91.2	687
Father completed primary school (% pupils whose father attended school)	96.0			0.8	94.4	97.7	558	93.3*			1.1	91.1	95.6	572

Sometimes father writes something down (% pupils whose father lives in same house)	84.2			2.1	80.1	88.2	715	83.1			2.1	79.0	87.2	714
Sometimes father reads a book or newspaper (% pupils whose father lives in same house)	77.1			2.5	72.2	82.0	713	80.0			2.4	75.3	84.6	714
Mother lives in same house all or some of the time (% pupils)	85.6			1.7	82.1	89.0	783	90.4**			1.6	87.3	93.4	782
Mother attended school (% pupils whose mother lives in same house)	73.1			2.3	68.5	77.7	621	65.6**			2.9	59.9	71.2	633
Mother completed primary school (% pupils whose mother attended school)	81.9			3.7	74.5	89.2	402	81.2			3.1	75.2	87.3	378
Sometimes mother writes something down (% pupils whose mother lives in same house)	61.7			2.3	57.1	66.4	688	59.3			2.7	54.0	64.6	692
Sometimes mother reads a book or newspaper (% pupils whose mother lives in same house)	66.5			2.7	61.1	71.9	689	61.4			3.1	55.3	67.6	692
Pupil nutrition														
Normally eat at home before school (% pupils)	90.5			1.6	87.4	93.5	783	92.1			1.4	89.5	94.8	782
Normally eat something during long break (% pupils)	84.4			1.6	81.3	87.6	782	84.6			1.7	81.3	87.8	782
Education support at home														
Number of books at home (% pupils)														
None	13.3			1.6	10.2	16.5	779	13.3			1.9	9.5	17.1	778
1–2	5.8			1.2	3.5	8.1	779	4.0			1.0	2.1	5.9	778
3–10	44.6			2.8	39.2	50.1	779	40.8			2.5	35.8	45.8	778
More than 10	31.3			2.6	26.3	36.4	779	38.9**			2.7	33.5	44.2	778
Don't know	4.9			1.2	2.6	7.3	779	3.1			0.7	1.6	4.5	778
Receive help with homework at home (% pupils)														
Never	22.8			2.2	18.5	27.0	778	24.3			2.4	19.6	29.1	776
Less than once a month	9.0			1.7	5.7	12.2	778	9.3			1.8	5.6	12.9	776
Sometimes	51.3			2.8	45.7	56.9	778	44.7			3.0	38.8	50.5	776
More than once a week	17.0			2.0	13.1	20.9	778	21.8			2.3	17.2	26.3	776

Source: TDP impact evaluation endline survey, pupil test, and background instrument.

Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator. (2) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Annex Table 6: Teachers, characteristics, and class readiness

	Treatment													Control												
	Baseline							Endline							Baseline							Endline							
Indicator[1,2]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Raw DID
Characteristics of panel teachers and head teachers who teach P1–P6																												
Average age (years)	38.2	29.0	49.0	0.4	37.5	38.9	275	41.3***	31.0	53.0	0.4	40.6	42.1	275	37.7	27.0	50.0	0.6	36.6	38.8	216	41.0***	30.0	53.0	0.6	39.8	42.3	216	-0.2
Average teaching experience (years)	12.8	3.0	24.0	0.4	12.0	13.6	274	15.9***	7.0	27.0	0.4	15.1	16.7	274	13.1	4.0	27.0	0.5	12.1	14.1	217	16.2***	7.0	30.0	0.5	15.2	17.2	217	0.0
Average time in current school (years)	5.2	0.0	12.0	0.3	4.7	5.8	272	8.3***	3.0	15.0	0.2	7.8	8.8	272	5.9	1.0	13.0	0.4	5.2	6.6	209	8.6***	4.0	15.0	0.4	7.8	9.3	209	0.4
Female (%)	12.0			1.9	8.3	15.7	277	12.0			1.9	8.3	15.7	277	20.9			2.9	15.1	26.7	217	21.2			2.9	15.4	27.1	217	-0.3
Have NCE qualification (%)	67.3			2.8	61.8	72.8	277	73.2***			2.3	68.7	77.8	277	67.3			2.8	61.8	72.8	217	73.4***			2.5	68.5	78.4	217	-0.2
Have senior school certificate examination (SSCE) qualification (%)	30.9			2.5	26.0	35.8	277	54.5***			2.7	49.2	59.7	277	37.8			3.4	31.0	44.6	217	57.5***			3.1	51.3	63.6	217	3.9
Have Grade II qualification (%)	44.5			3.1	38.3	50.7	277	54.1***			2.8	48.5	59.7	277	47.0			3.3	40.4	53.6	217	53.6***			3.2	47.3	59.9	217	3.0
Received teaching training in last two years (baseline) or last three years (endline) (%)	51.0			2.4	46.3	55.7	277	97.4***			1.1	95.3	99.6	277	54.7			3.2	48.3	61.0	217	74.8***			2.6	69.6	80.0	217	26.3***
Training received by all panel teachers and head teachers																												
Received teaching training in last three years by provider (% panel teachers and head teachers)																												
TDP								92.3			1.4	89.6	95.0	318								29.5***			2.4	24.8	34.1	263	

TDP Reading Programme on sound groups								55.9			2.9	50.3	61.5	318								9.1***			1.7	5.7	12.4	263	
SUBEB based on TDP model								6.5			1.4	3.8	9.2	318								3.8			1.0	1.8	5.8	263	
UNICEF RANA early learning intervention								2.9			0.8	1.3	4.6	318								8.5***			1.9	4.8	12.2	263	
ESSPIN								14.8			1.7	11.4	18.2	318								25.2***			1.8	21.7	28.7	263	
Jolly Phonics								26.4			2.1	22.4	30.5	318								22.3			2.1	18.3	26.4	263	
The Global Partnership for Education								14.6			1.8	11.0	18.2	318								12.9			2.0	9.1	16.8	263	
Other SUBEB (not based on TDP model)								4.9			1.2	2.5	7.2	318								4.5			1.1	2.2	6.7	263	
UBEC								0.6			0.3	-0.1	1.2	318								0.7			0.6	-0.5	1.9	263	
National Teachers Institute								3.4			0.9	1.7	5.2	318								4.2			1.2	1.7	6.6	263	
Local government (LGEA)								3.4			1.2	1.1	5.7	318								0.6**			0.3	-0.1	1.2	263	
Other								6.6			1.3	4.0	9.1	318								7.7			1.5	4.8	10.6	263	
Teacher turnover and reasons																													
No longer at school at endline (% sample baseline teachers)								32.6			1.8	29.1	36.1	452								48.4***			2.3	43.9	52.8	450	
Will retire by endline (% sample baseline teachers)								1.9			0.5	0.9	2.9	453								2.7			0.7	1.2	4.2	450	
Transferred to another school since last school year (mean % teachers in a school)								6.5	0.0	16.7	0.9	4.8	8.3	134								6.6	0.0	20.0	0.8	5.0	8.1	133	
Source: TDP impact evaluation baseline and endline surveys, teacher and head teacher interviews, baseline-endline survey records.																													
Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator. (2) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1.																													

Annex Table 7: Schools, infrastructure, and the teaching and learning environment

	Treatment													Control													
	Baseline							Endline							Baseline							Endline							
Indicator[1,2]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Raw DID
School infrastructure																													
Has at least one toilet (% schools)								87.4			2.0	83.5	91.3	165								87.7			2.0	83.8	91.7	164	
Average number of pupils per toilet in a school								147.8	28.9	327.0	9.0	130.0	165.6	143								145.8	30.0	333.3	9.6	126.9	164.7	141	
Drinking water in school (% schools)								55.6			2.9	49.8	61.3	165								52.0			2.8	46.4	57.6	164	
Has a staff room (% schools)								46.6			3.1	40.5	52.7	165								41.4			3.1	35.3	47.6	164	
Average number of classrooms in a school								8.1	3.0	14.0	0.3	7.5	8.7	165								8.4	4.0	16.0	0.4	7.6	9.1	164	
Average number of classrooms in use in a school								7.4	3.0	13.0	0.3	6.8	7.9	165								7.4	3.0	12.0	0.3	6.7	8.1	164	
Average number of pupils per classroom in use in a school								90.3	43.3	151.3	2.8	84.8	95.8	164								95.8	41.3	166.2	5.8	84.5	107.2	160	
Teaching and learning environment from school records																													
Average number of P1–P6 pupils registered in a school	639.8	145.0	1489.0	33.2	574.2	705.4	163	7000.0***	174.0	16000.0	34.7	631.3	768.7	163	662.9	133.0	1524.0	38.4	586.7	739.0	161	692.5	196.0	1486.0	39.7	613.9	771.2	161	30.5
Average number of P1–P6 teachers employed in a school (excl. voluntary and temporary teachers)	12.4	4.0	26.0	0.7	11.0	13.9	165	11.5***	3.0	25.0	0.7	10.1	12.8	165	12.2	4.0	25.0	0.7	10.9	13.5	163	10.6***	3.0	26.0	0.6	9.5	11.7	163	0.6
Average pupil–teacher ratio in a school	59.5	23.7	110.0	2.2	55.2	63.8	163	72.6***	32.2	126.8	2.7	67.3	77.9	163	57.2	21.4	102.0	2.1	53.0	61.4	160	77.3***	30.9	135.1	3.9	69.6	84.9	160	-7*
Teaching and learning environment from observed classrooms																													
Average number of pupils in a classroom	38.9	11.0	75.0	1.6	35.7	42.1	245	45.4***	14.0	95.0	1.8	41.9	49.0	245	43.5	11.0	84.0	2.3	38.9	48.0	200	46.5	14.0	88.0	2.0	42.6	50.5	200	3.5
Multi-grade teaching (% observed lessons)	1.2			0.6	0.0	2.4	245	11.1***			2.1	7.0	15.3	245	6.9			1.9	3.1	10.6	200	10.3			2.4	5.4	15.1	200	6.5*

Co-teaching (% observed lessons)	8.8			2.1	4.7	12.9	245	2.4***			0.8	0.8	4.1	245	7.0			1.8	3.5	10.6	200	2.1***			0.4	1.3	3.0	200	-1.5
Availability of teaching and learning materials																												
Resources used during lesson (% observed lessons)																												
Teachers textbook	61.9			2.8	56.3	67.5	245	57.7			3.0	51.7	63.7	245	59.4			3.2	53.0	65.8	200	56.7			2.7	51.4	62.0	200	-1.5
Functioning blackboard/whiteboard	95.8			1.2	93.4	98.2	245	90.4**			2.0	86.5	94.3	245	95.0			1.7	91.5	98.4	200	93.2			1.5	90.2	96.2	200	-3.6
Chalk/marker	96.0			1.2	93.7	98.3	245	92.0*			1.6	88.8	95.2	245	98.5			0.7	97.2	99.8	200	92.8***			1.6	89.8	95.9	200	1.6
Poster, chart, or pictures	6.3			1.4	3.7	9.0	245	36.9***			2.7	31.6	42.2	245	7.8			1.8	4.2	11.3	200	13.5***			1.9	9.7	17.2	200	24.9***
Resources made by the teacher (e.g. flash card, handouts, etc.)	11.0			1.9	7.3	14.8	245	36.3***			2.8	30.7	41.9	245	16.9			2.5	12.0	21.8	200	28.2***			2.9	22.4	34.1	200	13.9***
Tools or objects from the local environment	7.0			1.4	4.2	9.7	245	22.0***			2.2	17.7	26.3	245	10.6			2.0	6.6	14.5	200	14.1			2.5	9.2	19.1	200	11.5***
Audio	0.0			0.0	0.0	0.0	245	0.8**			0.4	0.0	1.5	245	0.6			0.5	-0.4	1.6	200	0.3			0.2	-0.2	0.8	200	1.1
Video	0.0			0.0	0.0	0.0	245	0.0			0.0	0.0	0.0	245	0.0			0.0	0.0	0.0	200	0.0			0.0	0.0	0.0	200	0.0
Language of instruction																												
Speak Hausa at home (% pupils)								98.6			0.6	97.5	99.7	783								99.0			0.5	98.1	99.9	782	
Language(s) used during P1–P3 lessons excluding English lessons (% P1–P3 observed lessons exclu. Eng)																												
Hausa only								42.0			5.6	30.8	53.3	85								47.1			5.4	36.5	57.8	78	
English only								2.1			2.1	-2.0	6.3	85								2.6			2.0	-1.4	6.5	78	
Hausa and English								54.2			5.8	42.5	65.8	85								47.9			5.1	37.6	58.1	78	
Language(s) used during P4–P6 lessons excluding Hausa lessons (% P4–P6 observed lessons exclu. Hausa)																												
Hausa only								19.6			4.4	10.8	28.4	101								20.7			4.3	12.1	29.3	91	
English only								1.4			1.0	-0.5	3.4	101								1.0			0.6	-0.1	2.1	91	
Hausa and English								75.0			4.7	65.6	84.4	101								69.9			4.6	60.8	79.1	91	

Source: TDP impact evaluation baseline and endline survey, head teacher interview, classroom observations, and pupil test and background instrument.

Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator. (2) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1.

B.3 SLM

Annex Table 8: SLM

	Treatment														Control													
	Baseline							Endline							Baseline							Endline							
Indicator[1,2]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Raw DID
Background characteristics of head teachers																												
Average age (years)	44.2	34.0	52.0	0.4	43.4	45.1	164	44.3	34.0	53.0	0.4	43.4	45.1	164	45.4	33.0	54.0	0.4	44.5	46.3	164	45.9	35.0	55.0	0.5	45.0	46.8	164	-0.5
Average teaching experience (years)	19.8	9.0	31.0	0.5	18.8	20.8	164	18.9	11.0	29.0	0.4	18.1	19.7	164	21.4	10.0	33.0	0.5	20.5	22.4	163	21.8	11.0	32.0	0.5	20.9	22.7	163	-1.3
Average time in current school (years)	3.8	0.0	10.0	0.3	3.1	4.4	156	3.2	0.0	7.0	0.2	2.7	3.7	156	3.5	0.0	8.0	0.3	2.9	4.0	160	3.4	0.0	8.0	0.3	2.9	4.0	160	-0.5
Female (%)	3.6			1.2	1.2	5.9	164	2.9			1.1	0.8	5.0	164	2.3			1.0	0.4	4.2	164	4.0			1.5	1.0	7.0	164	-2.3
Have NCE qualification (%)	86.9			2.1	82.8	91.1	164	86.6			2.1	82.4	90.8	164	86.8			2.1	82.6	90.9	164	80.5**			2.5	75.5	85.5	164	5.9
Received teaching training in last two years (baseline) or last three years (endline) (%)	78.9			2.2	74.7	83.2	164	95.5***			1.4	92.8	98.3	164	79.3			2.2	74.9	83.7	164	78.9			2.2	74.5	83.4	164	17***
Received SLM training in last two years (baseline) or last three years (endline) (%)	22.5			2.6	17.4	27.5	164	93.5***			1.6	90.4	96.6	164	27.1			2.9	21.4	32.9	163	66.3***			2.5	61.3	71.3	163	31.9***
Received SLM training in last three years by provider (%)																													
TDP								90.0			1.8	86.4	93.6	164								19.5***			2.5	14.6	24.3	163	
UNICEF GEP3								4.7			1.3	2.2	7.2	164								15.3***			1.9	11.6	19.1	163	
ESSPIN								13.0			1.8	9.4	16.6	164								24.3***			2.0	20.3	28.3	163	

The Global Partnership for Education								27.8			2.1	23.8	31.8	164								28.0			2.0	24.0	31.9	163	
SUBEB								9.5			1.7	6.1	13.0	164								12.0			2.1	8.0	16.0	163	
UBEC								2.4			1.0	0.5	4.3	164								1.1			0.6	-0.1	2.4	163	
National Teachers Institute								0.0			0.0	0.0	0.0	164								0.6			0.5	-0.4	1.6	163	
Local government (LGEA)								2.3			0.9	0.6	4.0	164								2.9			1.0	0.8	4.9	163	
Other								1.2			0.7	-0.2	2.7	164								2.4			1.0	0.5	4.3	163	
Regularly teaches primary classes (%)								52.3			3.1	46.3	58.4	165																
Management of pupil and teacher attendance																														
Took action to improve pupil attendance last school year (% of head teachers)	98.8			0.7	97.4	100.2	158	93.6***			1.6	90.5	96.6	158	98.8			0.7	97.4	100.2	159	98.7			0.5	97.7	99.8	159	-5.2***
Actions taken to reduce pupil absence (% head teachers)																														
Involve SBMC in finding reasons for pupil non-attendance	75.4			3.0	69.3	81.4	147	83.7**			2.5	78.8	88.7	147	76.8			2.9	71.1	82.5	155	79.9			2.6	74.8	85.0	155	5.2
Discuss with teachers, pupils, or parents about reasons for pupil non-attendance	70.9			3.1	64.7	77.0	147	70.7			2.7	65.3	76.1	147	75.3			2.6	70.1	80.5	155	63.0***			3.0	57.1	68.8	155	12.2**
Provide financial support to pupils	3.2			1.1	0.9	5.4	147	13.0***			2.1	8.8	17.2	147	4.5			1.4	1.7	7.2	155	10.4***			2.0	6.5	14.4	155	3.8
Provide uniforms free of charge	6.7			1.6	3.4	9.9	147	12.9**			2.3	8.3	17.5	147	6.4			1.6	3.1	9.6	155	15.7***			2.5	10.9	20.6	155	-3.1
Provide textbooks, exercise books and stationery free of charge	13.0			2.3	8.3	17.6	147	23.8**			2.9	17.9	29.6	147	16.5			2.6	11.3	21.6	155	30.6***			2.8	25.0	36.1	155	-3.3
Address bullying	1.4			0.8	-0.2	3.0	147	0.6			0.5	-0.3	1.5	147	0.0			0.0	0.0	0.0	155	1.3*			0.8	-0.2	2.8	155	-2.1*
Address corporal punishment	2.6			1.1	0.5	4.7	147	0.7			0.6	-0.4	1.8	147	1.8			0.8	0.1	3.4	155	1.3			0.8	-0.2	2.8	155	-1.4
Improve quality of teaching/learning	6.0			1.5	2.9	9.0	147	6.5			1.6	3.4	9.7	147	7.0			1.7	3.6	10.5	155	6.3			1.8	2.7	9.9	155	1.3
Other actions	10.9			2.2	6.6	15.3	147	2.0***			0.9	0.2	3.8	147	12.2			2.2	7.9	16.5	155	3.6***			1.5	0.6	6.6	155	-0.4
Took action to improve teacher attendance last school year (% head teachers)	95.6			1.3	93.0	98.3	157	90.3**			2.0	86.5	94.2	157	94.3			1.6	91.2	97.4	156	93.9			1.4	91.1	96.7	156	-4.9
Actions taken to reduce teacher absence (% head teachers)																														
Ruling attendance book at opening time and following up on absences	46.8			3.7	39.4	54.2	136	50.3			3.3	43.7	56.9	136	48.4			3.6	41.2	55.5	139	45.5			3.3	38.9	52.1	139	6.3

Insist on written absence request	46.0			3.3	39.5	52.5	136	34.4***			3.0	28.5	40.3	136	38.4			2.8	32.8	44.0	139	30.4**			2.6	25.4	35.5	139	-3.6
Complete movement book during school hours	28.2			3.1	22.1	34.3	136	25.0			3.0	19.0	31.0	136	32.8			3.3	26.1	39.4	139	20.4***			2.6	15.2	25.6	139	9.2
Discuss with teachers about attendance	72.3			3.0	66.4	78.3	136	75.6			2.9	69.8	81.4	136	64.8			3.5	57.9	71.8	139	71.4			3.3	64.9	78.0	139	-3.4
Address pay-/salary-related grievances	11.2			2.1	7.1	15.4	136	11.1			2.3	6.5	15.6	136	8.2			1.9	4.4	12.0	139	6.6			1.6	3.4	9.8	139	1.4
Address childcare/maternity/paternity issues	3.6			1.3	1.0	6.1	136	2.1			1.0	0.2	4.1	136	6.9			2.0	2.8	10.9	139	4.7			1.3	2.0	7.3	139	0.8
Address issues related to school infrastructure/conditions	0.0			0.0	0.0	0.0	136	1.4*			0.8	-0.2	3.1	136	2.1			1.0	0.2	4.0	139	1.4			0.8	-0.2	2.9	139	2.2
Address lack of teaching materials	1.5			0.9	-0.2	3.2	136	3.2			1.3	0.5	5.9	136	2.0			1.0	0.2	3.9	139	5.5*			1.6	2.4	8.6	139	-1.8
Other actions	5.4			1.6	2.3	8.6	136	3.5			1.3	1.0	6.0	136	11.9			2.1	7.7	16.1	139	4.9***			1.5	1.9	7.7	139	5.2
Head teacher's support for teachers																													
Conducted lesson observations in last 10 days (% of head teachers)	78.8			2.6	73.5	84.0	165	72.3*			2.5	67.3	77.2	165	80.5			2.7	75.2	85.7	164	70.0***			3.0	64.2	75.9	164	3.9
Hold one or more formal head teacher–teacher meetings per month (% head teachers)	69.8			2.7	64.5	75.1	165	57.7***			3.0	51.7	63.7	165	65.3			3.1	59.3	71.4	164	49.5***			3.0	43.6	55.5	164	3.7
Hold one or more formal head teacher–teacher meetings per week (% of head teachers)	32.2			3.0	26.4	38.1	165	23.6**			2.4	18.8	28.3	165	27.3			2.8	21.8	32.7	164	27.5			2.9	21.8	33.2	164	-8.9*
Topics discussed during formal head teacher–teacher meetings (% head teacher reporting holding these meetings)																													
Teacher absenteeism/lateness	73.3			2.9	67.6	79.1	161	72.5			2.8	67.0	78.1	161	67.2			3.2	60.8	73.6	156	71.6			2.7	66.2	77.0	156	-5.2
Pupil attendance	63.6			3.0	57.6	69.6	161	62.8			2.7	57.4	68.2	161	61.6			3.7	54.3	68.9	156	68.0			2.9	62.2	73.9	156	-7.3
Teachers' pay/salary	3.1			1.1	1.0	5.2	161	1.9			0.9	0.1	3.7	161	1.9			0.9	0.1	3.7	156	1.9			0.9	0.1	3.6	156	-1.2
Lack of teaching/learning materials	6.6			1.6	3.5	9.7	161	30.2***			2.5	25.3	35.2	161	13.1			2.3	8.5	17.7	156	31.0***			2.8	25.4	36.6	156	5.7
School building conditions/repairs	4.4			1.2	2.1	6.7	161	9.2**			1.7	5.9	12.6	161	6.2			1.6	3.1	9.3	156	10.8**			2.0	7.0	14.7	156	0.2
Teaching practice/pedagogy	42.7			3.2	36.4	49.1	161	43.4			3.1	37.3	49.4	161	47.3			3.5	40.3	54.3	156	37.2**			3.3	30.7	43.7	156	10.7*
Individual students' needs	12.5			2.1	8.2	16.7	161	26.4***			2.8	20.9	31.9	161	11.5			2.3	7.0	16.1	156	26.3***			3.2	20.0	32.7	156	-0.9
Parents/community	13.8			2.3	9.2	18.5	161	23.8***			2.5	18.9	28.6	161	17.9			2.7	12.4	23.3	156	22.5			2.4	17.8	27.2	156	5.3

Training	3.2			1.1	1.0	5.3	161	5.8			1.5	2.7	8.8	161	2.6			1.4	-0.2	5.3	156	8.7***			1.8	5.1	12.3	156	-3.6
Professional development	25.1			2.7	19.8	30.3	161	13.2***			2.3	8.7	17.7	161	28.7			3.1	22.5	35.0	156	7.0***			1.7	3.6	10.3	156	9.9**
Other	10.4			2.1	6.4	14.5	161	6.5			1.6	3.4	9.6	161	6.6			1.6	3.4	9.8	156	3.1*			1.0	1.0	5.1	156	-0.3
Head teacher turnover																													
No longer at school at endline (% baseline head teachers)								58.8			3.0	52.8	64.7	164								59.4			3.1	53.3	65.6	164	
Will retire by endline (% baseline head teachers)								2.4			1.0	0.5	4.3	165								6.6**			1.8	3.1	10.2	164	
Head teacher incentives and motivation																													
Received salary on time last school year (% head teachers)																													
Always on time								70.4			2.1	66.2	74.6	165								71.7			2.2	67.5	76.0	164	
Usually/mostly on time								20.3			2.2	15.9	24.6	165								17.1			2.2	12.8	21.3	164	
Usually/mostly delayed								8.7			1.6	5.5	12.0	165								10.0			1.7	6.8	13.3	164	
Always delayed								0.6			0.5	-0.4	1.6	165								1.2			0.7	-0.2	2.5	164	
Received correct salary amount for last three payments (% head teachers)																													
All three payment amounts correct								87.2			1.8	83.6	90.8	165								85.5			1.7	82.1	89.0	164	
Some payment amounts not correct								12.8			1.8	9.2	16.4	165								13.3			1.7	9.9	16.7	164	
No payment amount correct								0.0			0.0	0.0	0.0	165								1.2*			0.7	-0.2	2.5	164	
Considers workload (% head teachers)																													
Appropriate								56.3			3.1	50.3	62.4	165								61.2			3.1	55.1	67.3	163	
Excessive								43.7			3.1	37.6	49.7	165								38.8			3.1	32.8	44.9	163	
Reasons for excessive workload (% of head teachers reporting excessive workload)[3]																													
Not enough teachers at the school								57.7			5.7	46.3	69.2	71								69.1			4.9	59.3	79.0	63	
Teach too many classes								40.0			5.4	29.2	50.8	71								38.1			4.6	28.9	47.3	63	
Too many administrative and clerical duties								55.8			5.4	45.0	66.7	71								41.8*			5.5	30.9	52.8	63	
There are many meetings with the LGEA/SUBEB								8.0			2.9	2.3	13.8	71								6.1			2.8	0.6	11.6	63	
I have to cover classes for absent teachers								17.7			4.6	8.5	26.9	71								15.6			2.8	10.1	21.1	63	

Other								1.4			1.5	-1.6	4.5	71								6.2			3.3	-0.4	12.8	63	
Absent one day or more in last five days, self-reported (% head teachers)	11.8			2.2	7.5	16.1	164	24.4***			2.6	19.2	29.6	165	16.1			2.5	11.3	21.0	164	16.4			2.3	11.8	21.0	163	12.4***
Reasons for absence last five days (% of head teachers reporting absent)																													
Elections/campaigning	0.0			0.0	0.0	0.0	19	0.0			0.0	0.0	0.0	40	0.0			0.0	0.0	0.0	26	0.0			0.0	0.0	0.0	28	0.0
Transport	0.0			0.0	0.0	0.0	19	2.5			3.2	-4.1	9.2	40	7.6			6.4	-6.0	21.2	26	7.5			4.6	-2.2	17.2	28	2.6
Teacher strikes	0.0			0.0	0.0	0.0	19	0.0			0.0	0.0	0.0	40	0.0			0.0	0.0	0.0	26	0.0			0.0	0.0	0.0	28	0.0
Other mass strikes	0.0			0.0	0.0	0.0	19	0.0			0.0	0.0	0.0	40	0.0			0.0	0.0	0.0	26	0.0			0.0	0.0	0.0	28	0.0
Own or family illness	46.1			11.5	22.1	70.0	19	27.2			7.4	11.8	42.6	40	44.8			7.1	29.8	59.8	26	25.9			7.2	10.8	41.1	28	0.0
Late or non-payment of salary	11.8			8.2	-5.2	28.9	19	0.0			0.0	0.0	0.0	40	7.6			6.2	-5.5	20.6	26	0.0			0.0	0.0	0.0	28	-4.3
Training	10.5			6.3	-2.7	23.7	19	39.9***			5.5	28.5	51.3	40	10.7			4.9	0.5	21.0	26	20.9			5.5	9.4	32.5	28	19.1*
Meeting or event at LGA / SUBEB	10.5			6.3	-2.7	23.7	19	18.1			7.3	2.9	33.4	40	0.0			0.0	0.0	0.0	26	10.3			7.0	-4.4	25.0	28	-2.7
Social or religious obligations	0.0			0.0	0.0	0.0	19	7.4**			3.2	0.6	14.2	40	7.6			6.4	-6.0	21.2	26	10.6			4.0	2.3	19.0	28	4.3
Epidemic/disease outbreak	0.0			0.0	0.0	0.0	19	0.0			0.0	0.0	0.0	40	0.0			0.0	0.0	0.0	26	0.0			0.0	0.0	0.0	28	0.0
Weather-related reasons	0.0			0.0	0.0	0.0	19	0.0			0.0	0.0	0.0	40	0.0			0.0	0.0	0.0	26	0.0			0.0	0.0	0.0	28	0.0
Other	21.1			10.2	-0.2	42.3	19	7.4			5.4	-3.8	18.6	40	21.8			6.4	8.3	35.2	26	24.4			6.9	9.8	38.9	28	-16.3
Absent one day or more in last term, self-reported (% head teachers)	47.7			3.0	41.8	53.5	163	65.0***			2.9	59.4	70.7	164	53.3			3.4	46.5	60.1	163	62.3**			3.2	56.0	68.7	163	8.3
Reasons for absence last term (% of head teachers reporting absent)																													
Elections/campaigning	0.0			0.0	0.0	0.0	76	0.0			0.0	0.0	0.0	106	0.0			0.0	0.0	0.0	87	0.0			0.0	0.0	0.0	101	0.0
Transport	1.2			1.0	-0.8	3.2	76	3.7			1.5	0.7	6.7	106	6.5			2.2	2.2	10.9	87	1.9*			1.1	-0.3	4.1	101	7.1**
Teacher strikes	0.0			0.0	0.0	0.0	76	0.0			0.0	0.0	0.0	106	0.0			0.0	0.0	0.0	87	0.0			0.0	0.0	0.0	101	0.0
Other mass strikes	0.0			0.0	0.0	0.0	76	0.0			0.0	0.0	0.0	106	0.0			0.0	0.0	0.0	87	0.0			0.0	0.0	0.0	101	0.0
Own or family illness	63.8			4.8	54.2	73.4	76	40.0***			3.7	32.7	47.3	106	48.8			4.6	39.7	57.9	87	58.9			4.2	50.6	67.1	101	-33.8***

Late or non-payment of salary	3.9			1.5	0.9	7.0	76	1.0*			0.8	-0.6	2.6	106	2.3			1.4	-0.4	5.0	87	0*			0.0	0.0	0.0	101	-0.7
Training	16.3			3.8	8.7	23.9	76	65.2***			3.9	57.6	72.9	106	17.0			3.4	10.3	23.7	87	23.6			3.4	16.9	30.3	101	42.3***
Meeting or event at LGA / SUBEB	11.5			3.0	5.6	17.4	76	14.8			2.7	9.5	20.2	106	21.3			3.7	14.0	28.7	87	11.8**			3.0	5.8	17.7	101	12.9**
Social or religious obligations	7.4			2.3	2.9	12.0	76	6.5			2.0	2.7	10.4	106	9.8			2.7	4.5	15.1	87	14.4			2.8	8.8	20.0	101	-5.5
Epidemic/disease outbreak	0.0			0.0	0.0	0.0	76	0.0			0.0	0.0	0.0	106	0.0			0.0	0.0	0.0	87	0.0			0.0	0.0	0.0	101	0.0
Weather-related reasons	1.2			1.0	-0.7	3.2	76	0.0			0.0	0.0	0.0	106	0.0			0.0	0.0	0.0	87	2.9**			1.4	0.1	5.7	101	-4.1**
Other	9.0			2.7	3.7	14.2	76	2.9**			1.4	0.1	5.6	106	11.5			2.9	5.7	17.3	87	3.0**			1.4	0.1	5.8	101	2.4
SUBEB and LGEA support to schools																												
Number of times SUBEB or LGEA supervisor visited school last year (% schools)																												
Never	0.0			0.0	0.0	0.0	156	2.6**			1.0	0.5	4.6	156	0.6			0.5	-0.4	1.7	158	3.7**			1.2	1.2	6.2	158	-0.5
At least once a year but not more than once a month	13.8			2.2	9.5	18.2	156	40.1***			3.1	33.9	46.3	156	15.0			2.5	10.1	19.9	158	36.0***			3.4	29.3	42.7	158	5.2
Two or three times a month	55.1			3.2	48.7	61.4	156	33.4***			2.9	27.6	39.2	156	59.5			3.4	52.9	66.2	158	37.6***			3.3	31.1	44.1	158	0.3
More than three times a month	31.1			3.1	25.0	37.2	156	24.0*			2.7	18.7	29.2	156	24.9			2.7	19.6	30.2	158	22.7			2.9	16.9	28.5	158	-5.0
Source: TDP impact evaluation baseline and endline surveys, head teacher interview, and baseline-endline survey records.																												
Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator. (2) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1. (3) Head teachers reported the two main reasons why their workload is excessive.																												

B.4 Teachers

Annex Table 9: Teachers

	Treatment		Control		
	Baseline	Endline	Baseline	Endline	

Indicator[1,2,3]	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Estimate	P10	P90	SE	Lower 95CI	Upper 95CI	N	Raw DID
Teacher pedagogy in the classroom (among panel P1–P6 teachers and head teachers who were observed)																													
Teacher involves pupils in positive interaction (mean % of lesson time)	24.3	9.7	40.7	0.8	22.8	25.9	251	29.1***	13.0	45.6	0.8	27.6	30.7	251	23.3	6.9	40.5	0.9	21.4	25.1	204	25.9**	10.6	40.5	0.8	24.4	27.4	204	2.2
Proportion of lesson time spent on each type of teacher talk (mean % of lesson time)																													
Instructs/presents/ dictates	18.0	0.0	50.0	1.3	15.3	20.6	251	22.6***	0.0	50.0	1.1	20.4	24.9	251	18.0	0.0	42.9	1.3	15.4	20.7	204	20.6	0.0	40.0	1.0	18.6	22.7	204	2.1
Chants	11.6	0.0	40.0	1.0	9.6	13.6	251	9.0**	0.0	26.7	0.8	7.4	10.6	251	9.1	0.0	30.0	0.9	7.3	10.9	204	11.1	0.0	33.3	0.9	9.2	13.0	204	-4.6***
Explains	29.4	0.0	66.7	1.5	26.4	32.3	251	24.6**	0.0	50.0	1.0	22.6	26.5	251	26.7	0.0	58.3	1.6	23.5	30.0	204	26.9	0.0	50.0	1.2	24.4	29.3	204	-5.0*
Closed question/response	10.1	0.0	25.0	0.7	8.8	11.5	251	10.6	0.0	27.3	0.7	9.3	11.9	251	13.3	0.0	33.3	1.0	11.3	15.4	204	10.3**	0.0	25.0	0.8	8.6	11.9	204	3.6**
Open question/response	4.3	0.0	16.7	0.5	3.3	5.3	251	5.1	0.0	20.0	0.5	4.1	6.1	251	3.8	0.0	16.7	0.5	2.9	4.7	204	4.6	0.0	15.4	0.5	3.6	5.5	204	0.1
Assists/group discussion	6.7	0.0	25.0	0.7	5.4	8.1	251	9.0**	0.0	33.3	0.8	7.4	10.6	251	7.7	0.0	28.6	1.0	5.7	9.7	204	6.3	0.0	20.0	0.7	4.9	7.7	204	3.7**
None of the above	19.9	0.0	50.0	1.3	17.3	22.5	251	19.1	0.0	46.7	1.0	17.2	21.1	251	21.3	0.0	50.0	1.4	18.6	24.1	204	20.3	0.0	50.0	1.2	17.9	22.7	204	0.2
Proportion of lesson time spent on each teacher activity (mean % of lesson time)																													
Writes on/reads from blackboard	42.6	10.0	75.0	1.4	39.9	45.3	251	26.6***	0.0	58.3	1.3	24.1	29.1	251	41.0	9.1	66.7	1.5	37.9	44.0	204	33.1***	6.7	66.7	1.7	29.7	36.4	204	-8.1***
Demonstrates/displays work	7.7	0.0	25.0	0.8	6.1	9.2	251	10.9**	0.0	30.0	1.0	9.0	12.9	251	8.0	0.0	28.6	1.0	6.0	10.1	204	10.1	0.0	33.3	0.9	8.4	11.9	204	1.2
Moves around among pupils	20.1	0.0	50.0	1.2	17.8	22.5	251	34.6***	0.0	66.7	1.3	32.1	37.2	251	20.3	0.0	50.0	1.3	17.6	22.9	204	32.2***	0.0	64.3	1.4	29.4	34.9	204	2.6
Uses textbook	6.1	0.0	25.0	0.8	4.4	7.7	251	3.1***	0.0	11.1	0.4	2.3	3.8	251	4.0	0.0	16.7	0.6	2.8	5.2	204	3.7	0.0	13.3	0.5	2.6	4.7	204	-2.6**
Uses materials (printed/improvised)	3.0	0.0	9.1	0.5	2.0	4.0	251	10.8***	0.0	33.3	0.9	9.1	12.6	251	3.0	0.0	11.1	0.5	2.0	3.9	204	6.9***	0.0	25.0	0.9	5.2	8.6	204	3.9***
None of the above	20.5	0.0	50.0	1.1	18.3	22.6	251	14.0***	0.0	41.7	1.0	12.0	16.0	251	23.7	0.0	55.6	1.4	20.9	26.5	204	14.1***	0.0	36.4	1.2	11.7	16.5	204	3.1
Proportion of lesson time spent on each pupil activity (mean % of lesson time)																													
Group discussion/presentation	1.2	0.0	0.0	0.3	0.5	1.8	251	8.1***	0.0	33.3	0.9	6.4	9.9	251	1.3	0.0	0.0	0.3	0.6	1.9	204	3.2***	0.0	12.5	0.6	2.0	4.3	204	5.0***

Group or pair work	4.5	0.0	16.7	0.9	2.7	6.3	251	8.2***	0.0	26.7	0.8	6.6	9.8	251	5.3	0.0	20.0	0.8	3.6	6.9	204	3.8	0.0	14.3	0.6	2.7	5.0	204	5.1***
Respond to open question	7.5	0.0	25.0	0.9	5.6	9.4	251	5.4*	0.0	20.0	0.5	4.4	6.5	251	6.6	0.0	25.0	0.8	5.0	8.2	204	4.9*	0.0	18.2	0.6	3.7	6.1	204	-0.3
Respond to closed question	14.6	0.0	40.0	1.1	12.4	16.8	251	9.4***	0.0	25.0	0.6	8.1	10.6	251	16.6	0.0	50.0	1.6	13.4	19.9	204	8.7***	0.0	27.3	0.9	7.0	10.5	204	2.6
Individual work	21.7	0.0	66.7	1.6	18.4	24.9	251	7.4***	0.0	26.7	0.8	5.9	8.9	251	21.6	0.0	60.0	1.9	17.8	25.5	204	11.7***	0.0	40.0	1.0	9.7	13.6	204	-4.3
None of the above	49.9	0.0	91.7	2.0	45.9	53.9	251	61.5***	33.3	88.9	1.3	59.0	64.0	251	47.8	10.0	90.0	2.1	43.6	52.1	204	67.7***	40.0	100.0	1.5	64.8	70.6	204	-8.3**
At the end of the lesson teacher (% of lessons observed until they ended)																													
Summarised day's lesson	46.0			4.2	37.7	54.3	117	48.2			4.3	39.6	56.7	117	52.0			4.3	43.3	60.8	109	41.8			5.0	31.8	51.8	109	12.4
Revisited lesson's objectives	18.7			3.7	11.2	26.2	117	16.0			3.4	9.1	22.9	117	30.7			4.0	22.6	38.8	109	29.7			4.7	20.2	39.1	109	-1.7
Gave pupils homework	19.7			3.3	13.0	26.4	117	27.7*			3.8	20.0	35.3	117	28.5			4.0	20.4	36.6	109	20.5			3.7	13.2	27.9	109	16**
Teacher used praise more than reprimands (% of lessons)	76.6			2.6	71.4	81.7	251	91.3***			1.4	88.5	94.0	251	82.7			2.6	77.6	87.8	204	92.2***			1.5	89.3	95.2	204	5.2
Use of teaching aids and materials in the classroom																													
Resources used during lesson (% of lessons)																													
Teacher's textbook	61.9			2.8	56.3	67.5	245	57.7			3.0	51.7	63.7	245	59.4			3.2	53.0	65.8	200	56.7			2.7	51.4	62.0	200	-1.5
Functioning blackboard/whiteboard	95.8			1.2	93.4	98.2	245	90.4**			2.0	86.5	94.3	245	95.0			1.7	91.5	98.4	200	93.2			1.5	90.2	96.2	200	-3.6
Chalk/marker	96.0			1.2	93.7	98.3	245	92.0*			1.6	88.8	95.2	245	98.5			0.7	97.2	99.8	200	92.8***			1.6	89.8	95.9	200	1.6
Poster, chart, or pictures	6.3			1.4	3.7	9.0	245	36.9***			2.7	31.6	42.2	245	7.8			1.8	4.2	11.3	200	13.4***			1.9	9.7	17.2	200	24.9***
Resources made by the teacher (e.g. flash card, handouts, etc.)	11.0			1.9	7.3	14.8	245	36.3***			2.8	30.7	41.9	245	16.9			2.5	12.0	21.8	200	28.2***			2.9	22.4	34.1	200	13.9***
Tools or objects from the local environment	7.0			1.4	4.2	9.7	245	22.0***			2.2	17.7	26.3	245	10.6			2.0	6.6	14.5	200	14.1			2.5	9.2	19.1	200	11.5***
Audio	0.0			0.0	0.0	0.0	245	0.8**			0.4	0.0	1.5	245	0.6			0.5	-0.4	1.6	200	0.3			0.2	-0.2	0.8	200	1.1
Video	0.0			0.0	0.0	0.0	245	0.0			0.0	0.0	0.0	245	0.0			0.0	0.0	0.0	200	0.0			0.0	0.0	0.0	200	0.0
TDP materials available in classroom (% of lessons)																													
Teachers' guide for subject taught								18.4			2.1	14.2	22.5	208															
Teachers' guide in pedagogy								7.7			1.1	5.6	9.9	246															
Lesson plan for the subject taught								22.0			2.6	16.9	27.2	208															

Flash cards (any subject)								25.0			2.4	20.3	29.7	247															
TDP English poster								12.3			1.9	8.6	16.0	247															
TDP maths poster								12.4			2.0	8.4	16.5	247															
TDP science poster								6.9			1.6	3.7	10.1	247															
TDP reading benchmark/I can read poster								2.9			0.8	1.3	4.5	247															
TDP materials used in classroom (% of lessons where TDP materials available)																													
Teachers' guide for subject taught								75.2			5.9	62.2	88.1	42															
Teachers' guide in pedagogy								77.4			12.3	49.7	105.2	21															
Lesson plan for the subject taught								85.7			3.1	78.9	92.4	50															
Flash cards (any subject)								93.1			3.0	87.0	99.2	69															
TDP poster for subject taught								73.0			7.0	57.8	88.2	45															
Lesson durations[4]																													
P1–P3 lesson duration (% of P1–P3 lessons of each length)																													
10 to 15 minutes								7.7			2.2	3.3	12.1	130								10.1			2.7	4.7	15.5	100	
16 to 30 minutes								49.2			4.0	41.4	57.1	130								35.7**			4.5	26.8	44.6	100	
31 to 45 minutes								20.7			3.6	13.5	27.9	130								33.6**			4.6	24.5	42.8	100	
More than 45 minutes								22.4			3.5	15.4	29.3	130								20.6			4.0	12.7	28.5	100	
P4–P6 lesson duration (% of P4–P6 lessons of each length)																													
9 to 21 minutes								23.0			3.7	15.6	30.4	101								23.2			4.1	15.1	31.3	96	
22 to 39 minutes								51.3			4.9	41.7	61.0	101								52.3			4.3	43.7	60.9	96	
40 to 45 minutes								3.1			2.0	-1.0	7.1	101								4.1			2.6	-1.0	9.3	96	
More than 45 minutes								22.6			4.6	13.4	31.8	101								20.4			4.1	12.2	28.6	96	
Teacher absenteeism from school, from school records																													
Average daily teacher absence from school (mean % of teachers)	12.3	0.0	26.7	0.8	10.8	13.9	165	26.5***	0.0	56.0	1.4	23.7	29.3	165	14.8	0.0	35.0	0.8	13.3	16.3	163	30.0***	0.0	60.0	1.5	27.1	32.9	163	-1.0
Teacher absenteeism from school, self-reported																													
Absent one day or more in last two days, self-reported (% teachers)	24.0			2.4	19.2	28.8	246	37.1***			2.8	31.5	42.7	246	24.4			2.6	19.2	29.6	191	34.7***			3.4	28.0	41.5	191	2.7

Reasons for absence last five days (% of teachers reporting absent)																													
Elections/campaigning								0.0			0.0	0.0	0.0	94								0.0			0.0	0.0	0.0	70	
Transport								1.0			0.6	-0.2	2.2	94								3.3			1.7	-0.2	6.7	70	
Teacher strikes								0.0			0.0	0.0	0.0	94								1.0			1.1	-1.1	3.2	70	
Other mass strikes								0.5			0.5	-0.6	1.6	94								0.0			0.0	0.0	0.0	70	
Own or family illness								56.6			4.4	47.8	65.4	94								56.2			5.3	45.7	66.8	70	
Late or non-payment of salary								0.0			0.0	0.0	0.0	94								0.9			0.9	-1.0	2.7	70	
Training								6.5			2.6	1.3	11.6	94								12.8**			1.6	9.6	16.0	70	
Meeting or event at LGA/SUBEB								3.4			1.5	0.4	6.4	94								1.9			1.4	-1.0	4.7	70	
Social or religious obligations								13.7			3.9	5.9	21.5	94								12.7			4.6	3.5	21.9	70	
Epidemic/disease outbreak								0.0			0.0	0.0	0.0	94								0.0			0.0	0.0	0.0	70	
Weather-related reasons								0.0			0.0	0.0	0.0	94								0.0			0.0	0.0	0.0	70	
Collect salary								5.2			2.6	-0.1	10.4	94								5.7			1.8	2.2	9.2	70	
Maternity leave								1.2			1.0	-0.7	3.2	94								0.0			0.0	0.0	0.0	70	
Other income-generating activity								3.9			2.0	0.0	7.9	94								4.9			2.7	-0.4	10.3	70	
Study leave/exam								6.3			3.5	-0.7	13.2	94								2.4			0.6	1.2	3.5	70	
Other								3.3			2.8	-2.4	8.9	94								2.7			2.3	-1.8	7.2	70	
Absent one day or more in last term, self-reported (% teachers)	52.9			2.6	47.8	58.1	244	69.2***			2.5	64.3	74.2	244	52.8			3.2	46.5	59.1	190	70.7***			3.0	64.7	76.7	190	-1.6
Reasons for absence last term (% of head teachers reporting absent)																													
Elections/campaigning								0.0			0.0	0.0	0.0	165								0.0			0.0	0.0	0.0	139	
Transport								7.3			2.1	3.1	11.5	165								7.7			1.4	5.0	10.4	139	
Teacher strikes								0.0			0.0	0.0	0.0	165								0.0			0.0	0.0	0.0	139	

Other mass strikes								0.0			0.0	0.0	0.0	165								0.0			0.0	0.0	0.0	139	
Own or family illness								73.8			3.2	67.4	80.2	165								75.3			3.5	68.4	82.1	139	
Late or non-payment of salary								0.0			0.0	0.0	0.0	165								1.0			0.6	-0.2	2.1	139	
Training								16.0			2.5	11.1	21.0	165								10.3*			2.1	6.2	14.4	139	
Meeting or event at LGA/SUBEB								2.8			1.4	0.0	5.6	165								1.8			1.6	-1.3	4.9	139	
Social or religious obligations								12.3			2.8	6.7	17.9	165								12.4			2.3	7.9	17.0	139	
Epidemic/disease outbreak								0.0			0.0	0.0	0.0	165								0.0			0.0	0.0	0.0	139	
Weather-related reasons								4.0			1.2	1.7	6.3	165								0.6***			0.5	-0.4	1.6	139	
Collect salary								5.6			2.0	1.8	9.5	165								8.6			1.8	5.0	12.2	139	
Maternity leave								1.6			1.2	-0.7	3.9	165								0.0			0.0	0.0	0.0	139	
Other income-generating activity								5.3			2.2	1.0	9.6	165								2.3			1.1	0.0	4.5	139	
Study leave/exam								1.9			1.0	-0.1	3.9	165								3.9			1.8	0.5	7.4	139	
Other								2.5			1.0	0.5	4.6	165								2.7			1.1	0.6	4.8	139	
Teacher absenteeism from classroom, from classroom attendance observation																															
Teacher and pupil presence in P1–P6 classrooms after roll-call (mean % of classrooms)																															
Teacher and pupils present								36.8	0.0	83.3	1.8	33.3	40.3	166								36.6	0.0	83.3	2.2	32.4	40.9	163	
Teacher absent, pupils present								53.6	0.0	100.0	2.0	49.7	57.5	166								52.4	0.0	100.0	2.3	47.9	56.9	163	
Teacher present, pupils absent								1.0	0.0	0.0	0.5	-0.1	2.0	166								0.6	0.0	0.0	0.6	-0.5	1.7	163	
Teacher and pupils absent								8.6	0.0	33.3	1.6	5.4	11.8	166								10.4	0.0	50.0	1.6	7.2	13.6	163	
Teacher and pupil presence in P1–P6 classrooms after long break (mean % of classrooms)																															
Teacher and pupils present								48.3	0.0	100.0	2.1	44.2	52.4	165								48.3	0.0	100.0	1.9	44.7	52.0	162	
Teacher absent, pupils present								47.1	0.0	100.0	2.0	43.1	51.1	165								47.7	0.0	84.5	1.8	44.1	51.3	162	

Teacher present, pupils absent								0.8	0.0	0.0	0.5	-0.2	1.9	165								0.3	0.0	0.0	0.1	0.0	0.6	162	
Teacher and pupils absent								3.8	0.0	8.3	0.9	2.0	5.6	165								3.7	0.0	0.0	0.9	2.0	5.4	162	
Teacher incentives and motivation																													
Received salary on time last school year (% of teachers)																													
Always on time								72.4			2.2	68.1	76.6	249								78.2*			2.4	73.4	82.9	197	
Usually/ mostly on time								14.7			1.7	11.4	18.0	249								11.9			1.8	8.4	15.5	197	
Usually/mostly delayed								12.3			1.9	8.7	16.0	249								9.3			1.9	5.7	13.0	197	
Always delayed								0.6			0.5	-0.4	1.6	249								0.6			0.5	-0.4	1.5	197	
Received correct salary amount for last three payments (% of teachers)																													
All three payment amounts correct								87.1			1.6	83.9	90.2	249								89.9			2.0	85.9	93.9	197	
Some payment amounts not correct								10.5			1.6	7.4	13.5	249								8.3			2.1	4.3	12.3	197	
No payment amount correct								2.5			0.8	0.8	4.1	249								1.8			0.8	0.3	3.3	197	
Considers workload (% of teachers)																													
Appropriate								68.0			2.7	62.7	73.3	248								63.4			3.3	56.8	70.0	197	
Excessive								32.0			2.7	26.7	37.3	248								36.6			3.3	30.0	43.2	197	
Reasons for excessive workload (% of teachers reporting excessive workload)[5]																													
Not enough teachers at the school								72.5			5.7	61.0	83.9	76								69.5			5.8	57.8	81.2	67	
Too many pupils in my classes								24.3			5.7	12.9	35.8	76								24.0			5.1	13.7	34.3	67	
Teach too many classes								63.2			6.8	49.6	76.8	76								65.8			6.0	53.7	78.0	67	
Too many administrative and clerical duties								9.1			3.2	2.6	15.6	76								15.9			4.6	6.7	25.1	67	
Have to cover classes for absent teachers								6.8			3.1	0.5	13.1	76								0**			0.0	0.0	0.0	67	
Lack of planning by head teacher								0.0			0.0	0.0	0.0	76								0.0			0.0	0.0	0.0	67	
Other								5.7			3.9	-2.0	13.5	76								4.8			4.1	-3.5	13.2	67	
Source: TDP impact evaluation baseline and endline surveys, classroom observation, head teacher interview (school records checks), teacher interview, and classroom attendance.																													

Notes: (1) Blank cells mean that the estimate is not applicable to this type of indicator. (2) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1. (3) All estimates from classroom observations exclude lessons shorter than nine minutes. (4) Lessons where the observation ended at minute 36 but the lesson was still ongoing are excluded, this affects 29 lessons. These were lessons at the very start of fieldwork when lesson observations were limited to 36 minutes. (5) Teachers reported the two main reasons why their workload is excessive.

Annex C EQUALS matrix

The table below describes how the evaluation report fulfils each of the EQUALS evaluation criteria, and identifies where each point is addressed in the report's two volumes.

Section	#	Question	Where is this addressed?	Comments for DFID
I. STRUCTURE AND CLARITY	1	The product is accessible to the intended audience (e.g. free of jargon, written in plain English, logical use of chapters, appropriate use of tables, graphs, and diagrams).	Volumes I and II.	The endline report is split into two volumes. Volume I is set up to be accessible to a non-technical audience and presents the impact evaluation findings at output, outcome, and impact levels, while Volume II covers all the technical aspects of the impact evaluation.
	2	It is clear who has carried out the evaluation.	Chapter 1, Volume I.	
	3	An executive summary is included, and it can stand alone as an accurate summary of the main product.	Volume I.	
	4	The annexes contain – at the least – the original terms of reference (TOR), the evaluation framework (including evaluation questions), a bibliography, and a list of consultees.	Annex A, Volume II.	There were no TOR as such for this evaluation. The original design of the impact evaluation was set out in the 2014 Impact Evaluation Framework. This was later amended as described in the 2017 Impact Evaluation Endline Plan, which constitutes the final intended design of this evaluation. The Endline Plan was agreed with DFID, EQUALS, and TDP. The Impact Evaluation Matrix is presented in Volume II, Annex A.
	5	Annexes increase the usefulness of the product.	Volumes I and II.	
	6	Any departures from the original TOR have been adequately explained and justified.	Chapter 1, Volume I.	Please note that the 2017 Impact Evaluation Endline Plan, constitutes the final intended design of this evaluation and is the key reference document for this evaluation.
	7	The product is of publishable quality.		
II. CONTEXT	1	The product provides a relevant and sufficient description of the intervention to be evaluated. At the least, this should include detail on the intervention's anticipated impact, outcomes and outputs, target groups, timescale, geographical coverage, and the extent to which the intervention aimed to address issues of equity, poverty, and exclusion.	Chapter 1 and Chapter 3, Volume I.	TDP does not explicitly aim to address issues of equity, poverty, and exclusion. The report does not touch on these issues further, except to disaggregate learning outcomes by pupil wealth (Chapter 7, Volume I).

	2	The product describes the intervention logic and/or TOC.	Chapter 1 and Annex A, Volume I. Annex A, Volume II.	During the development of the evaluation framework prior to the evaluation baseline, the evaluation team developed a detailed, expanded programme TOC; this is described in detail in the 2014 Impact Evaluation Framework. The programme TOC and the evaluation matrix, with the research questions covered by this evaluation, are included as annexes.
	3	The product provides a relevant and sufficient description of the local, national, and/or international development context within which the intervention was operating.	Chapter 4, Volume I.	
	4	The product identifies key linkages between the evaluated intervention and other relevant projects / programmes / donors. If no linkages are identified, the product justifies why other projects / programmes / donors were not relevant to the evaluation.	Section 2.1.3, Volume I, and Section 3.1, Volume II.	
	5	There is an assessment of the policy context for the intervention and this includes reference to poverty reduction strategies, gender equality, environmental protection, and human rights.	Chapter 4, Volume I.	Poverty reduction strategies, human rights, and environmental protection are not relevant for this evaluation and are not referenced.
	6	The product describes the extent to which the intervention has been managed and delivered against the Paris Declaration principles.		This was not part of the evaluation framework or endline plan for this evaluation.
III. PURPOSE, SCOPE, AND OBJECTIVES	1	The product describes what information is needed through the evaluation, and how that information will be used.	Chapter 1, Volume I.	
	2	The product describes whether the evaluation is for accountability and/or learning purposes.	Chapter 1, Volume I.	This evaluation has formative, summative and learning purposes. Volume I provides recommendations for TDP and general lessons based on the findings from the evaluation.
	3	The product describes the target audience(s) for the evaluation.	Chapter 1, Volume I.	
	4	The product justifies the timing of the evaluation.	Chapter 1, Volume I.	
	5	The product clearly outlines what aspects of the intervention are and are not to be covered by the evaluation.	Chapters 1 and 2, Volume I.	
	6	The evaluation's objectives are specific and realistic. They are clearly related to the evaluation purpose.	Chapters 1 and 2, Volume I.	

IV. EVALUATION METHODOLOGY AND DESIGN	1	The evaluation framework is clearly explained. It establishes the evaluation questions, data sources, and methods for data collection.	Chapter 1, Volume I and Annex A, Volume II (evaluation objectives and questions), Chapter 2, Volume I and Chapters 2–7, Volume II (data sources, data collection, and questionnaires).	Summary sections at the end of Chapters 4–7 in Volume I also list the evaluation questions relevant to each topic.
	2	The product describes and justifies which evaluation criteria are applied (e.g. OECD Development Assistance Committee (DAC)). This includes discussion around which criteria were not relevant for this evaluation.	Chapter 1, Volume I.	This evaluation addressed the effectiveness, impact, and sustainability OECD-DAC criteria. The efficiency criterion was assessed by the 2017 TDP Implementation Review conducted as part of this evaluation. The relevance criterion was assessed through the evaluation baseline research.
	3	The evaluation methods are described and justified. These methods are appropriate for addressing the evaluation questions.	Chapter 2, Volume I, and Chapters 2–7, Volume II.	
	4	The methodology is appropriate for assessing the cross-cutting issues of gender, poverty, human rights, HIV/AIDS, environment, anti-corruption, capacity building, and power relations.	Section 4.2, Chapter 5, and Chapter 7, Volume I.	In this evaluation the impact-level findings are analysed along the lines of gender and poverty, and some contextual findings on power relations within households, and between schools and LGEAs and SUBEBs, are presented. Environment, HIV/AIDS, anti-corruption and capacity building are not directly relevant to this evaluation.
	5	The sampling strategy is described, and is appropriate. Primary and secondary data sources are appropriate, adequate and reliable. Sample sizes are adequate.	Chapter 2, Volume I, and Sections 3.2 and 5.1, Volume II.	
	6	The design provides for multiple lines of inquiry and/or triangulation of data.	Chapters 3– 7, Volume I.	The evaluation presents relevant findings from other national and state-level surveys and studies where available, and also from the 2017 TDP Implementation Review and 2016 TDP Formative Study.
	7	The methodology enables the collection and analysis of disaggregated data to show difference between groups.	Chapters 4–7, Volume I.	Sample size was designed to evaluate impact in the programme areas. Disaggregation by pupil background characteristics and state is provided for the key learning indicators. High sample attrition means disaggregation is not always possible for head

				teachers and teachers. Results for selected key indicators are disaggregated by state.
	8	Any methodological limitations are acknowledged and their impact on the evaluation discussed. The limitations are acceptable and/or they are adequately addressed.	Chapter 2 and Sections 3.3 and 5.5, Volume II.	
	9	Any departures from the TOR, inception phase and / or original evaluation design are adequately explained.	Chapter 1, Volume I.	
	10	The product discusses any inherent imbalances or biases that interviews and other data collection may have created.	Chapters 2, 3, and 5, Volume II.	
	11	The product describes how any bias has been overcome.	Chapters 2, 3, and 5, Volume II.	
V. IMPLEMENTATION	1	Instruments were tested and validated (e.g. pre-testing of questionnaires).	Chapters 4 and 5, Volume II.	
	2	Data was collected in an appropriate and respectful manner, taking into account cultural, ethical, and legal concerns.	Chapters 4, 5, and 9, Volume II.	
	3	There was an appropriate level of involvement from the various stakeholders in the design and implementation of the evaluation.	Chapters 4, 5, and 10, Volume II.	
	4	The evaluation process provided affected stakeholders with access to evaluation-related information in forms that respect people and honour confidentiality.	Chapter 10, Volume II.	
	5	The evaluation process was transparent enough to ensure its legitimacy.	Chapters 4, 5, and 10, Volume II.	
	6	Where primary stakeholders were not consulted due to the scope of the evaluation, relevant documentation drawing on secondary data sources were identified and referred to.		Primary stakeholders were consulted.
	7	Any summary or description of consultees takes into account ethical, privacy, and security concerns. (The document should only provide a summary of the number and level of staff interviewed, by organisation.)	Chapter 9, Volume II.	
	8	To what extent has the evaluation been implemented in accordance with Paris Declaration principles? Have issues of country ownership and management been addressed? To what extent has the evaluation used country systems? How far has the evaluation harmonised approaches with other donors? Has the evaluation	Chapter 10, Volume II.	

		contributed to building evaluation capacity within partner countries?		
VI. ANALYSIS	1	Information is presented, analysed, and interpreted systematically and logically.	Volumes I and II.	
	2	The analysis is presented against the evaluation questions and criteria.	Chapters 1 and 3–7, Volume I.	
	3	The evaluation is transparent about the sources and quality of information, and references or sources are provided.	Chapters 2–7, Volume II (data collection, checking, and quality assurance), and list of references used, Volume I.	The data collection process, data checking and cleaning process, data quality assurance, and development of the questionnaires are described in the report. The definitions of the various indicators are provided in the evaluation matrix. Any additional studies and sources that are used for the analysis are listed in the references.
	4	Evidence can be traced through the analysis and into findings and recommendations. There is sufficient cross-referencing.	Volumes I and II. Chapter 9, Volume I, links the findings to recommendations and lessons.	
	5	The analysis includes an appropriate reflection of the views of different stakeholders (reflecting diverse interests).	Volume I and Chapter 10, Volume II.	
	6	The analysis is disaggregated to show impact on, and outcomes affecting, the different stakeholder groups.	Chapters 5–7, Volume I.	
	7	The analysis explores the cross- cutting issues of gender, poverty, human rights, HIV/AIDS, environment, anti-corruption, capacity building, and power relations.	Section 4.2, Chapter 5, and Chapter 7, Volume I.	In this evaluation the impact-level findings are analysed along the lines of gender and poverty, and some contextual findings on power relations within households, and between schools and LGEAs and SUBEBs, are presented. Environment, HIV/AIDS, anti-corruption, and capacity building are not directly relevant to this evaluation.
VII. FINDINGS	1	The findings follow logically from the analysis.	Volume I.	Volume I of the endline report follows the logical chain of the TDP intervention from outputs to outcome to impact.
	2	The findings address the evaluation questions and criteria.	Chapters 3–7, Volume I	This evaluation addressed the effectiveness, impact, and

			(findings and evaluation questions), and Annex A, Volume II (evaluation matrix with all questions).	sustainability OECD-DAC criteria. The efficiency criterion was assessed by the 2017 TDP Implementation Review conducted as part of this evaluation. The relevance criterion was assessed through the evaluation baseline research. The evaluation questions are set out in the evaluation matrix.
	3	The relevance of the context (e.g. developmental, policy, institutional) is taken into account.	Chapter 4, Volume I.	The relevance of the context was examined in detail by the baseline research for the 2016 TDP Impact Evaluation Baseline Report.
	4	The evidence is clear and sufficiently triangulated.	Volume I.	
	5	The findings are useful and they are presented in ways that are accessible to different users.	Volumes I and II.	In Volume I means and proportions for the quantitative indicators are presented, and figures are used for key findings. In Volume II detailed statistical tables for each indicator presented in Volume I are included.
	6	The findings reflect diverse views and interests. If they do not, there are adequate explanations for omissions.	Chapters 4–8, Volume I.	
	7	There are appropriate and sufficient findings provided around the cross-cutting issues of gender, poverty, human rights, HIV/AIDS, environment, anti-corruption, capacity building, and power relations.	Section 4.2, Chapter 5, and Chapter 7, Volume I.	In this evaluation the impact-level findings are analysed along the lines of gender and poverty, and some contextual findings on power relations within households, and between schools and LGEAs and SUBEBs, are presented. Environment, HIV/AIDS, anti-corruption and capacity building are not directly relevant to this evaluation.
	8	Issues of attribution are considered.	Chapter 2, Section 6.2 and Section 7.1, Volume I, and Chapter 3, Volume II.	
	9	Unintended and unexpected findings are identified.	Chapters 4, 5, 6, 7, and 9, Volume I.	This is discussed as relevant in the different result chapters.
VIII. RECOMMENDATIONS	1	The recommendations follow logically from the findings and evidence cited.	Chapter 9, Volume I.	
	2	The recommendations are relevant to the evaluation and targeted at the intended users.	Chapter 9, Volume I.	

	3	The recommendations are prioritised and clearly presented, enabling individuals or departments to follow up on each specific recommendation.	Chapter 9, Volume I.	
IX. LESSONS	1	The lessons contribute to general knowledge and they are useful.	Chapter 9, Volume I.	
	2	The lessons are valid (i.e. they have not been generalised from single point findings).	Chapter 9, Volume I.	
	3	The lessons reflect the interests of different stakeholders, including different sexes.	Chapter 9, Volume I.	
	4	The lessons are presented separately, with a clear logical distinction between the findings, recommendations, and lessons learned.	Chapter 9, Volume I.	
X. USEFULNESS	1	The report addresses the needs of the TOR, and evaluation questions are adequately covered by the report. If this is not the case, departures from the TOR are justified.	Chapter 1, Volume I.	Please note that the 2017 Impact Evaluation Endline Plan, constitutes the final intended design of this evaluation and is the key reference document for this evaluation.
	2	The evaluation has been designed and managed to meet the information and decision making needs of the intended users.	Chapter 1, Volume I.	
	3	Stakeholders and end-users have been given opportunities to comment on the draft findings, recommendations, and lessons. The evaluation report reflects those comments and acknowledges disagreements.	Chapter 1, Volume I, and Chapter 10, Volume II	The detailed plan for the endline research, including major design features and choices, and the implementation review were discussed and agreed with DFID and the programme. The Steering Committee for the evaluation will review and provide feedback on the endline report, and feedback from state-level decision makers will be sought at dissemination events in the states by evaluation team members.
	4	There is a communications plan within the report. It suggests how dissemination of the evaluation results could lead to improved accountability.	Chapter 10, Volume II	
XI. INDEPENDENCE	1	Differences of opinion (within the evaluation team, or among stakeholders consulted) are fully acknowledged in the report.		All findings, recommendations, and lessons, have been discussed by the members of the evaluation team and there is agreement on these.
	2	Any conflicts of interest are openly discussed.		The evaluation is independent and there are no conflicts of interest.

	3	The report indicates whether the evaluation team was able to work freely and without interference.		The report team was able to work freely and without interference. This is not discussed further in the report.
	4	Information sources and their contributions were independent of other parties with an interest in the evaluation.		Information sources including teachers, head teachers, and staff of other programmes, may have an interest in the evaluation. However, all sources of information are made clear in the report.

Annex D Final sample design and weighting procedures for the TDP baseline survey

This section contains the sample design and weighting note prepared by David Megill in 2015 for the TDP baseline report.

D.1 Background

This annex will begin by briefly describing the implementation of TDP and the final sampling plan for the baseline survey, followed by the weighting procedures based on that sample design. Useful reference documents are the earlier report recommending a sampling plan for the baseline survey (Megill, 2014), the TDP Evaluation Framework and Plan (EDOREN, 2014), which includes in its annexes a description of methodology for calculating the minimum detectable effect for a DID estimate.

The TDP endline survey was conducted in 2017 after three years of the programme implementation. For this reason the baseline sample schools will be part of a panel to be followed up in the endline survey to measure trends in the indicators. A sample of control schools was also included in the baseline survey, so that the trends in the key indicators for TDP schools can be compared to those for the control group. This will involve a DID analysis, as described in the reference documents.

The sampling and weighting procedures described in this report were developed in collaboration with various OPM staff, including Sourovi De and Matthew Powell, as well as Bukola Oyinloye of TDP. The sampling consultant appreciates their collaboration. This technical assistance was provided through the EDOREN Project, funded by DFID.

D.2 Implementation of TDP during the first phase and population for evaluation study

During the first phase TDP was implemented in public schools of 14 LGAs within each of the three states (Jigawa, Katsina, and Zamfara). Within each LGA the schools were clustered based on geographical proximity in order to facilitate the training and periodic meetings of the teachers in each cluster, and to create a broader peer network within the locality. Within each LGA, two clusters of 12 primary schools each were identified: one cluster was randomly assigned to the TDP treatment group and the other to the control group. This strategy was related to the evaluation plan for measuring the impact of the TDP intervention in the treatment schools compared to a similar control group without the intervention. In this way a total of 42 clusters were assigned to the treatment group in the three states, with a corresponding total of 42 control clusters in the same LGAs. With 12 primary schools in each cluster, TDP covered a total of 504 schools in the three states, and the control group also included 504 schools in these states. Within each school selected for TDP, the first phase intervention involved the training of four primary teachers: two in English and two in maths.

This TDP implementation was a type of quasi-experimental design, so the population being studied in the TDP baseline survey consists of the set 42 treatment clusters and 42 control clusters in the three states. Originally each cluster had 12 primary schools, but later it was found that a few of the schools did not have eligible Primary 3 pupils, who were the subject of the pupil tests as part of the evaluation. Therefore the final population of schools for some clusters had less than 12 eligible schools. The sample for the baseline survey was selected to represent the eligible schools in the clusters for the three states. Inferences can only be made for the frame of all eligible schools in the clusters for each state. Therefore the sample for the baseline survey was not designed to be representative at the state level.

Within the eligible schools of the treatment and control clusters in each state, the only teachers eligible to be included in the baseline survey for the treatment clusters were the four teachers receiving the TDP training and the head teacher. A similar group of four teachers was chosen in schools of the control clusters. However, some treatment schools had less than four eligible Grade 3 teachers, in which case all of them received TDP training in the treatment schools. In the case of the pupil tests, the population for the evaluation study consisted of all Grade 3 pupils who had classes led by one of the teachers chosen for the study in the treatment and the control schools.

D.3 Sample design for TDP baseline survey

Once the 14 treatment clusters and 14 control clusters were established in each state, the sampling frame consisted of all the eligible public primary schools in each cluster; most clusters had 12 eligible schools each, but a few clusters had fewer schools. In this case the eligible schools in each cluster were considered the PSUs selected at the first sampling stage for the baseline survey.

The stratification of the sampling frame for the TDP baseline survey is by individual treatment or control cluster, since an independent sample of schools was selected from each cluster in the frame. In this case these are not 'clusters' based on the classic sampling terminology: actually, each PSU (school) is a cluster of teachers and pupils.

The first sampling stage consisted of randomly selecting a sample of four schools from each of the 14 treatment clusters and 14 control clusters in each state. All of the four (or fewer) teachers who received TDP training in each sample treatment school and the corresponding group of up to four teachers in each control school were selected with certainty to be tested and observed for the baseline survey, as well as the head teacher from each of these sample schools.

For the pupil tests a sample of eight Primary 3 pupils was selected for the TDP baseline survey from a list of all the eligible Primary 3 pupils who had a class led by one of the eligible teachers who received TDP training in each treatment school or the corresponding teachers chosen in the control schools. In the case of small schools with fewer than eight eligible Primary 3 pupils, all were selected for the baseline survey.

D.4 Weighting procedures for TDP baseline survey

In order to make inferences from the TDP baseline survey data it was necessary to assign appropriate weights to each sample school, teacher, and pupil. The weights are equal to the inverse of the overall sampling probabilities, taking into account each stage of selection. The school, teacher, and pupil weights will be calculated at the school level. Based on the sample design described in the previous section, the probability and corresponding weight for the sample schools would be calculated as follows:

$$p_{Sh} = \frac{n_h}{N_h} \quad \text{and} \quad W_{Sh} = \frac{N_h}{n_h},$$

where:

p_{Sh} = probability of selection for the sample schools in cluster (stratum) h;

n_h = number of sample primary schools successfully enumerated in cluster h for the TDP baseline survey; generally $n_h = 4$;

N_h = total number of eligible primary schools with Grade 3 pupils in cluster h; generally $N_h = 12$; and

W_{Sh} = weight of sample schools in cluster h.

In the case of clusters in which fewer than four sample schools were successfully enumerated for the TDP baseline survey, this formula automatically adjusts the weight for nonresponse.

Each sample school has one head teacher, so the head teacher has the same weight as the school. Since all of the four teachers receiving TDP training in each sample treatment school and the corresponding group of four teachers chosen in each sample control school are included in the TDP baseline survey, the teacher weights are generally equal to the school weights. In the case of small schools with fewer than four eligible teachers, the teacher weight would also be equal to the school weight if all these teachers are successfully tested and observed. However, there are a few cases of sample schools where some eligible teachers could not be enumerated, in which case it will be necessary to adjust the weight for nonresponse. In this case the teacher weight would be calculated as follows:

$$W_{Thi} = W_{Sh} \times \frac{T_{hi}}{T'_{hi}},$$

where:

W_{Thi} = weight for teachers in the i-th sample school in cluster (stratum) h;

T_{hi} = number of eligible teachers included in the study for the i-th sample school in cluster h; generally $T_{hi} = 4$; and

T'_{hi} = number of eligible teachers with completed interviews in the i-th sample school in cluster h.

In the case of the teacher observations, the weights would be calculated in a similar way as the teacher interview weights, but in this case T'_{hi} would be the number of eligible teachers who had been successfully observed.

The weights for the sample Grade 3 pupils who are tested involve components from two sampling stages. The first component of the weight would be the school weight defined previously. The second component would be the inverse of the within-school probability of selection for the sample pupils. In this case the pupil weights can be defined as follows:

$$W_{Phi} = W_{Sh} \times \frac{P_{hi}}{p_{hi}},$$

where:

W_{Phi} = weight for Grade 3 sample pupils who were tested in the i-th sample school in cluster h;

P_{hi} = number of eligible Grade 3 pupils in the i-th sample school in cluster h; and

p_{hi} = number of sample Grade 3 pupils with completed tests in the i-th sample school in cluster h.

Annex E Additional descriptive statistics and list of covariates

Table 10.2: Treatment receipt and assignment (pupil-level)

	Treatment not assigned	Treatment assigned
Treatment not received	711	12
Treatment received	71	771

Table 10.3: Treatment receipt and assignment (teacher-level)

	Treatment not assigned	Treatment assigned
Treatment not received	193	4
Treatment received	13	247

Table 10.4: Treatment receipt and assignment (school-level)

	Treatment not assigned	Treatment assigned
Treatment not received	151	5
Treatment received	13	161

Table 10.5: List of covariates

Outcome	Group of variables	Variables included
Test scores (maths, English, science)	Outcome baseline measurement	Outcome baseline value
	Pupil background characteristics	Gender Dummy for speaking Hausa at home Asset index Dummy for electricity at home Family size Number of rooms at home Dummy for whether parents received primary school education Dummy for whether the pupil eats something at home before school Dummy for whether there is a toilet at home

	School background characteristics	Dummy for whether the school needs major repairs Dummy for whether the school has power supply Number of pupils registered Number of teachers employed Pupil–teacher ratio Dummy for whether a SBMC exists Dummy for whether the school receives support from external organisations Dummy for whether the LGEA conducted a visit to the school more than three times per month Dummy for whether the roof was in a good state Dummy for whether the inner walls were in a good state Dummy for whether the outer walls were in a good state Dummy for whether the playground was in a good state Dummy for whether the windows were in a good state
	Head teacher background characteristics	Age Gender Years of experience, Dummy for whether the head teacher has conducted lesson observation Dummy for whether the head teacher had meetings with teachers Dummy for whether the head teacher has NCE qualification Dummy for whether the head teacher implemented activities to reduce pupils' absenteeism Dummy for whether the head teacher implemented activities to reduce teacher absenteeism Dummy for whether the head teacher attended trainings Dummy for whether the head teacher reported having received the salary on time Dummy for whether the head teacher was absent at least one day during the previous five days Dummy for whether the head teacher was absent at least one day during the last term
	Non-TDP contamination	Dummies for the following programmes: SUBEB, ESSPIN, RANA, Jolly Phonics

	Geographical controls	State dummies LGA dummies
Teachers' positive interaction	Outcome baseline measurement	Outcome baseline value
	Teacher background characteristics	Age Gender Years of experience Dummy for whether the head teacher has NCE qualification Dummy for whether the teacher attended trainings Dummy for whether the teacher reported to have received the salary on time
	School background characteristics	Dummy for whether the school needs major repairs Dummy for whether the school has power supply Number of pupils registered Number of teachers employed Pupil–teacher ratio Dummy for whether the school receives support from external organisations Dummy for whether the LGEA conducted a visit to the school more than three times per month Dummy for whether the head teacher had meetings with teachers Dummy for whether the head teacher had meetings with teachers Dummy for whether the head teacher implemented activities to reduce pupils' absenteeism Dummy for whether the head teacher implemented activities to reduce teacher absenteeism
	Non-TDP contamination	Dummies for the following programmes: SUBEB, ESSPIN, RANA, Jolly Phonics
	Geographical controls	State dummies LGA dummies
Teacher absenteeism	Outcome baseline measurement	Outcome baseline value
	School background characteristics	Number of pupils registered Pupil–teacher ratio Dummy for whether the head teacher implemented activities to reduce pupils' absenteeism

		Dummy for whether the head teacher implemented activities to reduce teacher absenteeism Dummy for whether the LGEA conducted a visit to the school more than three times per month Dummy for whether the school receives support from external organisations Dummy for whether the school needs major repairs Dummy for whether the school has power supply Dummy for whether the school has drinking water Dummy for whether there is a staffroom in school Number of classrooms used for teaching activities Number of toilets working Dummy for whether the roof was in a good state Dummy for whether the inner walls were in a good state Dummy for whether the outer walls were in a good state Dummy for whether the playground was in a good state Dummy for whether the windows were in a good state
	Head teacher background characteristics	Age Gender Years of experience in total Years of experience in current school Dummy for whether the head teacher has conducted a lesson observation Dummy for whether the head teacher had meetings with teachers Dummy for whether the head teacher has NCE qualification Dummy for whether the head teacher attended trainings
	Non-TDP contamination	Dummies for the following programmes: SUBEB, ESSPIN, RANA, Jolly Phonics
	Geographical controls	State dummies LGA dummies

Annex F Estimation results

F.1 Teacher positive interaction: IV and OLS results

	1	2	3
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	0.040*	0.029*	
	(0.016)	(0.011)	
Treatment assignment			0.030*
			(0.012)
Gender	0.009	0.008	0.010
	(0.015)	(0.015)	(0.015)
Age	-0.002*	-0.002*	-0.002*
	(0.001)	(0.001)	(0.001)
Total teaching experience in ANY school in 2014 (years)	0.000	0.001	0.001
	(0.001)	(0.001)	(0.001)
Teacher has NCE qualification	0.033**	0.033**	0.034**
	(0.011)	(0.011)	(0.011)
Teacher attended teaching-related training in last two years	0.026	0.026	0.026
	(0.014)	(0.014)	(0.014)
Receives salary on time	-0.028	-0.028	-0.030*
	(0.015)	(0.015)	(0.015)
Number of teachers employed (excluding voluntary/temporary teachers)	0.004***	0.004***	0.003***
	(0.001)	(0.001)	(0.001)
Number of class 1–6 pupils registered	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)
One or more formal head teacher–teacher meetings per week	0.007	0.008	0.008
	(0.011)	(0.011)	(0.011)
Conducted lesson observations last 10 days	-0.018	-0.018	-0.019
	(0.013)	(0.013)	(0.013)
Head teacher took action to improve pupil absenteeism last school year	-0.091*	-0.092*	-0.097*
	(0.035)	(0.036)	(0.042)
Head teacher took action to improve teacher absenteeism last school year	0.033	0.034	0.036
	(0.019)	(0.020)	(0.020)
LGEA visit more than three times in a month	0.001	0.001	-0.002
	(0.011)	(0.011)	(0.011)
Pupil–teacher ratio	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)
School receives support from other organisation/programme	-0.005	-0.004	-0.004
	(0.012)	(0.012)	(0.012)
School has electricity	0.021	0.020	0.017
	(0.017)	(0.017)	(0.016)
School needs major repair	0.006	0.005	0.007
	(0.016)	(0.016)	(0.015)

School got SUBEB-led INSET	-0.011	-0.017	-0.013
	(0.015)	(0.014)	(0.015)
School got RANA	0.009	0.008	0.008
	(0.025)	(0.025)	(0.025)
School got Jolly Phonics	0.004	0.004	0.006
	(0.012)	(0.012)	(0.012)
School got ESSPIN	0.011	0.014	0.017
	(0.018)	(0.019)	(0.019)
Constant	0.261***	0.268***	0.275***
	(0.052)	(0.054)	(0.056)
Observations	443	443	443
R-squared	0.174	0.177	0.177
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included.			
**** p<0.001, ** p<0.01, * p<0.05*			

F.2 Teacher positive interaction: panel and DID results

	1	2	3
Estimation technique	**Panel RE**	**Panel IV RE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	0.029*	0.044**	0.004
	(0.012)	(0.016)	(0.012)
DID			0.030
			(0.015)
Time	0.028**	0.021	-0.001
	(0.010)	(0.011)	(0.014)
Gender			-0.010
			(0.011)
Head teacher age	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)
Total teaching experience in ANY school in 2014 (years)	0.002	0.002	0.001
	(0.001)	(0.001)	(0.001)
Teacher has NCE qualification			0.014
			(0.009)
Teacher attended teaching-related training in last two years			0.016
			(0.009)
Receives salary on time			0.024*
			(0.012)
Number of teachers employed, excluding voluntary/temporary teachers			0.002**
			(0.001)
Number of class 1–6 pupils registered			-0.000
			(0.000)

One or more formal head teacher–teacher meetings per week (% of head teachers)			0.005
			(0.009)
Conducted lesson observations last 10 days (% of head teachers)			0.007
			(0.009)
Head teacher took action to improve pupil absenteeism last school year (% of head teachers)			-0.028
			(0.024)
Head teacher took action to improve teacher absenteeism last school year (% head teachers)			0.024
			(0.015)
Pupil–teacher ratio			0.000*
			(0.000)
LGEA visit more than three times in a month			-0.033***
			(0.009)
School receives support from other organisation/programme (% of schools)			0.000
			(0.000)
School has electricity			-0.029
			(0.015)
School needs major repair			-0.021
			(0.012)
School has electricity	-0.000	0.002	-0.029
	(0.013)	(0.013)	(0.015)
School got SUBEB-led INSET	-0.014	-0.009	-0.015
	(0.012)	(0.013)	(0.012)
School got RANA	-0.003	-0.002	0.021
	(0.025)	(0.025)	(0.020)
School got Jolly Phonics	0.002	0.002	0.001
	(0.009)	(0.009)	(0.009)
School got ESSPIN	0.031**	0.030**	0.034
	(0.011)	(0.011)	(0.017)
Constant	0.248***	0.246***	0.181***
	(0.025)	(0.025)	(0.039)
Observations	909	909	867
Number of IDs	457	457	0.145
Standard errors in parentheses. Survey weights included.			
**** p<0.001, ** p<0.01, * p<0.05*			

F.3 Teacher absenteeism: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	-3.107	-2.707	
	(2.099)	(1.890)	
Treatment assignment			-2.784
			(1.877)
Number of class 1–6 pupils registered	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)
Head teacher took action to improve pupil absenteeism last school year	1.518	1.669	2.291
	(5.868)	(5.840)	(5.975)
Head teacher took action to improve teacher absenteeism last school year	5.696	5.679	5.710
	(4.470)	(4.476)	(4.470)
LGEA visit more than three times a month	0.323	0.283	0.318
	(1.851)	(1.856)	(1.848)
Pupil–teacher ratio	-0.013	-0.013	-0.011
	(0.025)	(0.025)	(0.025)
School receives support from other organisation/programme	1.449	1.435	1.424
	(2.065)	(2.071)	(2.068)
School has electricity	7.163**	7.195**	7.435**
	(2.504)	(2.508)	(2.523)
School needs major repairs	-1.320	-1.342	-1.470
	(2.849)	(2.850)	(2.852)
School roof good condition	2.353	2.367	2.227
	(2.413)	(2.415)	(2.437)
Class inner walls good condition	-10.952***	-10.969***	-10.848***
	(2.576)	(2.575)	(2.590)
Class outer walls good condition	6.829**	6.828**	6.632**
	(2.294)	(2.293)	(2.294)
School windows good condition	2.754	2.726	3.000
	(2.646)	(2.646)	(2.648)
School playground good condition	-0.022	-0.039	-0.355
	(2.219)	(2.217)	(2.231)
Head teacher gender	8.236	8.245	8.348
	(4.947)	(4.943)	(4.955)
Head teacher age	-0.136	-0.138	-0.149
	(0.121)	(0.121)	(0.121)
Total head teacher teaching experience in any school (years)	-0.098	-0.095	-0.082
	(0.112)	(0.112)	(0.113)
Head teacher has NCE qualification	2.431	2.460	2.524
	(2.282)	(2.286)	(2.278)
Head teacher attended teaching-related training in last years	1.463	1.485	1.511
	(2.312)	(2.309)	(2.308)
One or more formal head teacher–teacher meetings per week	-2.503	-2.500	-2.388

	(1.843)	(1.846)	(1.851)
Conducted lesson observations last 10 days	3.049	3.042	2.914
	(2.069)	(2.067)	(2.059)
This school has an SBMC	-2.857	-2.888	-2.923
	(5.329)	(5.314)	(5.261)
School got SUBEB-led INSET	0.276	0.523	0.346
	(2.485)	(2.415)	(2.458)
School got RANA	-0.369	-0.284	-0.130
	(4.695)	(4.698)	(4.648)
School got Jolly Phonics	1.736	1.742	1.759
	(1.930)	(1.931)	(1.937)
School got ESSPIN	-5.242	-5.284	-5.489
	(3.355)	(3.342)	(3.330)
Average daily teacher absence from school (baseline)	0.203**	0.204**	0.204**
	(0.068)	(0.068)	(0.068)
Constant	19.174	18.901	19.139
	(10.958)	(10.891)	(10.984)
Observations	305	305	305
R-squared	0.182	0.184	0.184
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.4 Pupil learning – maths test scores: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	5.619	11.750	
	(7.163)	(7.447)	
Treatment assignment			5.074
			(6.481)
Maths: Rasch score (scaled with survey weights)	0.300***	0.299***	0.300***
	(0.038)	(0.038)	(0.038)
Pupil is female	-10.766*	-10.667*	-10.767*
	(4.966)	(4.959)	(4.965)
Speak Hausa at home	29.568*	29.448*	29.910*
	(13.156)	(13.407)	(13.049)
Mean value of household asset index	-0.704	-0.776	-0.767
	(3.406)	(3.399)	(3.440)
Electricity at home	4.485	4.952	4.506
	(8.332)	(8.240)	(8.347)
Average family size [pupils]	-0.314	-0.307	-0.310
	(0.543)	(0.539)	(0.543)

Average no. of rooms in pupil's house [pupils]	3.931	3.920	3.931
	(2.141)	(2.112)	(2.156)
At least one parent/guardian has completed primary school	0.364	0.233	0.347
	(7.182)	(7.162)	(7.200)
Normally eat something during long break	0.787	0.501	0.885
	(5.839)	(5.882)	(5.869)
Flush toilet at home	31.795***	31.759***	31.798***
	(8.865)	(8.860)	(8.877)
School needs major repair	-1.554	-2.338	-1.081
	(10.348)	(10.334)	(10.356)
School has electricity	1.106	1.922	0.780
	(8.954)	(9.153)	(8.881)
Number of class 1–6 pupils registered	0.014	0.016	0.015
	(0.009)	(0.009)	(0.009)
This school has an SBMC	87.599**	87.411*	87.991**
	(32.887)	(35.862)	(32.773)
School receives support from other organisation/programme	6.436	6.135	6.653
	(6.832)	(6.866)	(6.871)
Number of teachers employed, excluding voluntary/temporary teachers	-0.276	-0.355	-0.290
	(0.481)	(0.479)	(0.489)
LGEA visit more than three times in a month	-4.369	-5.153	-4.243
	(6.838)	(6.836)	(6.834)
Pupil–teacher ratio	-0.128	-0.137	-0.133
	(0.134)	(0.131)	(0.137)
Head teacher's age	-0.285	-0.302	-0.282
	(0.709)	(0.709)	(0.711)
Total teaching experience in ANY school in 2014 (years)	-0.633	-0.564	-0.653
	(0.518)	(0.510)	(0.522)
Head teacher's gender	-14.366	-14.752	-14.208
	(14.848)	(15.017)	(14.840)
Conducted lesson observation last 10 days	14.423	13.827	14.902
	(7.794)	(7.796)	(7.769)
One or more formal head teacher–teacher meetings per week	-4.676	-4.808	-4.788
	(6.527)	(6.580)	(6.575)
Head teacher has NCE qualification	-17.625	-17.183	-17.751
	(9.966)	(9.863)	(10.048)
Head teacher took action to improve pupil absenteeism last school year	-2.860	-1.458	-5.277
	(18.390)	(18.519)	(19.059)
Head teacher took action to improve teacher absenteeism last school year	21.749	21.168	22.820
	(14.638)	(14.682)	(14.993)
Head teacher attended teaching-related training	-5.757	-4.645	-6.012
	(8.312)	(8.228)	(8.285)
Head teacher receives salary on time	43.693**	44.997**	43.718**
	(13.752)	(14.060)	(13.836)
Head teacher absent one day or more last five days [head teacher]	8.159	7.781	8.539
	(9.973)	(9.818)	(9.919)

Head teacher absent one day or more last term [head teacher]	7.129	7.629	7.130
	(6.223)	(6.191)	(6.245)
School roof good condition [schools]	-13.619	-14.455	-12.925
	(8.018)	(7.871)	(8.142)
Class inner walls good condition [schools]	10.868	9.218	10.928
	(12.178)	(12.192)	(12.201)
Class outer walls good condition [schools]	7.542	9.234	7.649
	(14.338)	(14.458)	(14.435)
School playground good condition [schools]	3.866	3.488	4.482
	(7.350)	(7.300)	(7.433)
School windows good condition [schools]	-0.992	-0.413	-1.944
	(11.062)	(11.117)	(11.177)
School got SUBEB-led INSET	-4.656	-1.319	-4.109
	(12.774)	(13.210)	(13.338)
School got RANA	-20.429	-19.543	-20.939
	(11.838)	(11.779)	(11.973)
School got Jolly Phonics	-12.392	-12.300	-12.461
	(7.246)	(7.243)	(7.270)
School got ESSPIN	-13.334	-13.055	-13.246
	(14.493)	(14.537)	(14.557)
Constant	234.926***	131.754	200.413**
	(60.099)	(67.164)	(64.016)
Observations	1,378	1,378	1,378
R-squared	0.221	0.222	0.220
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.5 Pupil learning – maths test scores: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	20.107	5.483	-5.186
	(12.818)	(10.367)	(8.080)
Time	-13.878	3.888	-8.295
	(10.745)	(9.395)	(12.735)
DID			22.477*
			(10.754)
Pupil is female			-23.518***
			(4.315)
Speak Hausa at home			2.493
			(11.519)
Mean value of household asset index			12.311***

			(3.087)
Electricity at home			12.751
			(6.747)
Average family size			0.102
			(0.438)
Average no. of rooms in pupil's house			5.290**
			(1.912)
At least one parent/guardian has completed primary school			5.384
			(4.912)
Flush toilet at home			14.255
			(7.241)
School needs major repair			19.125**
			(7.244)
School has electricity	-7.975	-9.530	7.016
	(20.814)	(18.532)	(7.439)
Number of class 1–6 pupils registered			0.021**
			(0.007)
School receives support from other organisation/programme			-0.048
			(0.098)
Number of teachers employed, excluding voluntary/temporary teachers			-0.834
			(0.429)
LGEA visit more than three times in a month			-8.729
			(5.207)
Pupil–teacher ratio			-0.171*
			(0.066)
This school has an SBMC			1.755
			(50.503)
Head teacher's age	0.667	0.475	-0.222
	(0.821)	(0.709)	(0.551)
Total teaching experience in ANY school in 2014 (years)	0.060	-0.105	-0.086
	(0.763)	(0.612)	(0.525)
Head teacher's gender			12.060
			(7.912)
Conducted lesson observations last 10 days			10.392
			(5.686)
One or more formal head teacher–teacher meetings per week			-3.275
			(5.897)
Head teacher has NCE qualification			21.689**
			(7.296)
Head teacher took action to improve pupil absenteeism last school year			-27.943*
			(13.252)
Head teacher took action to improve teacher absenteeism last school year			-8.547
			(10.735)
Head teacher attended teaching-related training			3.848
			(7.283)
Head teacher receives salary on time			-23.861*

			(11.385)
Head teacher absent one day or more last five days			1.803
			(6.805)
Head teacher absent one day or more last term			6.871
			(5.494)
School got SUBEB-led INSET	29.771	-3.039	-8.101
	(19.177)	(11.246)	(9.417)
School got RANA	-65.711	-17.188	14.115
	(42.233)	(14.966)	(16.356)
School got Jolly Phonics	-21.645*	-12.171	7.329
	(11.904)	(9.462)	(5.634)
School got ESSPIN	46.632***	46.015***	-2.246
	(13.522)	(10.356)	(9.100)
Constant	471.083***	459.501***	459.812***
	(29.794)	(24.663)	(64.360)
Observations	3,119	3,119	2,871
R-squared	0.062		0.155
State FE	NO	NO	YES
LGA FE	NO	NO	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.6 Pupil learning – English test scores: IV and OLS results

	1	2	3
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	2.689	0.898	
	(7.379)	(6.671)	
Treatment assignment			2.427
			(6.652)
Maths: Rasch score (scaled with survey weights)	0.326***	0.327***	0.327***
	(0.034)	(0.034)	(0.035)
Pupil is female	-10.104	-10.133	-10.103
	(5.335)	(5.330)	(5.335)
Speak Hausa at home	47.413*	47.450*	47.575*
	(18.474)	(18.496)	(18.491)
Mean value of household asset index	-2.668	-2.651	-2.705
	(3.182)	(3.185)	(3.193)
Electricity at home	13.505	13.364	13.511
	(7.472)	(7.445)	(7.474)
Average family size	-0.628	-0.630	-0.626
	(0.522)	(0.521)	(0.521)
Average no. of rooms in pupil's house	-0.863	-0.862	-0.866
	(2.066)	(2.070)	(2.068)

At least one parent/guardian has completed primary school	2.890	2.930	2.886
	(6.344)	(6.340)	(6.348)
Normally eat something during long break	-2.355	-2.277	-2.316
	(6.029)	(6.022)	(6.017)
Flush toilet at home	29.958***	29.968***	29.960***
	(7.223)	(7.224)	(7.226)
School needs major repair	-16.689	-16.466	-16.471
	(10.625)	(10.569)	(10.658)
School has electricity	-1.789	-2.024	-1.942
	(9.817)	(9.766)	(9.773)
Number of class 1–6 pupils registered	0.007	0.006	0.007
	(0.008)	(0.008)	(0.008)
This school has an SBMC	102.395***	102.451***	102.597***
	(16.689)	(16.861)	(16.528)
School receives support from other organisation/programme	5.658	5.745	5.759
	(6.325)	(6.285)	(6.275)
Number of teachers employed, excluding voluntary/temporary teachers	0.622	0.646	0.616
	(0.463)	(0.461)	(0.469)
LGEA visit more than three times in a month	-3.298	-3.079	-3.245
	(6.065)	(6.065)	(6.071)
Pupil–teacher ratio	0.077	0.080	0.075
	(0.116)	(0.115)	(0.118)
Head teacher's age	-0.321	-0.316	-0.319
	(0.638)	(0.636)	(0.636)
Total teaching experience in ANY school in 2014 (years)	-0.281	-0.300	-0.290
	(0.529)	(0.516)	(0.521)
Head teacher's gender	-25.905	-25.811	-25.838
	(13.288)	(13.209)	(13.253)
Conducted lesson observation last 10 days	-4.088	-3.919	-3.861
	(6.934)	(6.892)	(6.904)
One or more formal head teacher–teacher meetings per week	-8.371	-8.335	-8.426
	(6.738)	(6.723)	(6.773)
Head teacher has NCE qualification	2.212	2.088	2.153
	(8.524)	(8.588)	(8.549)
Head teacher took action to improve pupil absenteeism last school year	-35.576	-35.967	-36.711
	(23.453)	(23.587)	(23.954)
Head teacher took action to improve teacher absenteeism last school year	45.718**	45.883**	46.229**
	(16.392)	(16.347)	(16.388)
Head teacher attended teaching-related training	1.826	1.504	1.706
	(7.206)	(7.177)	(7.117)
Head teacher receives salary on time	49.389***	48.992***	49.388***
	(12.832)	(12.953)	(12.842)
Head teacher absent one day or more last five days [head teacher]	-9.626	-9.524	-9.451
	(9.364)	(9.395)	(9.428)
Head teacher absent one day or more last term [head teacher]	9.226	9.089	9.234
	(6.075)	(6.039)	(6.058)

School roof good condition [schools]	-13.714	-13.472	-13.390
	(7.557)	(7.485)	(7.748)
Class inner walls good condition [schools]	0.858	1.333	0.886
	(12.107)	(12.119)	(12.106)
Class outer walls good condition [schools]	12.769	12.268	12.814
	(11.341)	(11.387)	(11.370)
School playground good condition [schools]	-4.464	-4.357	-4.169
	(7.391)	(7.373)	(7.491)
School windows good condition [schools]	-1.044	-1.209	-1.500
	(9.335)	(9.237)	(9.418)
School got SUBEB-led INSET	2.808	1.841	3.074
	(10.775)	(10.797)	(11.184)
School got RANA	6.133	5.881	5.896
	(13.579)	(13.617)	(13.621)
School got Jolly Phonics	-17.129*	-17.162*	-17.168*
	(7.156)	(7.145)	(7.160)
School got ESSPIN	16.506	16.423	16.543
	(13.747)	(13.734)	(13.742)
Constant	178.984***	102.769	101.696
	(51.415)	(58.079)	(57.546)
Observations	1,375	1,375	1,375
R-squared	0.232	0.232	0.232
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.7 Pupil learning – English test scores: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	-4.395	0.379	7.441
	(12.108)	(10.659)	(7.985)
Time	0.526	1.503	-7.893
	(10.202)	(8.060)	(11.463)
DID			-2.276
			(10.146)
Pupil is female			-19.584***
			(4.941)
Speak Hausa at home			13.882
			(12.777)
Mean value of household asset index			9.833***
			(2.708)
Electricity at home			15.299**

			(5.875)
Average family size			-0.019
			(0.423)
Average no. of rooms in pupil's house			4.425*
			(1.939)
At least one parent/guardian has completed primary school			7.101
			(5.640)
Flush toilet at home			11.690
			(6.810)
School needs major repair			9.583
			(7.683)
School has electricity	-2.794	-9.250	2.549
	(18.153)	(15.805)	(6.745)
Number of class 1–6 pupils registered			0.019***
			(0.005)
School receives support from other organisation/programme			0.120
			(0.095)
Number of teachers employed, excluding voluntary/temporary teachers			-0.316
			(0.347)
LGEA visit more than three times in a month			-6.341
			(5.699)
Pupil–teacher ratio			-0.081
			(0.059)
This school has an SBMC			8.235
			(65.401)
Head teacher's age	1.066	0.190	-0.239
	(0.826)	(0.878)	(0.622)
Total teaching experience in ANY school in 2014 (years)	0.218	0.455	0.175
	(0.716)	(0.837)	(0.562)
Head teacher's gender			14.658
			(13.047)
Conducted lesson observations last 10 days			10.002
			(5.467)
One or more formal head teacher–teacher meetings per week			-6.859
			(6.319)
Head teacher has NCE qualification			14.945*
			(6.439)
Head teacher took action to improve pupil absenteeism last school year			-43.469***
			(12.343)
Head teacher took action to improve teacher absenteeism last school year			12.253
			(12.304)
Head teacher attended teaching-related training			4.904
			(7.639)
Head teacher receives salary on time			-17.985
			(12.612)
Head teacher absent one day or more last five days			4.068

			(7.226)
Head teacher absent one day or more last term			2.444
			(5.023)
School got SUBEB-led INSET	20.895	13.824	-6.954
	(13.312)	(13.081)	(10.647)
School got RANA	-36.366	2.103	7.050
	(37.159)	(14.067)	(12.747)
School got Jolly Phonics	-32.574***	-18.326	1.451
	(11.084)	(9.646)	(5.699)
School got ESSPIN	67.583***	70.293***	23.546*
	(10.650)	(11.116)	(11.150)
Constant	448.321***	455.461***	401.338***
	(27.837)	(26.645)	(74.388)
Observations	3,114	3,114	2,867
R-squared	0.075		0.134
State FE	NO	NO	YES
LGA FE	NO	NO	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.8 Pupil learning – science and technology test scores: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	-2.429	3.086	
	(7.478)	(6.604)	
Treatment assignment			-2.193
			(6.749)
Maths: Rasch score (scaled with survey weights)	0.158***	0.157***	0.158***
	(0.029)	(0.029)	(0.030)
Pupil is female	-25.190***	-25.097***	-25.189***
	(5.451)	(5.454)	(5.453)
Speak Hausa at home	30.568	30.429	30.412
	(23.229)	(23.132)	(23.213)
Mean value of household asset index	4.531	4.454	4.557
	(2.905)	(2.916)	(2.904)
Electricity at home	-7.236	-6.808	-7.241
	(8.333)	(8.288)	(8.328)
Average family size	0.063	0.069	0.061
	(0.454)	(0.455)	(0.454)
Average no. of rooms in pupil's house	0.145	0.137	0.146
	(2.159)	(2.144)	(2.156)

At least one parent/guardian has completed primary school	6.844	6.737	6.852
	(6.535)	(6.536)	(6.532)
Normally eat something during long break	9.394	9.148	9.357
	(6.242)	(6.229)	(6.228)
Flush toilet at home	36.524***	36.495***	36.524***
	(7.869)	(7.895)	(7.868)
School needs major repair	-26.086*	-26.781*	-26.285*
	(10.296)	(10.408)	(10.467)
School has electricity	-11.231	-10.500	-11.090
	(8.742)	(8.843)	(8.660)
Number of class 1–6 pupils registered	0.005	0.006	0.005
	(0.008)	(0.008)	(0.008)
This school has an SBMC	94.305*	94.142*	94.131*
	(44.877)	(47.750)	(44.902)
School receives support from other organisation/programme	8.277	8.001	8.183
	(6.375)	(6.361)	(6.326)
Number of teachers employed, excluding voluntary/temporary teachers	0.117	0.046	0.123
	(0.491)	(0.486)	(0.498)
LGEA visit more than three times in a month	-2.055	-2.763	-2.110
	(6.419)	(6.427)	(6.405)
Pupil–teacher ratio	-0.239*	-0.247*	-0.237*
	(0.111)	(0.109)	(0.113)
Head teacher's age	0.587	0.569	0.585
	(0.766)	(0.767)	(0.764)
Total teaching experience in ANY school in 2014 (years)	-1.389*	-1.326*	-1.380*
	(0.613)	(0.612)	(0.608)
Head teacher's gender	-21.741	-22.080	-21.802
	(11.988)	(12.164)	(11.988)
Conducted lesson observation last 10 days	11.394	10.848	11.186
	(6.408)	(6.415)	(6.309)
One or more formal head teacher–teacher meetings per week	0.186	0.080	0.239
	(5.882)	(5.907)	(5.894)
Head teacher has NCE qualification	-12.462	-12.059	-12.402
	(9.529)	(9.423)	(9.507)
Head teacher took action to improve pupil absenteeism last school year	-8.585	-7.344	-7.550
	(21.391)	(21.156)	(21.748)
Head teacher took action to improve teacher absenteeism last school year	8.708	8.218	8.253
	(15.063)	(15.082)	(15.260)
Head teacher attended teaching-related training	3.853	4.877	3.974
	(7.399)	(7.342)	(7.362)
Head teacher receives salary on time	23.235	24.431	23.228
	(13.705)	(13.677)	(13.683)
Head teacher absent one day or more last five days	17.174	16.795	17.001
	(9.680)	(9.597)	(9.801)
Head teacher absent one day or more last term	8.366	8.838	8.373
	(6.315)	(6.337)	(6.308)

School roof good condition	-11.627	-12.393	-11.930
	(7.172)	(7.057)	(7.025)
Class inner walls good condition	3.744	2.259	3.715
	(14.782)	(14.711)	(14.762)
Class outer walls good condition	-1.363	0.203	-1.395
	(13.484)	(13.499)	(13.489)
School playground good condition	31.616***	31.283***	31.350***
	(6.658)	(6.660)	(6.767)
School windows good condition	7.768	8.242	8.165
	(9.711)	(9.801)	(9.992)
School got SUBEB-led INSET	-0.835	2.178	-1.068
	(10.875)	(10.870)	(11.207)
School got RANA	-42.350***	-41.513***	-42.113***
	(12.127)	(12.090)	(12.256)
School got Jolly Phonics	-20.214*	-20.129*	-20.181*
	(8.283)	(8.283)	(8.258)
School got ESSPIN	6.186	6.433	6.146
	(17.056)	(17.082)	(17.039)
Constant	319.411***	227.342**	232.348**
	(68.565)	(76.823)	(74.993)
Observations	1,380	1,380	1,380
R-squared	0.238	0.239	0.238
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.9 Pupil learning – science and Technology test scores: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	19.330	13.701	-3.462
	(15.833)	(11.270)	(8.577)
Time	-11.004	-5.274	-11.223
	(15.082)	(9.788)	(14.861)
DID			13.738
			(12.934)
Pupil female			-22.134***
			(3.322)
Speak Hausa at home			8.794
			(16.536)
Mean value of household asset index			3.311
			(2.439)
Electricity at home			5.332

			(5.437)
Average family size			-0.172
			(0.377)
Average no. of rooms in pupil's house			2.915
			(1.483)
At least one parent/guardian has completed primary school			8.420
			(5.665)
Flush toilet at home			18.377**
			(6.626)
School needs major repair			7.752
			(7.204)
School has electricity	4.291	-11.937	-6.295
	(17.506)	(17.618)	(7.383)
Number of class 1–6 pupils registered			0.027***
			(0.006)
School receives support from other organisation/programme			0.001
			(0.105)
Number of teachers employed, excluding voluntary/temporary teachers			-0.520
			(0.343)
LGEA visit more than three times in a month			-9.813
			(6.236)
Pupil–teacher ratio			-0.137
			(0.084)
This school has an SBMC			0.371
			(82.102)
Head teacher's age	-0.954	-0.286	-1.226*
	(1.020)	(0.713)	(0.567)
Total teaching experience in ANY school in 2014 (years)	1.517	0.644	1.029
	(1.064)	(0.681)	(0.543)
Head teacher's gender			19.723
			(11.625)
Conducted lesson observations last 10 days			8.088
			(7.311)
One or more formal head teacher–teacher meetings per week			-1.759
			(6.321)
Head teacher has NCE qualification			3.586
			(7.946)
Head teacher took action to improve pupil absenteeism last school year			-31.379*
			(14.916)
Head teacher took action to improve teacher absenteeism last school year			-0.418
			(10.148)
Head teacher attended teaching-related training			17.008*
			(7.458)
Head teacher receives salary on time			-2.180
			(7.863)
Head teacher absent one day or more last five days			-12.864

			(7.741)
Head teacher absent one day or more last term			20.001**
			(6.426)
School got SUBEB-led INSET	46.189**	42.097**	1.982
	(17.960)	(16.331)	(8.941)
School got RANA	-91.790**	-39.265*	23.515
	(42.242)	(18.970)	(13.582)
School got Jolly Phonics	-25.126*	-0.110	-0.152
	(15.026)	(11.050)	(5.781)
School got ESSPIN	41.756***	22.898	-4.501
	(14.361)	(13.622)	(12.789)
Constant	510.193***	480.816***	496.289***
	(34.876)	(24.976)	(90.481)
Observations	3,121	3,121	2,873
R-squared	0.070		0.140
State FE	NO	NO	YES
LGA FE	NO	NO	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.10 Proportion of pupils in bottom band – maths: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	0.080*	0.051	
	(0.036)	(0.034)	
Treatment assignment			0.072*
			(0.033)
Dummy for bottom band at baseline	0.223***	0.222***	0.223***
	(0.028)	(0.028)	(0.028)
Pupil is female	0.057*	0.056*	0.056*
	(0.025)	(0.025)	(0.025)
Speak Hausa at home	-0.150*	-0.153**	-0.142*
	(0.059)	(0.058)	(0.059)
Mean value of household asset index	0.011	0.011	0.010
	(0.016)	(0.016)	(0.015)
Electricity at home	-0.088*	-0.090*	-0.087*
	(0.036)	(0.036)	(0.036)
Average family size	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)
Average no. of rooms in pupil's house	-0.008	-0.008	-0.008
	(0.008)	(0.007)	(0.007)
At least one parent/guardian has completed primary school	0.041	0.042	0.041

	(0.030)	(0.030)	(0.029)
Normally eat something during long break	0.001	0.002	0.002
	(0.031)	(0.031)	(0.031)
Flush toilet at home	-0.083**	-0.082**	-0.083**
	(0.031)	(0.031)	(0.031)
School needs major repair	0.020	0.024	0.027
	(0.042)	(0.043)	(0.042)
School has electricity	0.048	0.044	0.044
	(0.042)	(0.043)	(0.041)
Number of class 1–6 pupils registered	-0.000**	-0.000**	-0.000*
	(0.000)	(0.000)	(0.000)
This school has an SBMC	-0.415***	-0.410***	-0.408***
	(0.070)	(0.071)	(0.070)
School receives support from other organisation/programme	-0.114***	-0.113***	-0.111***
	(0.033)	(0.033)	(0.033)
Number of teachers employed, excluding voluntary/temporary teachers	0.003	0.003	0.002
	(0.002)	(0.002)	(0.002)
LGEA visit more than three times in a month	0.070*	0.074**	0.072*
	(0.028)	(0.028)	(0.028)
Pupil–teacher ratio	0.001*	0.001*	0.001*
	(0.001)	(0.001)	(0.001)
Head teacher's age	0.002	0.002	0.002
	(0.003)	(0.003)	(0.003)
Total teaching experience in ANY school in 2014 (years)	0.002	0.002	0.002
	(0.003)	(0.003)	(0.003)
Head teacher's gender	-0.009	-0.007	-0.006
	(0.083)	(0.085)	(0.083)
Conducted lesson observation last 10 days	-0.032	-0.029	-0.025
	(0.036)	(0.036)	(0.036)
One or more formal head teacher–teacher meetings per week	0.008	0.009	0.007
	(0.029)	(0.029)	(0.029)
Head teacher has NCE qualification	0.024	0.022	0.022
	(0.041)	(0.040)	(0.039)
Head teacher took action to improve pupil absenteeism last school year	0.177*	0.170*	0.143
	(0.076)	(0.075)	(0.075)
Head teacher took action to improve teacher absenteeism last school year	-0.029	-0.027	-0.014
	(0.048)	(0.048)	(0.049)
Head teacher attended teaching-related training	0.047	0.042	0.044
	(0.042)	(0.042)	(0.041)
Head teacher receives salary on time	-0.034	-0.040	-0.034
	(0.046)	(0.047)	(0.046)
Head teacher absent one day or more last five days	-0.026	-0.024	-0.021
	(0.050)	(0.050)	(0.051)
Head teacher absent one day or more last term	-0.024	-0.026	-0.024
	(0.031)	(0.031)	(0.031)
School roof good condition [schools]	0.034	0.038	0.044

	(0.044)	(0.043)	(0.043)
Class inner walls good condition [schools]	-0.137*	-0.129*	-0.136*
	(0.062)	(0.062)	(0.062)
Class outer walls good condition [schools]	0.082	0.074	0.083
	(0.065)	(0.066)	(0.064)
School playground good condition [schools]	-0.015	-0.013	-0.006
	(0.036)	(0.036)	(0.037)
School windows good condition [schools]	-0.007	-0.010	-0.021
	(0.058)	(0.059)	(0.060)
School got SUBEB-led INSET	0.026	0.010	0.034
	(0.043)	(0.042)	(0.044)
School got RANA	-0.000	-0.004	-0.007
	(0.052)	(0.051)	(0.052)
School got Jolly Phonics	-0.037	-0.037	-0.038
	(0.033)	(0.033)	(0.033)
School got ESSPIN	0.139**	0.137**	0.140**
	(0.050)	(0.051)	(0.050)
Constant	0.447*	0.641**	0.627**
	(0.210)	(0.225)	(0.225)
Observations	1,377	1,377	1,377
R-squared	0.172	0.172	0.174
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
*** p<0.001, ** p<0.01, * p<0.05			

F.11 Proportion of pupils in bottom band – maths: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	0.059	0.054	-0.021
	(0.047)	(0.042)	(0.034)
Time	-0.494***	-0.513***	-0.479***
	(0.042)	(0.035)	(0.050)
DID			0.036
			(0.041)
Pupil is female			0.123***
			(0.018)
Speak Hausa at home			0.116
			(0.072)
Mean value of household asset index			-0.031*
			(0.014)
Electricity at home			-0.093***
			(0.025)

Average family size			-0.001
			(0.002)
Average no. of rooms in pupil's house			-0.011
			(0.007)
At least one parent/guardian has completed primary school			-0.003
			(0.022)
Flush toilet at home			-0.035
			(0.028)
School needs major repair			-0.088*
			(0.035)
School has electricity	0.064	0.044	0.010
	(0.074)	(0.049)	(0.033)
Number of class 1–6 pupils registered			-0.000***
			(0.000)
School receives support from other organisation/programme			-0.000
			(0.000)
Number of teachers employed, excluding voluntary/temporary teachers			0.003**
			(0.001)
LGEA visit more than three times in a month			0.046
			(0.028)
Pupil–teacher ratio			0.001***
			(0.000)
This school has an SBMC			0.065
			(0.051)
Head teacher's age	0.001	0.003	0.005*
	(0.003)	(0.003)	(0.002)
Total teaching experience in ANY school in 2014 (years)			-0.002
			(0.002)
Head teacher's gender			0.013
			(0.044)
Conducted lesson observations last 10 days			-0.021
			(0.027)
One or more formal head teacher–teacher meetings per week			-0.012
			(0.026)
Head teacher has NCE qualification			-0.054*
			(0.026)
Head teacher took action to improve pupil absenteeism last school year			0.105
			(0.061)
Head teacher took action to improve teacher absenteeism last school year			0.039
			(0.050)
Head teacher attended teaching-related training			-0.049
			(0.031)
Head teacher receives salary on time			0.014
			(0.043)
Head teacher absent one day or more last five days			-0.039
			(0.026)

Head teacher absent one day or more last term			0.010
			(0.021)
School got SUBEB-led INSET	0.020	0.050	0.032
	(0.044)	(0.051)	(0.037)
School got RANA	0.049	-0.049	-0.045
	(0.145)	(0.068)	(0.067)
School got Jolly Phonics	0.002	-0.013	-0.066**
	(0.049)	(0.038)	(0.025)
School got ESSPIN	-0.109**	-0.087*	0.025
	(0.048)	(0.041)	(0.051)
Constant	0.809***	0.813***	0.630**
	(0.116)	(0.087)	(0.201)
Observations	3,106	3,106	2,861
R-squared	0.454		0.334
State FE	NO	NO	YES
LGA FE	NO	NO	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.12 Proportion of pupils in top band – maths: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	0.024	0.022	
	(0.022)	(0.020)	
Treatment assignment			0.022
			(0.020)
Dummy for top band in maths at baseline	0.222***	0.222***	0.223***
	(0.061)	(0.061)	(0.061)
Pupil is female	-0.001	-0.001	-0.001
	(0.017)	(0.017)	(0.017)
Speak Hausa at home	0.077	0.077	0.080
	(0.052)	(0.052)	(0.051)
Mean value of household asset index	0.006	0.006	0.006
	(0.009)	(0.009)	(0.009)
Electricity at home	0.032	0.032	0.032
	(0.021)	(0.021)	(0.021)
Average family size	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)
Average no. of rooms in pupil's house	0.006	0.006	0.006
	(0.006)	(0.006)	(0.006)
At least one parent/guardian has completed primary school	-0.009	-0.009	-0.009
	(0.026)	(0.026)	(0.026)

Normally eat something during long break	-0.006	-0.006	-0.005
	(0.022)	(0.022)	(0.022)
Flush toilet at home	0.135***	0.135***	0.135***
	(0.028)	(0.028)	(0.028)
School needs major repair	-0.023	-0.023	-0.021
	(0.034)	(0.034)	(0.034)
School has electricity	0.033	0.033	0.032
	(0.032)	(0.032)	(0.032)
Number of class 1–6 pupils registered	0.000*	0.000*	0.000*
	(0.000)	(0.000)	(0.000)
This school has an SBMC	0.071	0.071	0.073
	(0.068)	(0.067)	(0.067)
School receives support from other organisation/programme	-0.052*	-0.052*	-0.051*
	(0.025)	(0.025)	(0.025)
Number of teachers employed, excluding voluntary/temporary teachers	-0.003*	-0.003*	-0.003*
	(0.001)	(0.001)	(0.001)
LGEA visit more than three times in a month	-0.002	-0.002	-0.001
	(0.022)	(0.022)	(0.022)
Pupil teacher ratio	-0.001**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)
Head teacher's age	0.003	0.003	0.003
	(0.002)	(0.002)	(0.002)
Total teaching experience in ANY school in 2014 (years)	-0.002	-0.002	-0.002
	(0.002)	(0.002)	(0.002)
Head teacher's gender	-0.132**	-0.132**	-0.131**
	(0.041)	(0.041)	(0.041)
Conducted lesson observation last 10 days	0.013	0.013	0.015
	(0.021)	(0.021)	(0.021)
One or more formal head teacher–teacher meetings per week	-0.032	-0.032	-0.032
	(0.020)	(0.020)	(0.020)
Head teacher has NCE qualification	-0.091*	-0.091*	-0.092*
	(0.035)	(0.035)	(0.035)
Head teacher took action to improve pupil absenteeism last school year	-0.003	-0.004	-0.014
	(0.063)	(0.063)	(0.064)
Head teacher took action to improve teacher absenteeism last school year	0.015	0.015	0.020
	(0.054)	(0.054)	(0.054)
Head teacher attended teaching-related training	-0.003	-0.004	-0.004
	(0.022)	(0.022)	(0.022)
Head teacher receives salary on time	0.087	0.086	0.087
	(0.045)	(0.045)	(0.045)
Head teacher absent one day or more last five days [head teacher]	0.007	0.007	0.009
	(0.024)	(0.024)	(0.024)
Head teacher absent one day or more last term [head teacher]	0.043*	0.043*	0.043*
	(0.018)	(0.018)	(0.018)
School roof good condition [schools]	-0.052*	-0.051*	-0.049*
	(0.024)	(0.024)	(0.024)

Class inner walls good condition [schools]	-0.038	-0.037	-0.037
	(0.048)	(0.049)	(0.048)
Class outer walls good condition [schools]	0.044	0.043	0.044
	(0.035)	(0.035)	(0.035)
School playground good condition [schools]	-0.005	-0.005	-0.002
	(0.020)	(0.020)	(0.020)
School windows good condition [schools]	0.032	0.031	0.028
	(0.031)	(0.031)	(0.031)
School got SUBEB-led INSET	-0.120***	-0.121***	-0.117***
	(0.031)	(0.030)	(0.032)
School got RANA	-0.101*	-0.101*	-0.103*
	(0.049)	(0.049)	(0.049)
School got Jolly Phonics	-0.028	-0.028	-0.028
	(0.024)	(0.024)	(0.024)
School got ESSPIN	-0.016	-0.016	-0.015
	(0.051)	(0.051)	(0.051)
Constant	0.023	-0.080	-0.079
	(0.159)	(0.159)	(0.159)
Observations	1,377	1,377	1,377
R-squared	0.154	0.154	0.154
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.13 Proportion of pupils in top band – maths: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	0.033	0.009	-0.017
	(0.034)	(0.029)	(0.021)
Time	0.049	0.081**	0.047
	(0.032)	(0.026)	(0.044)
DID			0.058
			(0.033)
Pupil is female			-0.018
			(0.013)
Speak Hausa at home			0.014
			(0.032)
Mean value of household asset index			0.014*
			(0.007)
Electricity at home			0.012
			(0.019)
Average family size			-0.001

			(0.001)
Average no. of rooms in pupil's house			0.013*
			(0.005)
At least one parent/guardian has completed primary school		-0.013	
			(0.015)
Flush toilet at home			0.060*
			(0.024)
School needs major repair			0.031
			(0.019)
School has electricity	-0.021	-0.048	0.045*
	(0.051)	(0.052)	(0.018)
Number of class 1–6 pupils registered			0.000
			(0.000)
School receives support from other organisation/programme			-0.000
			(0.000)
Number of teachers employed, excluding voluntary/temporary teachers		-0.002	
			(0.001)
LGEA visit more than three times in a month			-0.008
			(0.013)
Pupil–teacher ratio			-0.000
			(0.000)
This school has an SBMC			-0.027
			(0.032)
Head teacher's age	0.004	0.001	0.001
	(0.003)	(0.002)	(0.002)
Total teaching experience in ANY school in 2014 (years)	0.001	0.001	-0.000
	(0.003)	(0.002)	(0.002)
Head teacher's gender			0.047
			(0.029)
Conducted lesson observations last 10 days			0.023
			(0.017)
One or more formal head teacher–teacher meetings per week			-0.003
			(0.015)
Head teacher has NCE qualification			0.032
			(0.019)
Head teacher took action to improve pupil absenteeism last school year		-0.105***	
			(0.030)
Head teacher took action to improve teacher absenteeism last school year		0.032	
			(0.020)
Head teacher attended teaching-related training			0.028
			(0.016)
Head teacher receives salary on time			-0.039
			(0.026)

Head teacher absent one day or more last five days			0.041
			(0.023)
Head teacher absent one day or more last term			0.024
			(0.016)
School got SUBEB-led INSET	-0.012	-0.034	-0.058*
	(0.032)	(0.032)	(0.023)
School got RANA	-0.251	-0.076	0.033
	(0.166)	(0.047)	(0.050)
School got Jolly Phonics	-0.045	-0.005	0.018
	(0.032)	(0.027)	(0.016)
School got ESSPIN	0.125***	0.052	0.006
	(0.033)	(0.028)	(0.027)
Constant	-0.128	-0.031	-0.078
	(0.087)	(0.065)	(0.121)
Observations	3,106	3,106	2,861
R-squared	0.078		0.087
State FE	NO	YES	YES
LGA FE	NO	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.14 Proportion of pupils in bottom band – English: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	-0.006	-0.007	
	(0.021)	(0.018)	
Treatment assignment			-0.006
			(0.019)
Dummy for bottom band in English at baseline	0.094***	0.094***	0.094***
	(0.018)	(0.018)	(0.018)
Pupil is female	0.017	0.017	0.017
	(0.016)	(0.016)	(0.016)
Speak Hausa at home	-0.235*	-0.235*	-0.236*
	(0.112)	(0.112)	(0.112)
Mean value of household asset index	0.015	0.015	0.015
	(0.009)	(0.009)	(0.009)
Electricity at home	-0.086**	-0.086**	-0.086**
	(0.026)	(0.026)	(0.026)
Average family size	-0.000	-0.000	-0.000
	(0.001)	(0.001)	(0.001)
Average no. of rooms in pupil's house	0.007	0.007	0.007
	(0.005)	(0.005)	(0.005)

At least one parent/guardian has completed primary school	-0.022	-0.022	-0.022
	(0.023)	(0.023)	(0.023)
Normally eat something during long break	-0.001	-0.001	-0.001
	(0.022)	(0.022)	(0.022)
Flush toilet at home	-0.036	-0.036	-0.036
	(0.019)	(0.019)	(0.019)
School needs major repair	-0.035	-0.035	-0.036
	(0.024)	(0.024)	(0.024)
School has electricity	0.075**	0.075**	0.076***
	(0.023)	(0.023)	(0.023)
Number of class 1–6 pupils registered	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
This school has an SBMC	-0.332***	-0.332***	-0.332***
	(0.054)	(0.054)	(0.055)
School receives support from other organisation/programme	-0.010	-0.010	-0.011
	(0.020)	(0.020)	(0.020)
Number of teachers employed, excluding voluntary/temporary teachers	-0.004***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)
LGEA visit more than three times in a month	-0.030*	-0.030*	-0.030*
	(0.014)	(0.014)	(0.014)
Pupil–teacher ratio	-0.001*	-0.001*	-0.001
	(0.000)	(0.000)	(0.000)
Head teacher's age	0.000	0.000	0.000
	(0.002)	(0.002)	(0.002)
Total teaching experience in ANY school in 2014 (years)	0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)
Head teacher's gender	0.040	0.040	0.040
	(0.040)	(0.040)	(0.040)
Conducted lesson observation last 10 days	0.054**	0.054**	0.053**
	(0.018)	(0.018)	(0.018)
One or more formal head teacher–teacher meetings per week	0.023	0.023	0.023
	(0.018)	(0.018)	(0.018)
Head teacher has NCE qualification	0.022	0.022	0.022
	(0.022)	(0.022)	(0.022)
Head teacher took action to improve pupil absenteeism last school year	0.083	0.083	0.085
	(0.068)	(0.068)	(0.068)
Head teacher took action to improve teacher absenteeism last school year	-0.136**	-0.136**	-0.137**
	(0.049)	(0.049)	(0.049)
Head teacher attended teaching-related training	0.029	0.029	0.030
	(0.019)	(0.019)	(0.019)
Head teacher receives salary on time	-0.057	-0.058	-0.057
	(0.032)	(0.032)	(0.032)
Head teacher absent one day or more last five days [head teacher]	0.004	0.004	0.004
	(0.025)	(0.025)	(0.025)
Head teacher absent one day or more last term [head teacher]	0.052***	0.052***	0.052***
	(0.015)	(0.015)	(0.015)

School roof good condition [schools]	-0.005	-0.005	-0.006
	(0.023)	(0.022)	(0.022)
Class inner walls good condition [schools]	-0.052*	-0.052*	-0.052*
	(0.026)	(0.025)	(0.026)
Class outer walls good condition [schools]	0.037	0.037	0.037
	(0.021)	(0.021)	(0.021)
School playground good condition [schools]	0.010	0.010	0.010
	(0.018)	(0.018)	(0.018)
School windows good condition [schools]	0.035	0.035	0.036
	(0.019)	(0.019)	(0.019)
School got SUBEB-led INSET	-0.029	-0.029	-0.030
	(0.026)	(0.025)	(0.027)
School got RANA	-0.002	-0.002	-0.001
	(0.041)	(0.041)	(0.040)
School got Jolly Phonics	0.001	0.001	0.001
	(0.017)	(0.017)	(0.017)
School got ESSPIN	0.022	0.022	0.022
	(0.034)	(0.034)	(0.034)
Constant	0.649***	0.674***	0.673***
	(0.159)	(0.161)	(0.161)
Observations	1,377	1,377	1,377
R-squared	0.133	0.133	0.133
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.15 Proportion of pupils in bottom band – English: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	-0.063	-0.040	0.030
	(0.050)	(0.049)	(0.036)
Time	-0.457***	-0.542***	-0.493***
	(0.040)	(0.037)	(0.040)
DID			-0.062
			(0.040)
Pupil is female			0.034
			(0.020)
Speak Hausa at home			-0.046
			(0.075)
Mean value of household asset index			-0.031**
			(0.010)
Electricity at home			-0.094***

			(0.020)
Average family size			-0.002
			(0.001)
Average no. of rooms in pupil's house			-0.005
			(0.005)
At least one parent/guardian has completed primary school		-0.031	
			(0.022)
Flush toilet at home			-0.017
			(0.026)
School needs major repair			-0.067*
			(0.031)
School has electricity	-0.030	-0.030	-0.028
	(0.074)	(0.064)	(0.032)
Number of class 1–6 pupils registered			-0.000*
			(0.000)
School receives support from other organisation/programme			0.000
			(0.000)
Number of teachers employed, excluding voluntary/temporary teachers		0.001	
			(0.001)
LGEA visit more than three times in a month			0.018
			(0.027)
Pupil–teacher ratio			0.000
			(0.000)
This school has an SBMC			0.161
			(0.139)
Head teacher's age	0.004	0.003	0.005*
	(0.004)	(0.003)	(0.002)
Total teaching experience in ANY school in 2014 (years)	-0.002	-0.003	-0.003
	(0.003)	(0.003)	(0.002)
Head teacher's gender			0.026
			(0.056)
Conducted lesson observations last 10 days			-0.046*
			(0.022)
One or more formal head teacher–teacher meetings per week			-0.009
			(0.029)
Head teacher has NCE qualification			-0.032
			(0.028)
Head teacher took action to improve pupil absenteeism last school year		0.144*	
			(0.066)
Head teacher took action to improve teacher absenteeism last school year		-0.072	
			(0.051)
Head teacher attended teaching-related training			-0.037
			(0.033)

Head teacher receives salary on time			0.036
			(0.048)
Head teacher absent one day or more last five days			-0.005
			(0.027)
Head teacher absent one day or more last term			0.020
			(0.019)
School got SUBEB-led INSET	-0.159***	-0.085	0.039
	(0.059)	(0.052)	(0.032)
School got RANA	0.108	-0.064	-0.009
	(0.136)	(0.090)	(0.044)
School got Jolly Phonics	0.071	0.070	-0.000
	(0.048)	(0.038)	(0.023)
School got ESSPIN	-0.176***	-0.176***	-0.071
	(0.054)	(0.041)	(0.045)
Constant	0.507***	0.656***	0.649**
	(0.134)	(0.118)	(0.226)
Observations	3,108	3,108	2,862
R-squared	0.510		0.371
State FE	NO	NO	YES
LGA FE	NO	NO	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.16 Proportion of pupils in top band – English: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	0.016	0.023	
	(0.017)	(0.014)	
Treatment assignment			0.014
			(0.015)
Dummy for top band in English at baseline	0.034	0.033	0.035
	(0.068)	(0.067)	(0.069)
Pupil is female	-0.028*	-0.028*	-0.028*
	(0.013)	(0.013)	(0.013)
Speak Hausa at home	-0.069	-0.069	-0.068
	(0.063)	(0.062)	(0.063)
Mean value of household asset index	0.010	0.010	0.010
	(0.007)	(0.007)	(0.007)
Electricity at home	-0.015	-0.015	-0.015
	(0.020)	(0.019)	(0.020)
Average family size	-0.004**	-0.004**	-0.004**

	(0.001)	(0.001)	(0.001)
Average no. of rooms in pupil's house	0.011*	0.011*	0.011*
	(0.006)	(0.006)	(0.006)
At least one parent/guardian has completed primary school	0.009	0.009	0.009
	(0.017)	(0.017)	(0.017)
Normally eat something during long break	0.037*	0.036*	0.037*
	(0.015)	(0.015)	(0.015)
Flush toilet at home	0.064**	0.064**	0.064**
	(0.021)	(0.021)	(0.021)
School needs major repair	-0.016	-0.017	-0.014
	(0.025)	(0.025)	(0.025)
School has electricity	0.010	0.011	0.009
	(0.023)	(0.024)	(0.024)
Number of class 1–6 pupils registered	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
This school has an SBMC	0.073	0.073	0.074
	(0.042)	(0.043)	(0.041)
School receives support from other organisation/programme	0.011	0.011	0.012
	(0.013)	(0.013)	(0.013)
Number of teachers employed, excluding voluntary/temporary teachers	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)
LGEA visit more than three times in a month	0.005	0.004	0.005
	(0.015)	(0.015)	(0.015)
Pupil–teacher ratio	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Head teacher's age	0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)
Total teaching experience in ANY school in 2014 (years)	-0.002	-0.001	-0.002
	(0.001)	(0.001)	(0.001)
Head teacher's gender	-0.056*	-0.057*	-0.056*
	(0.026)	(0.026)	(0.026)
Conducted lesson observation last 10 days	0.024	0.023	0.025
	(0.015)	(0.015)	(0.015)
One or more formal head teacher–teacher meetings per week	0.012	0.012	0.011
	(0.015)	(0.015)	(0.015)
Head teacher has NCE qualification	-0.048	-0.048	-0.048
	(0.025)	(0.025)	(0.025)
Head teacher took action to improve pupil absenteeism last school year	0.062	0.064	0.055
	(0.047)	(0.048)	(0.048)
Head teacher took action to improve teacher absenteeism last school year	0.024	0.023	0.027
	(0.029)	(0.029)	(0.030)
Head teacher attended teaching-related training	-0.008	-0.007	-0.009
	(0.015)	(0.015)	(0.015)
Head teacher receives salary on time	0.114***	0.115***	0.114***
	(0.031)	(0.031)	(0.031)
Head teacher absent one day or more last five days [head teacher]	0.017	0.017	0.019

	(0.019)	(0.019)	(0.019)
Head teacher absent one day or more last term [head teacher]	0.029*	0.030*	0.029*
	(0.015)	(0.015)	(0.015)
School roof good condition [schools]	-0.009	-0.010	-0.007
	(0.018)	(0.018)	(0.019)
Class inner walls good condition [schools]	-0.022	-0.024	-0.022
	(0.037)	(0.036)	(0.037)
Class outer walls good condition [schools]	0.042	0.044	0.042
	(0.037)	(0.037)	(0.037)
School playground good condition [schools]	-0.010	-0.010	-0.008
	(0.020)	(0.020)	(0.020)
School windows good condition [schools]	0.003	0.004	0.000
	(0.023)	(0.023)	(0.024)
School got SUBEB-led INSET	-0.040	-0.035	-0.038
	(0.023)	(0.022)	(0.024)
School got RANA	-0.037	-0.036	-0.038
	(0.035)	(0.035)	(0.035)
School got Jolly Phonics	-0.019	-0.019	-0.019
	(0.017)	(0.017)	(0.018)
School got ESSPIN	0.079*	0.079*	0.079*
	(0.032)	(0.032)	(0.032)
Constant	-0.066	-0.042	-0.032
	(0.131)	(0.134)	(0.134)
Observations	1,377	1,377	1,377
R-squared	0.109	0.109	0.109
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.17 Proportion of pupils in top band – English: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	0.003	0.002	0.004
	(0.027)	(0.021)	(0.015)
Time	0.029	0.033	0.005
	(0.027)	(0.019)	(0.020)
DID			0.011
			(0.021)
Pupil is female			-0.022*
			(0.010)
Speak Hausa at home			-0.022
			(0.027)

Mean value of household asset index			0.013*
			(0.005)
Electricity at home			-0.006
			(0.012)
Average family size			-0.003**
			(0.001)
Average no. of rooms in pupil's house			0.013**
			(0.004)
At least one parent/guardian has completed primary school		0.008	
			(0.012)
Flush toilet at home			0.029*
			(0.014)
School needs major repair			0.017
			(0.016)
School has electricity	-0.055	-0.070	0.011
	(0.052)	(0.037)	(0.013)
Number of class 1–6 pupils registered			0.000*
			(0.000)
School receives support from other organisation/programme			0.000
			(0.000)
Number of teachers employed, excluding voluntary/temporary teachers		-0.002**	
			(0.001)
LGEA visit more than three times in a month			-0.008
			(0.008)
Pupil–teacher ratio			-0.000**
			(0.000)
This school has an SBMC			0.019
			(0.024)
Head teacher's age	0.002	-0.001	0.001
	(0.002)	(0.001)	(0.001)
Total teaching experience in ANY school in 2014 (years)	0.000	-0.000	-0.001
	(0.002)	(0.001)	(0.001)
Head teacher's gender			0.021
			(0.029)
Conducted lesson observations last 10 days			0.023
			(0.012)
One or more formal head teacher–teacher meetings per week			0.008
			(0.013)
Head teacher has NCE qualification			0.014
			(0.013)
Head teacher took action to improve pupil absenteeism last school year		-0.051*	
			(0.025)
Head teacher took action to improve teacher absenteeism last school year		-0.001	

			(0.014)
Head teacher attended teaching-related training			0.009
			(0.013)
Head teacher receives salary on time			-0.035
			(0.024)
Head teacher absent one day or more last five days			0.013
			(0.018)
Head teacher absent one day or more last term			0.021*
			(0.008)
School got SUBEB-led INSET	-0.012	0.008	-0.003
	(0.023)	(0.028)	(0.022)
School got RANA	-0.065	-0.006	-0.015
	(0.071)	(0.028)	(0.025)
School got Jolly Phonics	-0.031	0.009	-0.003
	(0.025)	(0.021)	(0.011)
School got ESSPIN	0.078**	0.017	0.057**
	(0.030)	(0.024)	(0.019)
Constant	-0.043	0.053	-0.052
	(0.064)	(0.044)	(0.079)
Observations	3,108	3,108	2,862
R-squared	0.033		0.071
State FE	NO	YES	YES
LGA FE	NO	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.18 Proportion of pupils in bottom band – science and technology: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	0.005*	0.003	
	(0.002)	(0.002)	
Treatment assignment			0.005*
			(0.002)
Dummy for bottom band in science at baseline	0.006	0.006	0.006
	(0.003)	(0.003)	(0.003)
Pupil is female	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
Speak Hausa at home	-0.006	-0.007	-0.006
	(0.008)	(0.008)	(0.008)
Mean value of household asset index	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
Electricity at home	0.003	0.003	0.003

	(0.002)	(0.002)	(0.002)
Average family size	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Average no. of rooms in pupil's house	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
At least one parent/guardian has completed primary school	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)
Normally eat something during long break	0.003	0.003	0.003
	(0.002)	(0.002)	(0.002)
Flush toilet at home	-0.002	-0.002	-0.002
	(0.002)	(0.002)	(0.002)
School needs major repair	0.004	0.005	0.005
	(0.002)	(0.003)	(0.003)
School has electricity	0.003	0.002	0.002
	(0.002)	(0.002)	(0.002)
Number of class 1–6 pupils registered	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
This school has an SBMC	-0.073	-0.073	-0.073
	(0.049)	(0.051)	(0.049)
School receives support from other organisation/programme	-0.004	-0.003	-0.003
	(0.002)	(0.002)	(0.002)
Number of teachers employed, excluding voluntary/temporary teachers	-0.000*	-0.000	-0.000*
	(0.000)	(0.000)	(0.000)
LGEA visit more than three times in a month	0.000	0.000	0.000
	(0.002)	(0.001)	(0.002)
Pupil–teacher ratio	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Head teacher's age	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Total teaching experience in ANY school in 2014 (years)	0.000*	0.000	0.000
	(0.000)	(0.000)	(0.000)
Head teacher's gender	-0.002	-0.002	-0.002
	(0.002)	(0.002)	(0.002)
Conducted lesson observation last 10 days	-0.002	-0.002	-0.002
	(0.003)	(0.003)	(0.003)
One or more formal head teacher–teacher meetings per week	0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)
Head teacher has NCE qualification	-0.000	-0.001	-0.000
	(0.002)	(0.002)	(0.002)
Head teacher took action to improve pupil absenteeism last school year	-0.001	-0.002	-0.003
	(0.005)	(0.005)	(0.004)
Head teacher took action to improve teacher absenteeism last school year	0.003	0.003	0.004
	(0.003)	(0.003)	(0.003)
Head teacher attended teaching-related training	0.003	0.002	0.003
	(0.003)	(0.003)	(0.003)
Head teacher receives salary on time	0.005	0.004	0.005

	(0.005)	(0.005)	(0.005)
Head teacher absent one day or more last five days [head teacher]	0.001	0.001	0.002
	(0.003)	(0.003)	(0.003)
Head teacher absent one day or more last term [head teacher]	0.004	0.004	0.004
	(0.002)	(0.002)	(0.002)
School roof good condition [schools]	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)
Class inner walls good condition [schools]	-0.002	-0.001	-0.002
	(0.003)	(0.002)	(0.002)
Class outer walls good condition [schools]	-0.002	-0.003	-0.002
	(0.002)	(0.002)	(0.002)
School playground good condition [schools]	-0.000	-0.000	0.000
	(0.003)	(0.003)	(0.002)
School windows good condition [schools]	-0.001	-0.001	-0.002
	(0.003)	(0.003)	(0.002)
School got SUBEB-led INSET	0.010*	0.009	0.011*
	(0.005)	(0.005)	(0.005)
School got RANA	-0.005	-0.005	-0.005
	(0.003)	(0.003)	(0.003)
School got Jolly Phonics	-0.005	-0.005	-0.005
	(0.003)	(0.003)	(0.003)
School got ESSPIN	-0.003	-0.004	-0.003
	(0.002)	(0.003)	(0.002)
Constant	0.095	0.082	0.079
	(0.051)	(0.052)	(0.051)
Observations	1,382	1,382	1,382
R-squared	0.028	0.029	0.030
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			

F.19 Proportion of pupils in bottom band – science and technology: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	-0.007	-0.036	-0.000
	(0.032)	(0.037)	(0.022)
Time	-0.153***	-0.156***	-0.159***
	(0.030)	(0.028)	(0.021)
DID			-0.006
			(0.025)
Pupil is female			0.019

			(0.012)
Speak Hausa at home			0.026
			(0.042)
Mean value of household asset index			-0.013
			(0.007)
Electricity at home			-0.014
			(0.013)
Average family size			-0.000
			(0.001)
Average no. of rooms in pupil's house			-0.005*
			(0.002)
At least one parent/guardian has completed primary school			-0.041*
			(0.017)
Flush toilet at home			0.016
			(0.016)
School needs major repair			-0.015
			(0.019)
School has electricity	-0.013	-0.014	0.036*
	(0.037)	(0.049)	(0.015)
Number of class 1–6 pupils registered			-0.000
			(0.000)
School receives support from other organisation/programme			-0.000
			(0.000)
Number of teachers employed, excluding voluntary/temporary teachers			-0.000
			(0.001)
LGEA visit more than three times in a month			-0.021
			(0.014)
Pupil–teacher ratio			-0.000
			(0.000)
This school has an SBMC			-0.140
			(0.298)
Head teacher's age	0.000	-0.002	0.001
	(0.003)	(0.003)	(0.001)
Total teaching experience in ANY school in 2014 (years)	-0.002	-0.002	-0.002
	(0.003)	(0.002)	(0.001)
Head teacher's gender			0.025
			(0.028)
Conducted lesson observations last 10 days			-0.019
			(0.014)
One or more formal head teacher–teacher meetings per week			0.035**
			(0.013)
Head teacher has NCE qualification			-0.037*
			(0.018)
Head teacher took action to improve pupil absenteeism last school year			-0.039
			(0.037)
Head teacher took action to improve teacher absenteeism last school year			0.084*

			(0.037)
Head teacher attended teaching-relate training			-0.011
			(0.019)
Head teacher receives salary on time			0.050*
			(0.023)
Head teacher absent one day or more last five days			0.076***
			(0.022)
Head teacher absent one day or more last term			-0.014
			(0.013)
School got SUBEB-led INSET	0.008	-0.065	-0.014
	(0.044)	(0.052)	(0.019)
School got RANA	0.087**	0.099**	-0.066**
	(0.035)	(0.036)	(0.025)
School got Jolly Phonics	0.033	0.003	-0.012
	(0.031)	(0.035)	(0.013)
School got ESSPIN	-0.200***	-0.172***	0.035
	(0.041)	(0.045)	(0.028)
Constant	0.212**	0.354***	0.442
	(0.094)	(0.082)	(0.312)
Observations	3,122	3,122	2,875
R-squared	0.209		0.170
State FE	NO	NO	YES
LGA FE	NO	NO	YES
Standard errors in parentheses. Survey weights included			
*** p<0.001, ** p<0.01, * p<0.05			

F.20 Proportion of pupils in top band – science and technology: IV and OLS results

	(1)	(2)	(3)
Estimation technique	**IV**	**OLS**	**OLS**
Estimate	**LATE**	**ATET**	**ITT**
Treatment receipt	-0.030	0.004	
	(0.027)	(0.025)	
Treatment assignment			-0.027
			(0.025)
Dummy for top band in science at baseline	0.045	0.044	0.044
	(0.033)	(0.033)	(0.033)
Pupil is female	-0.051*	-0.050*	-0.051*
	(0.021)	(0.021)	(0.021)
Speak Hausa at home	-0.021	-0.017	-0.024
	(0.072)	(0.072)	(0.072)
Mean value of household asset index	0.013	0.013	0.013
	(0.012)	(0.012)	(0.012)

Electricity at home	0.012	0.015	0.012
	(0.027)	(0.027)	(0.027)
Average family size	0.000	0.000	-0.000
	(0.001)	(0.001)	(0.001)
Average no. of rooms in pupil's house	-0.001	-0.001	-0.001
	(0.007)	(0.007)	(0.007)
At least one parent/guardian has completed primary school	-0.038	-0.038	-0.038
	(0.027)	(0.027)	(0.027)
Normally eat something during long break	0.036	0.034	0.035
	(0.027)	(0.027)	(0.027)
Flush toilet at home	0.145***	0.144***	0.145***
	(0.030)	(0.030)	(0.030)
School needs major repair	-0.045	-0.050	-0.048
	(0.032)	(0.033)	(0.033)
School has electricity	-0.068	-0.064	-0.067
	(0.036)	(0.037)	(0.036)
Number of class 1–6 pupils registered	0.000**	0.000**	0.000*
	(0.000)	(0.000)	(0.000)
This school has an SBMC	0.148*	0.148	0.146*
	(0.071)	(0.084)	(0.071)
School receives support from other organisation/programme	-0.004	-0.005	-0.005
	(0.025)	(0.025)	(0.024)
Number of teachers employed, excluding voluntary/temporary teachers	-0.002	-0.002	-0.002
	(0.002)	(0.002)	(0.002)
LGEA visit more than three times in a month	0.024	0.020	0.023
	(0.022)	(0.022)	(0.022)
Pupil–teacher ratio	-0.001**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)
Head teacher's age	0.002	0.002	0.002
	(0.002)	(0.002)	(0.002)
Total teaching experience in ANY school in 2014 (years)	-0.005*	-0.004*	-0.005*
	(0.002)	(0.002)	(0.002)
Head teacher's gender	-0.127*	-0.129*	-0.128*
	(0.051)	(0.053)	(0.051)
Conducted lesson observation last 10 days	0.041	0.038	0.039
	(0.026)	(0.026)	(0.026)
One or more formal head teacher–teacher meetings per week	-0.025	-0.026	-0.025
	(0.024)	(0.024)	(0.023)
Head teacher has NCE qualification	-0.088*	-0.086*	-0.088*
	(0.038)	(0.037)	(0.037)
Head teacher took action to improve pupil absenteeism last school year	-0.043	-0.034	-0.031
	(0.088)	(0.088)	(0.088)
Head teacher took action to improve teacher absenteeism last school year	0.038	0.035	0.032
	(0.064)	(0.064)	(0.064)
Head teacher attended teaching-related training	-0.018	-0.012	-0.017
	(0.029)	(0.029)	(0.029)

Head teacher receives salary on time	-0.034	-0.027	-0.034
	(0.045)	(0.045)	(0.045)
Head teacher absent one day or more last five days [head teacher]	0.049	0.047	0.047
	(0.033)	(0.032)	(0.033)
Head teacher absent one day or more last term [head teacher]	0.053*	0.056*	0.053*
	(0.023)	(0.023)	(0.023)
School roof good condition [schools]	-0.034	-0.039	-0.038
	(0.032)	(0.031)	(0.031)
Class inner walls good condition [schools]	-0.011	-0.020	-0.011
	(0.050)	(0.050)	(0.050)
Class outer walls good condition [schools]	-0.003	0.006	-0.003
	(0.069)	(0.069)	(0.069)
School playground good condition [schools]	0.052	0.050	0.049
	(0.028)	(0.028)	(0.029)
School windows good condition [schools]	0.067*	0.070*	0.072*
	(0.032)	(0.032)	(0.033)
School got SUBEB-led INSET	-0.006	0.013	-0.009
	(0.042)	(0.042)	(0.042)
School got RANA	-0.069	-0.065	-0.067
	(0.049)	(0.049)	(0.049)
School got Jolly Phonics	-0.083**	-0.083**	-0.083**
	(0.030)	(0.030)	(0.029)
School got ESSPIN	0.051	0.052	0.050
	(0.070)	(0.070)	(0.070)
Constant	0.263	0.002	0.043
	(0.178)	(0.208)	(0.203)
Observations	1,382	1,382	1,382
R-squared	0.191	0.193	0.194
State FE	YES	YES	YES
LGA FE	YES	YES	YES
Standard errors in parentheses. Survey weights included			
*** p<0.001, ** p<0.01, * p<0.05			

F.21 Proportion of pupils in top band – science and technology: panel and DID results

	(1)	(2)	(3)
Estimation technique	**Panel FE**	**Panel IV FE**	**DID**
Estimate	**ATET**	**LATE**	**ATET**
Treatment receipt	0.071	0.038	-0.032
	(0.059)	(0.042)	(0.034)
Time	-0.040	0.003	0.027
	(0.051)	(0.036)	(0.056)
DID			0.043
			(0.049)

Pupil is female			-0.054***
			(0.015)
Speak Hausa at home			0.046
			(0.067)
Mean value of household asset index			-0.014
			(0.010)
Electricity at home			0.027
			(0.021)
Average family size			0.001
			(0.001)
Average no. of rooms in pupil's house			0.007
			(0.005)
At least one parent/guardian has completed primary school		0.003	
			(0.015)
Flush toilet at home			0.088***
			(0.023)
School needs major repair			0.068**
			(0.024)
School has electricity	-0.009	-0.085	-0.037
	(0.098)	(0.076)	(0.035)
Number of class 1–6 pupils registered			0.000***
			(0.000)
School receives support from other organisation/programme			-0.000
			(0.000)
Number of teachers employed, excluding voluntary/temporary teachers		-0.003**	
			(0.001)
LGEA visit more than three times in a month			-0.016
			(0.024)
Pupil–teacher ratio			-0.001*
			(0.000)
This school has an SBMC			-0.209
			(0.150)
Head teacher's age	-0.003	-0.002	-0.004
	(0.004)	(0.003)	(0.002)
Total teaching experience in ANY school in 2014 (years)	0.005	0.002	0.003
	(0.004)	(0.003)	(0.002)
Head teacher's gender			0.069
			(0.044)
Conducted lesson observations last 10 days			0.070**
			(0.022)
One or more formal head teacher–teacher meetings per week			0.008
			(0.022)
Head teacher has NCE qualification			0.015
			(0.033)

Head teacher took action to improve pupil absenteeism last school year		-0.138*	
			(0.057)
Head teacher took action to improve teacher absenteeism last school year	0.044		
			(0.040)
Head teacher attended teaching-related training			0.029
			(0.027)
Head teacher receives salary on time			0.002
			(0.030)
Head teacher absent one day or more last five days			-0.017
			(0.023)
Head teacher absent one day or more last term			0.068**
			(0.024)
School got SUBEB-led INSET	0.163**	0.113*	0.008
	(0.075)	(0.055)	(0.029)
School got RANA	-0.216	-0.052	0.041
	(0.155)	(0.076)	(0.037)
School got Jolly Phonics	-0.052	0.012	-0.034
	(0.058)	(0.039)	(0.021)
School got ESSPIN	0.135**	0.063	0.028
	(0.054)	(0.044)	(0.043)
Constant	0.175	0.153	0.159
	(0.152)	(0.092)	(0.229)
Observations	3,122	3,122	2,875
R-squared	0.045		0.119
State FE	NO	NO	YES
LGA FE	NO	NO	YES
Standard errors in parentheses. Survey weights included			
**** p<0.001, ** p<0.01, * p<0.05*			