

EQUIP-Tanzania Impact Evaluation

Endline Quantitative Technical Report, Volume II

Methods and Supplementary Evidence



FINAL REPORT

Georgina Rawle, Michele Binci, Gunilla Pettersson Gelandner, Jana Harb, Paul Jasper, Safa Khan, Deo Medardi, Alessio Romarri, Michelle Rorich, and Nicola Ruddle.

25 January 2019

Acknowledgements

Please consult Volume 1 of this report to see a list of acknowledgments

Table of contents

Acknowledgements	i
List of tables and figures	v
List of abbreviations	viii
1 Introduction to Volume II	1
1.1 Overview	1
1.2 Structure of this Volume	1
Part D: Methods	3
2 Mixed methods approach	4
2.1 Impact evaluation objectives	4
2.2 Impact evaluation methods	4
3 Quantitative impact evaluation design and endline adjustments	6
3.1 Rationale for quasi-experimental design	6
3.2 Sampling strategy, sample size and instruments	6
3.2.1 Sampling strategy	6
3.2.2 Sample size	8
3.2.3 Survey weights	9
3.2.4 Survey instruments	9
3.3 Fieldwork timing and model	12
3.4 Quantitative analysis	12
3.4.1 Impact estimation	12
3.4.2 Descriptive analysis of change in programme schools	13
3.4.3 Quality assurance	13
3.5 Risks and limitations	14
3.5.1 Contamination risk from other programmes	14
3.5.2 Confounding theory of change failure with implementation failure	17
3.5.3 Limitations to the quantitative component	17
Part E Supplementary evidence	20
4 Impact estimates	21
4.1 Impact identification strategy	21
4.2 Combining DID and PSM	21
4.3 How results are presented in Volume I	23
4.4 Caveats - Addressing weaknesses in the analysis	25
4.5 Results	26
Presentation of results	27
Proportion of pupils in the top performance band for Kiswahili	29
Proportion of pupils in the bottom performance band for Kiswahili	32
Kiswahili Rasch Scale	35
Proportion of pupils in the top performance band for Mathematics	38
Proportion of pupils in the bottom performance band for Mathematics	41
Mathematics Rasch Scale	44
Teacher school absenteeism	47
Teacher classroom absenteeism	50
Proportion of teachers who report participation in performance appraisal	53
4.6 Comment on effect size	55

5	Supplementary descriptive trends in programme areas	57
5.1	Pupil learning and background characteristics	57
5.1.1	Pupil Kiswahili raw test score indicators	57
5.1.2	Pupil maths raw test score indicators	59
5.1.3	Pupils' background characteristics, including disability	60
5.1.4	Pupil Kiswahili scale scores for different subpopulations	62
5.1.5	Pupil maths scale scores for different subpopulations	62
5.1.6	Trends in absolute numbers of pupils achieving at different performance bands in 17 programme districts	63
5.2	Teacher performance	66
5.2.1	EQUIP-T school-based in-service teacher training	66
5.2.2	Profile of INCO	69
5.2.3	Difficulties with EQUIP-T training	70
5.2.4	Teacher access to curriculum, syllabi and teacher guides	72
5.2.5	Early grade teacher background characteristics	73
5.2.6	Teacher performance indicators by gender and age	74
5.3	SLM	76
5.3.1	Teacher management: teacher absence	76
5.3.2	Teacher management: EQUIP-T impact on teachers receiving performance appraisals	77
5.4	Turnover in education posts at the school and ward level	78
5.4.1	Teacher turnover	79
5.4.2	INCO turnover	83
5.4.3	Head teacher turnover	83
5.4.4	WEO turnover	85
	References	86
Annex A	Impact evaluation districts	91
Annex B	Stakeholder engagement and impact evaluation governance	93
B.1	Stakeholder engagement	93
B.2	Reference Group	95
B.3	Impact evaluation governance and quality assurance	96
Annex C	Ethical considerations	98
C.1	Ethical protocols at endline	98
C.2	Principles	98
C.2.1	Respect for persons	98
C.2.2	Beneficence	99
C.2.3	Justice	99
Annex D	Quantitative survey fieldwork	100
D.1	Personnel	100
D.2	Fieldwork preparation	100
D.2.1	Pre-test	101
D.2.2	Permits and reporting	101
D.2.3	Fieldwork manual	101
D.3	Training and pilot	102
D.4	Fieldwork organisation	102
D.4.1	Fieldwork plan	102
D.4.2	Fieldwork model	102
D.5	Fieldwork implementation	103
D.5.1	Replacements	103
D.5.2	Response rates per instrument	103
D.6	Quality control and data checking protocols	103

D.6.1	Selection and supervision of enumerators	104
D.6.2	CAPI built-in routing and validations	104
D.6.3	Secondary consistency checks and cleaning in Stata	104
D.6.4	Monitoring fieldwork progress and performance indicators	104
D.6.5	Field visits by fieldwork management team including back-checking of data	105
D.6.6	Integration of Analysis and Survey Team	105
D.7	Fieldwork challenges	105
Annex E	Measurement of pupil learning outcomes	107
E.1	Summary of the content of the pupil tests	107
E.1.1	Rationale for using EGRA- and EGMA- type tests and matching to curriculum criteria	107
E.1.2	Kiswahili	108
E.1.3	Mathematics	108
E.2	Notes on traditional test analysis in the IE	109
E.3	Application of the Rasch model in the impact evaluation	109
E.4	Rasch analysis of Kiswahili baseline, midline and endline pupil test data	110
E.4.1	Overall treatment of Kiswahili items in the Rasch analysis	111
E.4.2	Steps taken in estimating Kiswahili item difficulty	111
E.4.3	Steps taken in estimating person abilities in Kiswahili	114
E.4.4	Kiswahili performance band descriptors	115
E.5	Rasch analysis of maths baseline, midline and endline pupil test data	117
E.5.1	Overall treatment of maths items in the Rasch analysis	117
E.5.2	Steps taken in estimating maths item difficulties	117
E.5.3	Steps taken in estimating person abilities in maths	119
E.5.4	Maths performance band descriptors	120
Annex F	Definition of key quantitative indicators	122
F.1	Chapter 3 Pupil learning and background characteristics	122
F.2	Chapter 4 Teacher performance	124
F.3	Chapter 5 School leadership and management	141
F.4	Chapter 6 Community participation and demand for accountability	156
F.5	Chapter 7 Conducive learning environments for marginalised children, particularly for girls and children with disabilities	161
Annex G	Statistical tables of results from programme areas	168
Annex H	Implementation of other large education programmes	169
H.1	LANES	169
H.2	BRN-Ed/EPforR	171
H.3	Tusome Pamoja	172

List of tables and figures

Figure 1: Visual representation of second PSM with DID combination.....	23
Figure 2: Impact of EQUIP-T on pupil learning	24
Figure 3: Example PSM comparisons.....	25
Figure 4: Kiswahili top band: Second stage results (Main strategy)	29
Figure 5: Kiswahili top band: Matched outcomes at baseline and endline.....	30
Figure 6: Kiswahili top band- Second stage results (Strategy 2)	31
Figure 7: Kiswahili bottom band: Second stage results (Strategy 1).....	32
Figure 8: Kiswahili bottom band: Matched outcomes at baseline and endline	33
Figure 9: Kiswahili bottom band- Second stage results (Strategy 2)	34
Figure 10: Kiswahili Rasch scale: Second stage results (Strategy 1)	35
Figure 11: Kiswahili Rasch scale: Matched outcome at baseline and endline	36
Figure 12: Kiswahili Rasch Scale- Second stage results (Strategy 2)	37
Figure 13: Mathematics top band: Second stage results (Strategy 1)	38
Figure 14: Mathematics top band: Matched outcome at baseline and endline	39
Figure 15: Mathematics top band- Second stage results (Strategy 2)	40
Figure 16: Mathematics bottom band: Second stage results (Strategy 1)	41
Figure 17: Mathematics bottom band: Matched outcome at baseline and endline	42
Figure 18: Mathematics bottom band- Second stage results (Strategy 2)	43
Figure 19: Mathematics Rasch scale: Second stage results (Strategy 1).....	44
Figure 20: Mathematics Rasch scale : Matched outcome at baseline and endline	45
Figure 21: Mathematics Rasch Scale- Second stage results (Strategy 2)	46
Figure 22: Teacher school absenteeism: Second stage results (Strategy 1).....	47
Figure 23: Teacher school absenteeism: Matched outcomes at baseline and endline	48
Figure 24: Teacher School Absenteeism- Second stage results (Strategy 2).....	49
Figure 25: Teacher classroom absenteeism: Second stage results (Strategy 1)	50
Figure 26: Teacher classroom absenteeism: Matched outcome at baseline and endline	51
Figure 27: Teacher Classroom Absenteeism- Second stage results (Strategy 2).....	52
Figure 28: Teacher performance appraisal: Second stage results (Strategy 1)	53
Figure 29: Teacher performance appraisal: Matched outcome at baseline and endline	54
Figure 30: Teacher Performance Appraisal- Second stage results (Strategy 2)	55
Figure 31: Topics covered in school-based training sessions from 2015 to 2017 (trends in programme areas)	66
Figure 32: Completion of EQUIP-T training modules by schools and early grade teachers.....	68
Figure 33: Challenges reported by schools with EQUIP-T training.....	71
Figure 34: Standards 1 to 3 teacher turnover since baseline and midline (trends in programme areas)	79
Figure 35: Standards 1-7 teacher turnover between midline and endline.....	80
Figure 36: Standards 1-7 teacher turnover between baseline and midline	81
Figure 37 ICC for oral reading passage subtest, endline	113
Figure 38 ICC for reading comprehension subtest, endline.....	114
Figure 39 Kiswahili person-item distribution at endline.....	115
Figure 40 ICC for word problem 4, endline	118
Figure 41 ICC for subtraction question 7 from level 1, endline	119
Figure 42 Maths person-item distribution at endline.....	120
Table 1 Endline survey respondents, school-level sampling, and instruments	8
Table 2 Endline survey actual and intended sample sizes	8
Table 3 Quantitative survey instruments from midline.....	9
Table 4: In-service training received by teachers in treatment schools in the previous two years (trends in programme areas)	14
Table 5: In-service training received by teachers in control schools in the previous two years ¹ (trends in control areas).....	15
Table 6: Limitations to the quantitative component of the impact evaluation and mitigating factors.....	19

Table 7: Impact indicators for PSM-DID estimation.....	27
Table 8: Kiswahili top band: PSM-DID estimate.....	30
Table 9: Kiswahili bottom band: PSM-DID estimate.....	33
Table 10: Kiswahili Rasch Score: PSM-DID estimate.....	36
Table 11: Mathematics top band: PSM-DID estimate.....	40
Table 12: Mathematics bottom band: PSM-DID estimate.....	42
Table 13: Mathematics Rasch Scale: PSM-DID estimate.....	45
Table 14: Teacher school absenteeism: PSM-DID estimate.....	48
Table 15: Teacher classroom absenteeism: PSM-DID estimate.....	51
Table 16: Teacher performance appraisal: PSM-DID estimate.....	54
Table 17: Pupils' oral reading speed at baseline, midline and endline in programme schools (trends in programme areas).....	58
Table 18: Pupils' reading and listening comprehension skills at baseline, midline and endline in programme schools (trends in programme areas).....	58
Table 19 Pupils' writing skills at baseline, midline and endline in programme schools (trends in programme areas).....	59
Table 20: Pupils' skills in number comparison and missing numbers at baseline, midline and endline in programme schools (trends in programme areas).....	59
Table 21: Pupils' skills in addition and subtraction at baseline, midline and endline in programme schools (trends in programme areas).....	60
Table 22: Pupils' skills in multiplication and word problems at baseline, midline and endline in programme schools (trends in programme areas).....	60
Table 23 Pupils' background characteristics at baseline, midline and endline in programme schools (trends in programme areas).....	61
Table 24 Trends in average Kiswahili scale scores by gender, home language, poverty status and disability in programme schools (trends in programme areas).....	62
Table 25 Trends in average maths scale scores by gender, home language, poverty status and disability in programme schools (trends in programme areas).....	63
Table 26: Distribution and estimates (in absolute terms) of Standard 3 pupils by Kiswahili performance band in treatment areas, baseline, midline, and endline (trends in programme areas).....	64
Table 27: Distribution and estimates (in absolute terms) of Standard 3 pupils by maths performance band in treatment areas, baseline, midline, and endline (trends in programme areas).....	64
Table 28: Profile of facilitators of EQUIP-T school-based training sessions.....	67
Table 29: INCO post at school.....	69
Table 30: Profile of INCO.....	69
Table 31: Suggested improvements by schools to the EQUIP-T in-service training.....	71
Table 32: Challenges reported by Standards 1 to 3 teachers with EQUIP-T training (trends in programme areas).....	72
Table 33: Teacher access to curriculum, syllabi and teacher guides, self-reported by teachers (trends in programme areas).....	73
Table 34: Background characteristics and qualifications of Standards 1 to 3 teachers (trends in programme areas).....	74
Table 35: Teacher output indicators at endline by gender and age.....	75
Table 36: Teacher intermediate outcome indicators at baseline, midline, and endline by gender (trends in programme areas).....	75
Table 37: Teacher intermediate outcome indicators at baseline, midline, and endline by age (trends in programme areas).....	76
Table 38: Turnover of teachers between midline and endline disaggregated by gender.....	81
Table 39: Turnover of teachers between midline and endline disaggregated by age.....	82
Table 40: New teachers employed at the school (trends in programme areas).....	82
Table 41: Turnover in INCO post.....	83
Table 42: Head teacher turnover and reasons (trends in programme areas).....	83
Table 43: Head teacher previous job and location (trends in programme areas).....	84
Table 44: Turnover in WEO post.....	85
Table 45 Impact evaluation treatment and control districts.....	91

Table 46: Stakeholder consultations and events—from dissemination of midline findings to plans for dissemination of quantitative endline findings	94
Table 47 EQUIP-T Impact Evaluation Reference Group members (December 2018)	96
Table 48: Endline quantitative impact evaluation team members and roles	97
Table 49: Comparison of estimated Kiswahili item locations from independent baseline, midline and endline Rasch analyses	113
Table 50: Kiswahili performance band descriptors	116
Table 51 Maths performance band descriptors	121
Table 52: LANES activities in 2014 and 2015	169
Table 53: LANES activities in 2016 and 2017	170
Table 54: BRN-Ed programme activities in 2014 and 2015.....	171
Table 55 EPforR activities in 2016 and 2017	171
Table 56 Tusome Pamoja (TP, let’s read together) activities in 2016 and 2017	172
Box 1 Detailed procedure for within-school pupil sampling	7
Box 2 Reasons for not reporting descriptive trends from control schools, teachers or pupils	13
Box 3 Large scale primary education development programmes, apart from EQUIP-T.....	14
Box 4: Assumptions related to the interpretation of the impact estimates.....	17
Box 5: EQUIP-T impact on teacher performance appraisals	78
Box 6: National 3R assessment targets	107

List of abbreviations

3Rs	Reading, writing, and arithmetic
ADEM	Agency for the Development of Education Management
ATT	Average treatment effect on the treated
BL	Baseline
BRN-Ed	Big Results Now in Education
CAPI	Computer aided personal interviewing
CENA	Community Education Needs Assessment
CSO	Civil society organisation
DFID	Department for International Development
DID	Difference-in-differences
DSI	District school inspector
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EPforR	Education Program for Results
EQUIP-T	Education Quality Improvement Programme in Tanzania
ESDP	Education Sector Development Plan
ETP	Education and Training Policy
IDELA	International development and early grade learning assessment
IE	Impact evaluation
IGA	Income-generating activity
INCO	In-service training coordinator
ISS	Institutional strengthening and sustainability
JUU	<i>Jiamini Uwezo Unao</i> ('Be confident, you can')
LANES	Literacy and Numeracy Education Support Programme
LGA	Local Government Authority
MA	Managing Agent
ML	Midline
MOEST	Ministry of Education, Science and Technology

MOEVT	Ministry of Education and Vocational Training ¹
OECD-DAC	Organisation for Economic Co-operation and Development – Development Assistance Committee
OPM	Oxford Policy Management
PO-RALG	President’s Office Regional Administration and Local Government
PSM	Propensity score matching
PTP	Parents–teachers partnership
RCT	Randomised control trial
RTI	Research Triangle International
SC	School Committee
SDP	School Development Plan
SIDA	Swedish Development Agency
SIS	School information system
SLM	School leadership and management
SPMM	School performance management meeting
SRP	School Readiness Programme
SQA	School quality assurance/assurer
STEP	Student Teacher Enrichment Program
TCF	Teacher Competency Framework
TDNA	Teacher development needs assessment
TLMs	Teaching and learning materials
TOR	Terms of reference
TZS	Tanzanian shilling
UNESCO	UN Educational, Scientific and Cultural Organization
UNICEF	UN Children’s Fund
URT	United Republic of Tanzania
USAID	US Agency for International Development
WEO	Ward Education Officer

¹ Before the change of government in 2015 MOEST was called the Ministry of Education and Vocational Training (MOEVT).

1 Introduction to Volume II

1.1 Overview

This is the second Volume of the endline quantitative impact evaluation report of the Education Quality Improvement Programme in Tanzania (EQUIP-T). EQUIP-T is a six-year, Government of Tanzania (GoT) programme with a budget of £90m funded by the UK Department for International Development (DFID). The aim of the programme is to increase the quality of primary education and improve pupil learning outcomes, in particular for girls. Initially, the programme was intended to be four years, with activities targeted at five, and later seven, of the most educationally disadvantaged regions in Tanzania.² In 2017 the programme was extended to 2020, and the extension included introducing some new subcomponents to the seven regions and a reduced package of interventions to two new regions.

The main aims of this first part of the endline evaluation are to estimate the impact of EQUIP-T on pupil learning achievement, and to assess the effectiveness of the school- and community-level EQUIP-T interventions, after nearly four years of implementation (44 months).

The results are intended to inform further adjustments to the programme before it finishes in January 2020, as well as to promote accountability and lesson learning for DFID and GoT. Its findings will also help to guide the design of the second part of the endline, which will include qualitative research in 2019.

This quantitative endline evaluation report is organised into two volumes. **Volume I (Results and Discussion)**, presents the main findings, conclusions, and recommendations. It also identifies a number of lessons learned which could be relevant to readers inside and outside of the Tanzania context involved in designing and implementing education programmes with similar objectives. **Volume II (Methods and Supplementary Evidence)** contains technical methods sections, as well as supplementary quantitative analysis to support the conclusions reached in Volume I. Readers interested in the more in-depth evidence base for the endline findings should read both Volumes.

1.2 Structure of this Volume

Volume I contains three parts: Part A: Impact evaluation objectives, background and methods; Part B: Endline findings; Part C: Conclusions, recommendations and lessons. Volume II is divided into two further parts, as follows:

- **Part D: Impact evaluation methods.** Full details of the overall impact evaluation design and methods are given in Volume II of the baseline evaluation report (OPM 2015b). Where details had changed between baseline and midline, Volume II of the midline evaluation report (OPM 2016b) set these out. Similarly in this part of the endline report, the full details of design and methods are not repeated but are referenced where appropriate. Instead the chapters summarise key design features and explain any changes made at endline as well as any specific risks and limitations to the endline analysis. The two chapters cover: mixed methods (Chapter 2); and quantitative methods (Chapter 3).
- **Part E: Supplementary endline evidence.** The two chapters in this part present additional quantitative results to support the main findings in Volume I. Chapter 4 discusses the impact estimates in greater detail than was possible in Volume I, and also explains the impact estimation methodology in technical detail. Chapter 5 covers additional descriptive material on trends in key

² There are 26 regions on mainland Tanzania.

indicators in programme treatment schools—this is structured according to the finding chapters in Volume I for easy reference (pupil learning, teacher performance, SLM, community engagement and accountability; and conducive learning environment for marginalised children).

These two parts are supplemented by eight annexes: a table showing the impact evaluation treatment and control districts (Annex A); processes for stakeholder engagement and governance in the impact evaluation (Annex B); endline survey fieldwork details and ethical protocols (Annexes C and D); a technical annex on the measurement of pupil learning using Rasch modelling (Annex E); definitions of the key quantitative indicators reported in Volume I (Annex F); statistical tables of results from the programme treatment districts (Annex G); and information on the implementation of other large education programmes in Tanzania (Annex H).

Part D: Methods



2 Mixed methods approach

2.1 Impact evaluation objectives

The main objectives of the impact evaluation are to: i) generate evidence on the impact of EQUIP-T on primary pupil learning outcomes, including any differential effects for boys and girls; ii) examine perceptions of effectiveness of different EQUIP-T components; iii) provide evidence on the fiscal affordability of scaling up EQUIP-T after the programme closes; and iv) communicate evidence generated by the impact evaluation to policy-makers and key education stakeholders.

2.2 Impact evaluation methods

The impact evaluation uses a mixed methods approach whereby quantitative and qualitative methods are integrated to provide robustness and depth to the research findings. This rests on both the integration of methodologies for better measurement, and the sequencing of information collection for better analysis. Two rounds of evaluation research have preceded this endline quantitative research: a baseline in 2014 before the programme interventions started, and a midline in 2016, almost two years into implementation. The application of mixed methods was fairly similar in both rounds of research, and relied on the sequencing of the quantitative and qualitative data collection within a relatively short period within a school year. In both rounds, the quantitative data was collected in April/May, and the qualitative data collection followed.³ For details on the use of mixed methods at midline, see OPM 2017b, chapter 2 pp4-6. The final baseline and midline evaluation reports present integrated quantitative and qualitative findings.

The application of mixed methods for the endline research is different to the approach taken so far, as it is complicated by the split in the timing of the quantitative and qualitative data collection across two school years (2018 and 2019). This has arisen because of a trade-off between two important objectives that came to light following the extension of EQUIP-T from 2018 to January 2020.⁴ Concerns about potential contamination of the impact measurement of the programme on pupil learning, led to the decision to collect the quantitative endline data in 2018, as originally planned. This timing means that the endline survey took place before EQUIP-T interventions happened at school-level in Singida—a district which contains two of the evaluation’s eight control districts. On the other hand, concern that some of EQUIP-T’s most relevant interventions to national priorities, such as those aimed at supporting girls and promoting inclusive education, are relatively new and thus may not have sufficient implementation time to be evaluated well in 2018, largely led to the decision to conduct the qualitative research in 2019. Moreover, some of these interventions are more suited to being evaluated using qualitative methods such as those related to empowerment, tackling cultural norms and taboo issues.

Overall then, the sequencing of the endline quantitative and qualitative components over a much longer period, means that the approach to ‘mixing’ evidence will be different. The quantitative part of the endline evaluation (this present report) focuses on measuring impact on pupil learning, and on providing quantitative evidence on the effectiveness of the school- and community-level interventions up to early 2018. These results will then feed into the design of the qualitative endline research in 2019, and help to prioritise research themes for follow-up (continuation of evaluation narrative). Qualitative research methods provide depth in evaluation evidence on a narrow number of themes,

³ The qualitative data was collected in April/May at midline, and June-August at baseline.

⁴ Following the extension of EQUIP-T with expanded duration, scope of activities and geography, DFID requested OPM to propose alternatives to the planned 2018 endline evaluation approach. OPM shared an Options Paper with DFID in July 2017, and held discussions with DFID in September 2017 to agree on an approach.

and there will be a consultative process in late 2018 or early 2019 to decide on where richer evidence is most warranted. As at midline, there will be two levels of endline qualitative research: school/community and regional/district—the latter will provide evaluation evidence on the district planning and management EQUIP-T component as this is not covered in the quantitative part of the endline evaluation.⁵

There will also be a separate cost study conducted in 2019, following on from the initial analysis done at midline (see OPM 2017b, chapter 5). At endline, the study will aim to analyse programme spending patterns, efficiency using cost to output ratios (for selected interventions), and affordability for government of potential scaling up of parts of the programme to other districts/regions. It will also explore whether it is possible to compare relevant costs of the programme with impact estimates, recognising that there may well be insurmountable constraints in the available data (see OPM 2017b for initial exploration of these issues).

In summary, the approach to ‘mixing’ different types of evaluation evidence at endline is sequential, with the qualitative analysis serving partly to deepen understanding of selected findings from the quantitative research (Greene *et al.*, 1989), and also to explore priority themes that are not amenable to quantitative methods. Put together with findings from the cost study on affordability, this should help the Government and its development partners to make decisions on whether to scale-up parts of the programme nationally.

The next chapter provides details on the design of the quantitative impact evaluation, including the sampling strategy, the instruments, and the analytical methods. It highlights any adaptations that have been made for the endline research.

⁵ Apart from some findings on WEO support to schools that is included in the evaluation of the SLM component (see chapter 5 in Volume I).

3 Quantitative impact evaluation design and endline adjustments

The core impact evaluation quantitative methodology involves measuring a consistent set of indicators in a panel of 200 schools, 100 treatment (EQUIP-T programme schools) and 100 control schools, over three rounds of research (2014, 2016 and 2018) at the same time of year (April/May). The core quasi-experimental methods used at baseline were replicated at midline, but some adjustments were made to the data collection instruments, school-level sampling of teachers, and the fieldwork model (OPM 2017b, pp7-12).⁶ At endline there needed to be some further adjustments to the data collection instruments, but the rest of the design, including the fieldwork model and protocols remained the same as at the midline.

The baseline evaluation report volume II (OPM 2015b, Chapter 3, pp2-28) explains the full details of the quantitative evaluation design including the rationale, sampling strategy, instrument development, analytical methods, as well as key methodological risks and limitations. This detail is not repeated here in full, but the core design features are summarised briefly in the sections below, together with an explanation of the adjustments made for the endline research, a table showing endline sample turnout, and a section covering the specific risks and limits to the quantitative evaluation at endline.

3.1 Rationale for quasi-experimental design

One of the main objectives of the impact evaluation is to be able to robustly attribute changes in key impact-level and outcome-level indicators to EQUIP-T as a whole. The EQUIP-T Managing Agent (MA) purposively selected the regions and districts into the programme on the basis of these being disadvantaged in terms of education and other social and economic indicators. In the absence of random assignment, a pure randomised control trial (RCT) was not possible, and the impact evaluation employed the best possible approach to simulate the RCT approach. In this case, this was to mimic randomisation using propensity score matching (PSM), and then to employ a PSM and difference-in-difference (DID) approach to estimating programme impact (see Chapter 4 for details of how the impact estimation was carried out in practice using baseline, midline and endline data).

3.2 Sampling strategy, sample size and instruments

3.2.1 Sampling strategy

Prior to sampling, a list of eligible treatment and control districts was established by excluding districts that are: (i) in Lindi and Mara as these are part of the EQUIP-T programme but not covered by the impact evaluation; (ii) receiving other education programmes that aim to influence the same outcomes as EQUIP-T including Big Results Now in Education (BRN-Ed), Kiufunza, UNICEF's school-based INSET programme and USAID's TZ21 programme; (iii) part of OPM's baseline pre-tests.

The sampling was carried out in four stages:

1. **Selection of control districts:** PSM was used to match eligible control districts to the 17 pre-selected eligible treatment districts.
2. **Selection of treatment schools:** schools in the treatment districts were selected using stratified random sampling.

⁶ Instead of sampling Standard 1-3 teachers for interview in each of the sample schools, as was done at baseline, all were interviewed to boost the overall sample size of early-grade teachers.

3. **Selection of control schools:** PSM was used to match eligible control schools to the sample of treatment schools.
4. **Selection of pupils within schools:** pupils were sampled within schools using systematic random sampling. The within-school sampling was assisted by selection tables automatically generated within the computer assisted survey instruments. Information on the detailed procedure followed by enumerators to select the 15 pupils is in Box 1. **Selection of teachers within schools:** at baseline, simple random sampling was used to draw a sample of Standards 1-3 teachers to be interviewed and to fill in a teacher development needs assessment (TDNA). At midline, instead of sampling Standards 1-3 teachers all of them were selected for interview, but the baseline sampling strategy remained for the TDNA. Simple random sampling was also used to select a sample of Standards 4-7 maths teachers to fill in a TDNA. At endline, all Standards 1-3 teachers were selected for interview. The TDNAs were not administered at endline (see Section 3.2.4 below for reasons).

Box 1 Detailed procedure for within-school pupil sampling

Enumerators were trained to use the following procedure for sampling Standard 3 pupils:

Collect the Standard 3 attendance registers for all streams, and check that they are filled in for 'today', then follow this sampling procedure:

1. Use a pencil. Have a rubber available.
2. Starting at 1, write a sequential series of numbers beside the names of all pupils who are present today.
3. If there is more than one stream in Standard 3, continue the number series on to the next registers.
4. The final number in your pencil number series is the number of pupils present today in Standard 3. Enter this number into the cell in the computer assisted personal interviewing (CAPI) instrument.
5. The CAPI instrument will automatically produce 15 pupil selection numbers in red font.¹
6. Look again at the pencil number series you marked on the register/s. Find the pupil name which corresponds to the first selection number. Write the pupil's name into the sampled pupils table.
7. Repeat the step above for the other 14 selection numbers. You will need to scroll in the table to see the spaces to enter all of the sampled pupils.²

Source: OPM (Midline Fieldwork Manual, April 2016). Notes: 1) The CAPI instrument automatically generates a random set of 15 different numbers of maximum value equal to the number in Step 4. 2) There is also a procedure for replacement in the event that any of the 15 pupils cannot take the test for a valid reason, for example being ill.

Table 1 shows the endline survey respondents, sampling and instruments. Two of the instruments — lesson observation and the small group teacher interview (a newly introduced instrument at endline) — were only conducted in treatment schools, because the information generated could not be used in the impact modelling and so collecting information in control schools was not necessary. More details are provided on the instruments in Section 3.2.4.

Overall, the sampling strategy yields a panel of schools (same schools visited during each round), and a repeated cross-section of Standard 3 pupils and Standards 1-3 teachers.

At baseline five out of 200 schools had to be replaced during fieldwork using a carefully controlled reserve list. Replacement was not necessary at midline or endline, and all schools that were interviewed at baseline were revisited and interviewed at midline and endline (see Annex D for endline fieldwork details).

Table 1 Endline survey respondents, school-level sampling, and instruments

Respondent	School-level sample	Instrument
Standard 3 pupils	Sample (15 pupils present on the day)	Adapted Early Grade Reading Assessment (EGRA) Adapted Early Grade Maths Assessment (EGMA) Pupil background
Parents of tested Standard 3 pupils	Sample (15 parents)	Poverty score card
Standards 1 to 3 Kiswahili and maths teachers	No sample ¹	Interview
In-service co-ordinator (INCO) together with some teachers that received in-service training	No sample ² – Treatment schools only	Small group interview
Head teacher	No sample	Interview, School records
Enumerator observation of Standard 2 lessons	Convenience sample ³ – Treatment schools only	Lesson observation
Enumerator observation	No sample	Head count (of teacher and pupil attendance)

Source: OPM EL survey. Note: (1) At baseline, a sample of Standards 1 to 3 teachers were interviewed. (2) The enumerator invited all teachers that have attended EQUIP-T in-service training away from school since baseline to join the group interview. (3) Enumerators selected Standard 2 3Rs lessons on the basis of opportunity to observe given the time the survey team were in the school.

3.2.2 Sample size

The theoretical justification for the choice of target sample sizes for each unit is explained in the baseline report volume II p8. Table 2 contains the endline survey's actual and intended sample sizes.

Table 2 Endline survey actual and intended sample sizes

Sampling unit	Treatment sample			Control sample		
	Target sample	Actual sample	Actual/Target (%)	Target sample	Actual sample	Actual/Target (%)
Regions	5	5	100	7	7	100
Districts	17	17	100	8	8	100
Schools ¹	100	100	100	100	100	100
Std. 3 pupils (tested both in Kiswahili and maths)	1,500	1,499	99.9	1,500	1,500	100
Parents of tested pupils (poverty scorecards)	1,500	1,495	99.7	1,500	1,497	99.8
Stds. 1–3 Kiswahili/maths teacher interviews ²	441	435	98.6	455	454	99.8
Teachers' group interview on in-service training	100	99	99.0	n.a.	n.a.	n.a.
Std. 2 lesson obs.maths ³	100	95	95.0	n.a.	n.a.	n.a.
Std. 2 lesson obs.Kiswahili ³	100	101	101	n.a.	n.a.	n.a.

Source: Evaluation endline survey. Notes: (1) The school instruments are: head teacher interview, data collection from school records, and head count of teacher and pupil attendance. (2) The samples include 16 head teachers (treatment) and 25 (control) who teach Kiswahili or maths to Stds. 1–3. All 7 teachers who were not interviewed were unavailable (absent on the day and could not be reached over the phone later). Some 11% of teachers were interviewed over the phone because they were absent on the day of the survey. (3) 95 maths (arithmetic) lessons and 101 Kiswahili lessons (either reading or writing) were observed. Some of these subjects were taught consecutively (without a break) in one class period. 138 separate class periods were observed.

Response rates are very high in the endline survey (Table 2). Actual sample sizes are close to target sample sizes for pupils, their parents and teachers. The lowest response rate is 95% for maths lesson observations. These response rates are similar to those obtained at baseline and midline—and slightly higher in most cases. Of the Standards 1-3 teachers due to be interviewed at endline, 11% were unavailable on the day of the survey but were later interviewed by phone—this share is similar to the 9% of teachers interviewed by phone at midline.

3.2.3 Survey weights

In order to obtain indicator estimates that are representative of the EQUIP-T programme areas (more specifically, the 17 districts that comprise the impact evaluation sample), to feed into the descriptive trends that support the analysis of the theory of change, estimates were weighted using survey weights. The survey weights are computed as the normalised values of the inverse probabilities of selection into the sample for each unit of observation. The formulae for computing the weights for different units (schools, pupils and teachers) are in the baseline impact evaluation report volume II (p11). At baseline there were two sets of teacher weights, one for sampled teachers (those who were interviewed or took TDNAs) and one for roster teachers (for indicators which use data on all teachers in a school). At midline this was extended to three sets of teachers weights because all Standards 1-3 teachers were interviewed rather than a sample. At endline only two sets of teacher weights were needed: one for roster teachers (all teachers in a school), and another for Standards 1-3 teachers (all were interviewed).

The survey weights were applied within a survey set up in Stata (the statistical programme used to analyse the data) that takes into account clustered sampling, stratification and finite population corrections.

3.2.4 Survey instruments

The endline survey uses a set of instruments that retain most of the midline questions, but with some additions (and removals) to take into account changes in programme context and design. The content of the survey instruments used at midline are summarised in Table 3 below. The enumerators administered all of the instruments on tablets using Computer Assisted Personal Interviewing (CAPI).

Table 3 Quantitative survey instruments from midline

Description of content	
1. Standard 3 pupil Kiswahili test (same pupils tested in both Kiswahili and mathematics)	
• Kiswahili literacy pupil test based on standard 1 and 2 curriculum requirements	Early Grade Reading Assessment (EGRA)
2. Standard 3 pupil mathematics test	
• Mathematics pupil test based on standard 1 and 2 curriculum requirements	Early Grade Mathematics Assessment (EGMA)
3. Standard 3 pupil background interview	
• Pupil background	Short pupil interview
• Pupil's school experience	
4. Parents of Standard 3 tested pupil interview	
• Set of household characteristics (that can be used to convert scores into poverty likelihoods based on a pre-existing instrument)	Poverty score card
• Pupil background	
• Home support for schooling	
5. Standards 1-3 teacher interview	
• Background information: gender, age, years of teaching, qualifications	Teacher interview

Description of content	
<ul style="list-style-type: none"> Frequency/type of in-service training received Classroom teaching and pupil assessment practices Support for teaching: lesson planning, observation, meetings, PTPs Morale and other conditions of service 	
6. Teacher development needs assessment Kiswahili and mathematics	
<ul style="list-style-type: none"> Teacher Kiswahili and mathematics subject knowledge assessment based on the primary school curriculum (standards 1-7 with limited materials from standards 1 and 2) 	Teacher Development Needs Assessment (TDNA)
7. Standard 2 Kiswahili and mathematics lesson observations	
<ul style="list-style-type: none"> Inclusive behaviour of teachers with respect to gender and spatial location of pupils Key teacher behaviours in the classroom Pupils' reading and teacher support Availability of lesson plan Availability of desks, textbooks, exercise books, pencils, supplementary reading books during the lesson 	Classroom mapping Lesson observation
8. Head teacher interview, data collection from school records and headcount	
<ul style="list-style-type: none"> Background information on head teacher: gender, age, years of experience, qualifications In-service training on school leadership and management School background information: teachers, physical facilities, school timetable, number of days school open Teacher management School development plan, school information, school committee, parent-teacher partnerships, community engagement School resources: cash (including capitation grants) and in-kind External support for school leadership and management Morale and other conditions of service Teacher punctuality and attendance (by records and by headcount on the day) Pupil attendance (by records and by headcount on the day) 	Head teacher interview School records checks Enumerator observation

Source: OPM 2018, p.27-28

It is critical that the core information collected over multiple rounds remains the same so that estimates of key indicators can be reliably tracked over time. The midline instruments were not changed substantially for use at endline. Nonetheless, as the programme design and context has evolved since midline, additional information needed to be captured to inform the analysis of change, and some information became less relevant. Also, there were problems with some questions at midline, such as ambiguous wording, which only became apparent at the analysis stage and it made sense to address these issues in a set of revised instruments. There were two overarching changes to the suite of instruments, compared to the midline set, as follows:

- TDNA instruments dropped:** These were designed to measure teachers' Kiswahili and maths subject knowledge, and were introduced at baseline because one of the original objectives of the early-grade teachers' in-service training intervention was to strengthen subject knowledge. Teachers scored about 60% on average in both subjects at baseline, and the results did not change significantly by midline. This lack of change following programme implementation was not unexpected, as the final design of EQUIP-T's in-service training chose not to focus on subject knowledge. At endline it made more sense to direct data collection efforts on instruments that are more directly relevant to the programme's interventions, and so the decision was taken to drop the TDNA.
- New small-group interview with teachers (focused on in-service training):** The early-grade teacher in-service training is central to the programme's theory of change, and has absorbed a large share of the programme's spending. For this reason, it merits particular focus in the impact evaluation. Attendance at in-service training is already captured in early-grade teacher interviews, but given the high level of teacher turnover that was found at midline, getting a picture of the

delivery of in-service training at a school level over the duration of the programme is useful complementary data. By gathering a small group of teachers that have attended the different types of in-service training (3Rs curriculum, Kiswahili, maths and gender-responsive pedagogy), as well as the in-service training coordinator (INCO, typically the academic teacher), this instrument captures the delivery of the various residential in-service training courses, as well as the school-based in-service training sessions. Where possible, participants were encouraged to refer to records to provide information. This instrument required skilled facilitation, to help the respondents to reach consensus on the answers (in cases where there are no records), and to ensure that the interview was not dominated by one or two individuals. This approach to data gathering is commonly used at the community-level.

Apart from these two changes to the group of instruments, these are the main changes that have been made to the other midline instruments:

- **Parents of tested standard 3 pupils interview (score card):** addition of questions on their child's pre-school attendance (including school readiness programme (SRP)); communication with the school; awareness of the parent-teachers partnership (PTP); and corporal punishment.
- **Standards 1-3 teacher interview:** addition of specific questions on EQUIP-T in-service training modules completed since baseline; attendance at ward cluster reflection meetings and school performance management meetings (SPMMs); outstanding non-salary claims; removal of questions on receipt of salary.
- **Standard 2 lesson observation:** addition of observations related to gender-responsive pedagogy; use of maths learning materials (not textbooks); display of positive and safe learning campaign-related materials.
- **Head teacher interview and school records:** addition of questions related to initiatives to support pupil welfare (e.g. health, hygiene, safety and child protection); initiatives to support marginalised groups of pupils (girls, children with disabilities, pupils with learning difficulties, pupils that are vulnerable for other socio-economic reasons); new EQUIP-T interventions since midline (tablet-based SIS, business plans and income-generating activities (IGA), SPMMs, JUU clubs, pupil suggestion boxes); PTP grant spending patterns; more detail on head teacher's attendance at in-service training; outstanding non-salary claims; removal of questions on receipt of salary, missing ages of baseline pupils, and information for sampling teachers for TDNA.

The revisions to the midline instruments were trialled during a pre-test held in February 2018 (see Annex D for details).

The original development of the instruments, and their contents, is described in detail in the evaluation baseline report volume II (pp13-18). Given the importance of the measurement of pupil learning to the impact evaluation (improving learning achievement is the main goal of the programme), it is worth briefly summarising the test design process. The OPM design team worked with a national team of specialists comprising Kiswahili and maths specialists from the University of Dar es Salaam, primary school teachers and a Tanzanian test design specialist, to develop the two pupil tests. The team developed new items adapted from an existing Early Grade Reading Assessment (EGRA) and an Early Grade Maths Assessment (EGMA) that was being used to monitor the Government's BRN-Ed programme. Three pre-tests with purposive sampling were carried out to check item difficulty and discrimination, clarity of wording, protocols for accurate measurement, and child-friendliness. The test items are kept secure so that they can be re-administered each round.

3.3 Fieldwork timing and model

The endline fieldwork took place during the same months, April/May, as the baseline and midline fieldwork. Planning the endline fieldwork was made easier by the national adoption of standard dates for the mid-term break holiday just prior to fieldwork. At both baseline and midline, these dates varied by region, giving less flexibility in planning. There were no changes to the fieldwork model or protocols compared with midline, apart from for the new small-group teacher interview.

Teams of five or six enumerators visited schools over one day. Most of the data collection took place at school level, with the exception of the parent interview where enumerators went to parents' households to interview them. As at midline, the key risk of using a one day model is that head teachers and early-grade teachers may be absent for interview. This risk was mitigated by using revisits in some cases, and by using phone interviews. This worked well, and the overall response rate for teacher interviews was 99% (Table 2). The pupil sampling is not affected by a one-day fieldwork model as the sample is drawn from those present on the day of the visit.

3.4 Quantitative analysis

3.4.1 Impact estimation

The quasi-experimental design relies on a propensity score matching (PSM) with difference-in-differences (DID) technique to estimate programme impact on a small set of impact and outcome indicators (see list below). As explained in detail in the midline evaluation report volume II (OPM 2017b, p34), this is an innovative approach that brings together PSM and DID in the specific context of the EQUIP-T evaluation, which is based on a panel of schools but repeated cross-sections of pupils and teachers. It is important to note that in a PSM estimation, outcome indicators from treatment units (i.e. programme school teachers and pupils) are compared to outcome indicators from specific control units based on the propensity score. This implies that the estimated average treatment effect will be valid for the group of treatment observations only, which, in turn, means that PSM produces an estimate of the Average Treatment Effect on the Treated (ATT). Extrapolating this estimate beyond the population for which the treatment sample is representative is not valid.

Programme impact has been estimated on the following indicators, which are grouped under the relevant level of the results chain from the programme's theory of change.

EQUIP-T impact: Improved learning outcomes, especially for girls

- Proportion of Standard 3 pupils in the bottom Kiswahili performance band (%)
- Proportion of Standard 3 pupils in the top Kiswahili performance band (%)
- Proportion of Standard 3 pupils in the bottom mathematics performance band (%)
- Proportion of Standard 3 pupils in the top mathematics performance band (%)
- Mean pupil test score in Kiswahili (scaled as Rasch scores in logits)
- Mean pupil test score in maths (scaled as Rasch scores in logits)

These indicators have also been disaggregated by gender in a separate descriptive analysis to assess how learning gaps have changed between baseline and endline.

EQUIP-T intermediate outcomes 1 and 2: Improved teacher performance, and enhanced SLM

- Teacher school absenteeism (%)

- Teacher classroom absenteeism (%)
- Proportion of Standards 1-3 teachers who report participation in performance appraisal (%)

Chapter 4 contains further technical details on the impact estimation methods, including the use of a main identification strategy supported by a complementary strategy to ensure robust results. This chapter also contains supplementary results from the impact analysis.

3.4.2 Descriptive analysis of change in programme schools

A descriptive analysis of trends in the many quantitative indicators that were measured at baseline, midline and endline is used to help understand whether changes have happened or not in programme schools, as anticipated in the programme theory of change. This in turn helps to explain the results from the impact analysis. The descriptive analysis is guided by the Endline Evaluation Matrix Part I (provided in Annex C in Volume I) which contains the endline evaluation questions, linked to the theory of change that can be answered using quantitative evidence.

Note that this report does not contain descriptive trends for the control group of schools because directly comparing results from programme (treatment) and control schools, teachers and pupils would be misleading because of the way the control group was sampled (see Box 2 for further explanation).

Box 2 Reasons for not reporting descriptive trends from control schools, teachers or pupils

Given the quasi-experimental nature of the evaluation design, pupils, teachers and schools belonging to treatment and control groups are not immediately comparable. A number of modelling and analytical techniques are needed to ensure that selection bias is controlled for and the two groups can be compared for the measurement of programme impact (see Chapter 4 for details of the PSM and DID methods applied). This is also the reason why it is not advisable to present descriptive statistics pertaining to the control group only or any analysis that simply compares descriptive statistics in the treatment and control groups. Whilst the treatment group school sample is representative of EQUIP-T schools in the districts covered by the evaluation, the control group school sample has been purposefully matched to the treatment group; it is not, in itself, representative of any underlying population since it has not been selected through probability sampling. In other words, the control group descriptive statistics cannot be weighted based on their probability of selection and would therefore not be representative of any meaningful population. Similarly, juxtaposing treatment and control descriptive patterns and trends either separately at baseline, midline and endline or over time, would not be informative and would in fact be misleading. The only way in which treatment and control schools, teachers and pupils are comparable is through the impact estimation approach based on the PSM with DID analysis.

3.4.3 Quality assurance

The impact estimation analysis was reviewed internally by OPM's statistical methods team, and then by a UK-based academic researcher, familiar with these methods and their application in education. Two other reviewers (a senior Tanzanian academic and an ex-World Bank Senior Education Specialist for Tanzania), provided comments and feedback on the descriptive analysis and interpretation. This same external reviewing team provided feedback on the draft baseline and midline evaluation reports. Annex B provides further details on the overall quality assurance processes applied in this study.

3.5 Risks and limitations

3.5.1 Contamination risk from other programmes

As the baseline report volume II (p26) highlights, the most common risk in longitudinal surveys is potential contamination of the selected impact study areas by third party interventions that may affect the outcomes of interest to the evaluation.

In this case, the risk of contamination comes from several large-scale primary education development programmes that have been working to improve the quality of primary education under the Government's Education Sector Development Plan (ESDP) during the period that EQUIP-T has been operating. The impact evaluation identified these programmes and this risk from the outset, and have continued to monitor their main activities and geographical coverage, as well as collecting relevant information in the evaluation surveys, in order to assess the risk and understand its implications for the impact estimates produced in this study.

The three large-scale primary education programmes which potentially pose a contamination risk are: the Literacy and Numeracy Education Support Programme (LANES); the Education Program for Results (EPforR), formerly Big Results Now-Education (BRN-Ed); and Tusome Pamoja (let's read together)—a programme which started after the midline evaluation of EQUIP-T (see Box 3).

Box 3 Large scale primary education development programmes, apart from EQUIP-T

LANES: this runs from 2014/15 to 2018/19, funded by the Global Partnership for Education, with a budget of US\$95m. It has national coverage (26 mainland regions) except for a few activities.

EPforR formerly BRN-Ed: this runs from 2014/15 to 2020/21 funded by World Bank, SIDA, DFID and GOT, with a budget of US\$416m. It has national coverage, except for a few of the earlier activities under BRN-Ed.

Tusome Pamoja (let's read together): this runs from 2016 to 2021, funded by USAID. Its covers four mainland regions (Iringa, Morogoro, Mtwara and Ruvuma) and all districts in Zanzibar.

Source: OPM 2018

The main activities conducted by these programmes in some of the impact evaluation study areas (mainly in the study control districts) that are most closely related to improving early-grade pupil learning outcomes are training activities for teachers, and TLM distribution. Annex H contains tables which summarise LANES, EPforR/BRN-Ed and Tusome Pamoja programme implementation between the baseline and midline evaluation (2014 and 2015); and between the midline and endline evaluation (2016 and 2017).⁷ These tables set out programme activities and their geographical coverage in relation to the EQUIP-T regions/districts. This information, together with data collected in the evaluation surveys, particularly from teachers on the receipt of in-service training by provider (see Table 4 and Table 5 below) has been used to assess the contamination risks, as follows.

Table 4: In-service training received by teachers in treatment schools in the previous two years (trends in programme areas)

Indicator	Baseline		Midline		Endline	
	Estimate	N	Estimate	N	Estimate	N
Attended any in-service training over the last two years (% Stds 1-3 teachers)	8.47	327	82.58	384	96.16	418

⁷ This information comes from available programme implementation reports, supplemented by interviews with the National LANES coordinator. These tables were included in the midline and endline evaluation planning reports (OPM 2016a, OPM 2018).

Attended in-service training over the last two years provided by: (% Stds 1-3 teachers)						
EQUIP-T	0	327	81.88	384	94.16	418
LANES	0	327	0	384	42.87	418
BRN-Ed	2.72	327	0.12	384	4.49	418
STEP (under BRN-Ed)	0.36	327	0	384	13.38	418
Tusome Pamoja					0	418
Other	5.85	327	5.53	384	8.15	418

Sources: Evaluation baseline, midline and endline surveys (teacher interview).
Notes: (1) Weighted estimates. (2) This is for all interviewed teachers who teach maths or Kiswahili to Standards 1-3. (3) The relevant period for BL is 2012-2013, for ML 2014-2015 and for EL 2016-2017.

Table 5: In-service training received by teachers in control schools in the previous two years¹ (trends in control areas)

Indicator	Baseline		Midline		Endline	
	Estimate	N	Estimate	N	Estimate	N
Attended any in-service training over the last two years (% Stds 1-3 teachers)	4.58	349	54.41	397	70.73	427
Attended in-service training over the last two years provided by (% Stds 1-3 teachers)						
EQUIP-T	0	349	0.25	397	0.23	427
LANES	0	349	50.38	397	53.4	427
BRN-Ed	0.86	349	0.5	397	2.81	427
STEP (under BRN-Ed)	0.29	349	0	397	5.85	427
Tusome Pamoja					7.49	427
Other	3.44	349	6.3	397	13.58	427

Sources: Evaluation baseline, midline and endline surveys (teacher interview).
Notes: (1) **Unweighted estimates**. (2) This is for all interviewed teachers who teach maths or Kiswahili to Standards 1-3. (3) The relevant period for BL is 2012-2013, for ML 2014-2015 and for EL 2016-2017.

Contamination risk: baseline to midline

Between baseline and midline, the only programme identified as a contamination risk is LANES. It carried out a sub-set of activities in the study control districts: one-off in-service curriculum orientation training for Standards 1 and 2 teachers on the new Standards 1 and 2 curriculum, and one-off SLM training for head teachers and WEOs. The midline impact evaluation survey confirmed that teachers and head teachers in the study control areas had received this training.

Table 5 shows that about 50% of early grade teachers in control schools reported receiving training from LANES at midline. Both of the LANES training activities were one-off events—in the case of the teachers, this was a 10 day residential training held in Dodoma—with no follow-up school-based training. The likely contamination risk from these LANES activities was discussed at the evaluation's baseline reference group meeting, which included representatives from government ministries, departments and agencies responsible for education in Tanzania. The view from that meeting was that these initial LANES interventions were unlikely to pose a serious contamination risk, as the likely impact on pupil learning would be minimal (too dilute) prior to the midline round of the impact evaluation, without further inputs (OPM 2017b).

Contamination risk: midline to endline

Between midline and endline, the vast majority of LANES activities have affected all schools across the country, and hence the likely contamination risk is low (since any effects are likely to be similar in treatment and control schools). This includes curriculum orientation training for teachers on the new Standards 3 and 4 curriculum. The endline evaluation survey confirms that early grade teachers in both treatment schools (Table 4) and control schools (Table 5) received LANES training. The main

exceptions are one-off training for WEOs with financial support for school visits, and one-off training for school committees.

By contrast, Tusome Pamoja is likely to have contaminated the EQUIP-T impact estimates to some extent. This programme operates in Ruvuma, which contains one of the impact evaluation's control districts. Tusome Pamoja's interventions include a sub-set of similar activities to the EQUIP-T programme, notably in-service training for Standard 1 and 2 teachers on the new 3Rs curriculum, and provision of reading books for pupils. The endline survey confirms that about 7% of early grade teachers in control areas received training from Tusome Pamoja (Table 5)—and in fact this is confined to Ruvuma where 86% of early grade teachers reported being trained by Tusome Pamoja.

In assessing the extent of contamination from Tusome Pamoja in the impact estimates, it is important to bear in mind that by endline EQUIP-T had been operating for close to four years, while Tusome Pamoja had only been implementing at school level for about one and a half years. In addition, EQUIP-T has delivered a broader and more intensive set of interventions at school level than Tusome Pamoja had by endline. These factors suggest that the impact of EQUIP-T on pupil learning by endline is highly likely to outweigh any initial impact of Tusome Pamoja on pupil learning.

The EPforR programme is a national programme. It operates via a results-based financing mechanism whereby a group of development partners (DFID, World Bank and SIDA) reward the Government for achieving a set of disbursement-linked results on an annual basis. The targets relate to strengthening the overall education system, and, as such, do not pose a particular contamination risk for the EQUIP-T evaluation. However, the precursor to this programme, BRN-Ed, had school-level components which were implemented in a selection of districts. Indeed the sampling strategy for the EQUIP-T evaluation took care to exclude districts with known BRN-Ed school activities (see Section 3.2). This did not prevent contamination entirely however, perhaps partly due to teacher transfers. Early grade teachers in both treatment and control schools report receiving training from BRN-Ed or STEP (which was a teacher training programme under BRN-Ed)—see Table 4 and Table 5. While the prevalence of BRN-Ed or STEP training among the surveyed early grade teachers is very low at both baseline and midline, by endline 4% of treatment teachers reported receiving BRN-Ed compared with 3% of control teachers; while 13% of treatment teachers said they were trained by STEP compared with 6% of control teachers. This suggests that BRN-Ed and STEP activities have affected both treatment and control schools, to a roughly similar degree. Assuming therefore that any effects on pupil learning from these programmes are roughly similar in the treatment and control areas, the likely contamination is fairly minimal.

The main implication of the contamination risks identified above is on the interpretation of the EQUIP-T impact estimates. This is discussed next.

Implications of contamination for interpretation of the impact estimates

The discussion on contamination risks above suggests that LANES (baseline to midline) and Tusome Pamoja (midline to endline) are likely to have contaminated the EQUIP-T impact estimates to some extent because of their operations in the impact evaluation control districts. This has implications for the interpretation of the impact estimates presented in this report. Specifically, the impact identified by the analysis represents the effect that EQUIP-T as a package has had on outcome indicators, compared to a counterfactual situation where, in the same schools, the alternative LANES and Tusome Pamoja training and materials have been delivered. In other words, the analysis measures the added impact of all EQUIP-T-related interventions over and above the potential effect of the other LANES and Tusome Pamoja initiatives. If certain assumptions hold (Box 4) then the full impact of EQUIP-T may in reality be slightly higher than the estimates in this study.

Box 4: Assumptions related to the interpretation of the impact estimates

Although there are reasons to assume that the extent of contamination from LANES and Tusome Pamoja in the evaluation's control districts is fairly minimal, it is also reasonable to assume that if these interventions have had any effect on the outcomes being measured in this study (such as pupil literacy and numeracy levels), it is likely to have been positive. Under this assumption, outcome levels in the control group schools are, on average, higher than they would have been in a pure counterfactual situation (with no contamination). This in turn means that the impact of EQUIP-T, which is estimated by comparing treatment schools and control schools over time, is potentially a slight underestimation of the full impact of EQUIP-T as a whole. This ignores the possibility that the effects of the BRN-Ed/STEP training which took place to a limited extent in both treatment and control areas, had a greater positive effect on pupil learning in treatment areas than in control areas. If in fact the latter is true, then this would reduce any potential overestimation of EQUIP-T impact due to LANES or Tusome Pamoja contamination of the control areas.

3.5.2 Confounding theory of change failure with implementation failure

This is the risk that the programme does not have the expected effect because it is not implemented as intended, rather than because its underlying design and theory of change is flawed. For example, key activities don't happen or happen too slowly or that different activities happen. The design of the impact evaluation aims to address this by using the theory of change to frame the analysis (see, Endline Evaluation Matrix (Part I) in Volume I Annex C) in order to understand if changes have happened as anticipated or not, and to explain why. It is important to note, however, that the impact evaluation does not include a process evaluation, so its focus is on how effective the programme has been in meeting its objectives, rather than on the details of receipt of inputs and timing of activities, although the survey does collect basic information on implementation. The impact evaluation relies on information from the EQUIP-T MA in its annual reports on the status of implementation. This documentation gives an overview of the implementation of different components overall each year, and notes any major adjustments to programme design (which in turn has been used by the evaluation team in designing each round of research). The EQUIP-T annual reports have become less detailed over time, however, and the evaluation team have increasingly relied on interviews and email correspondence with EQUIP-T MA staff to fill in gaps related to the nature of interventions and their intended timing and volume. School- ward- or district-level implementation data is not readily available on component activities, and so a detailed analysis of the status of implementation in the areas under evaluation is not possible.

3.5.3 Limitations to the quantitative component

Some of the general limitations of the quantitative component are set out in Table 6 together with explanations of how these limitations have been addressed in the endline design and analysis. Some of the specific limitations to the endline quantitative analysis include:

- *Head teacher absence on day of survey:* In 40 of the 200 schools (20%), the head teacher was absent from school on the day of the survey. This meant that the assistant head teacher or another teacher at the school who is knowledgeable about school records and practices answered all school-related questions in the head teacher interview. Phone interviews were conducted with 38 of these 40 absent head teachers to collect the remaining information such as head teachers' training attendance, teacher management, morale and perceptions of the usefulness of school committee support, community support, and WEO support. At midline in 36 out of the 200 schools (18%) head teachers were absent on the day of the survey.
- *Teacher interviews conducted over the phone:* 11% of teacher interviews were conducted over the phone because these teachers were absent from school or unavailable to be interviewed on the day of the survey. This could have implications for the quality of the data as 30 to 40 minutes is considered a long period for a phone interview and also there are certain modules such as in-

service training attendance that rely on effective probing from enumerators in order to elicit accurate responses from the respondent. Additionally, it was not possible to ask these teachers the questions which required them to show written records (examples of pupil assessment, feedback on lesson observation and lesson plans). A similar share of phone interviews were conducted with teachers at midline (9%).

- *Pupils administered the scorecard questionnaire:* 8% of scorecard interviews were conducted with the pupils themselves as opposed to with their parents, guardians or other adult household members, as the latter were not available even after more than one visit attempt. This had two implications. Firstly, pupils were only asked the questions relating to the household characteristics, and as a result there was high item non-response rates on the other information relating to the education support pupils receive at home and to households' awareness of school noticeboards and PTPs. Secondly, this could have implications on the quality of the data collected as answers from the pupils will be less reliable than answers from adult household members. A similar share of pupils were administered the scorecard questions at midline (7%).
- *Recall bias with the data from the in-service training teacher group questionnaire:* The in-service training group questionnaire collects information on all of the EQUIP-T residential and school-based training sessions that teachers from the school had attended since 2014. The survey finds that only 27% of all school-based training sessions in a school had a record available. This means that the majority of information collected in this interview relies on the memory of respondents. This includes information on the date of the sessions, the topics and participants. In order to reduce the recall bias, the interview was not administered individually to the INCO but rather was conducted as a group interview with the INCO and a small group of other teachers who were knowledgeable about the training activities that had taken place and who themselves had attended the majority of the sessions. During the endline pre-test, this strategy was found to refresh the memory of participants and to improve the accuracy of the collected data.
- *Problems with comparing baseline and endline estimates of certain SLM indicators because of changes in administration:*
 - Teachers were asked to show written examples of their own pupil assessments, written feedback on their lesson plans, and written feedback on a lesson observed by the head teacher. At baseline, these examples were sought during the interview. At midline and endline these questions were asked at the end of the interview, because during the midline pre-test it was observed that requesting these examples during the interview was very disruptive. There is some suggestion from field feedback that some teachers were reluctant to look for evidence at the end of the interview because they wanted the interview to be finished. This may also have affected how they answered the previous questions on whether these actions had taken place (for example, answering 'Did you receive feedback from the head teacher on your lesson plans in the last 30 days?'). It is difficult to unpick the possible effect of this change in administration, but it means that baseline and endline results are not strictly comparable.
 - There was also some ambiguity in the meaning of the terms 'lesson observation' and 'written feedback' between baseline and midline. Although the wording of the related questions did not change between these rounds, the training of enumerators at midline emphasised more precise definitions of these terms, and that they needed to probe respondents to ensure that they captured the information accurately. This may have compromised the comparison between baseline and midline for indicators of lesson observation and written feedback to some extent. Between midline and endline, the training of enumerators on these terms was the same, and in addition the terms were defined in the questions. The most reliable estimates are likely to be those captured at endline, and the trends need to be interpreted with some caution.

Table 6: Limitations to the quantitative component of the impact evaluation and mitigating factors

Possible limitation	Why is this limiting and what mitigating factors were taken?
EQUIP-T regions and districts were purposively selected to target those performing weakly on selected education indicators	An RCT design was not possible for the impact evaluation due to purposively selected treatment regions and districts. A quasi-experimental PSM-DID approach was chosen instead to establish an appropriate counterfactual to assess EQUIP-T impact. This relies on the assumptions of PSM to mimic the experimental approach. A key assumption of PSM is that the information on observables is sufficient to match the control and treatment groups for the purposes of the evaluation. If the groups are matched on observables, but differ on unobservable and time-variant characteristics that affect the impact indicators, the estimate of impact will not be robust. <i>See Chapter 4 for further explanation of ways this risk has been minimised.</i>
Language spoken at home is not the language of instruction	Pupils that do not speak Kiswahili at home (as a main language) may be systematically disadvantaged by pupil tests conducted in Kiswahili. <i>Diagnostic tests on the endline pupil test data did not find any substantial differential item functioning related to home language.¹</i>
Not possible to substantially change survey instruments after the baseline	If there are substantial changes to the EQUIP-T programme design after the baseline the instruments may not be able to measure this. The indicators included have been carefully considered to ensure they capture key EQUIP-T outcomes and outputs as per the original design. <i>The endline survey was adjusted to accommodate some of the key changes and additions in the EQUIP-T programme design such as the introduction of the new 4B component, COL for teachers and head teachers and school information system. However, there are limits. For example, trend analysis of indicators related to component 4B is not possible and there are also other design features of the programme that could not be assessed such as the School Readiness Programme (although this was included to some extent in the midline qualitative research).</i>
The number of teachers per school is small in the control and treatment districts	The total sample of teachers was smaller than originally anticipated with implications for the power of detection of impact. A larger school sample size would have been required to address this issue but was not deemed possible by DFID for cost reasons. <i>See baseline evaluation report volume II (OPM 2015b, p19) for notes on this risk. See midline evaluation report volume II (OPM 2016b) for explanation of how the small sample sizes affect the impact estimation.</i>
Inaccurate identification of EQUIP-T interventions by respondents	Respondents do not always know the official name of programme interventions or that they come from EQUIP-T. Multiple names for the same intervention may be in use, and there is the possibility of respondents mixing up EQUIP-T interventions with other similar development interventions. <i>The instruments were carefully pre-tested at midline and endline, and some questions were adjusted to deal with naming confusion which arose at this design stage. Similarly during enumerator training and piloting many school visits took place to practice the survey protocols and to review data collected. Daily debriefs were held, and changes to the instruments and training manual were made as appropriate. Enumerators were also trained on the specifications of the most common trainings that had taken place in the impact evaluation regions such as EQUIP-T, LANES and Tusome Pamoja, in order to be able to effectively probe during the interview. This included information on where and when the training took place, and who delivered the training. The latter was important as some respondents reported the name of the training by the organisation or agency that delivered the training such as ADEM as opposed to the provider of the training.</i>
Notes: (1) The differential item functioning tests were carried out as part of a Rasch analysis of the pupil test data (see Annex E for information on measurement of pupil learning using the Rasch model of item response).	

Part E Supplementary evidence



4 Impact estimates

This Chapter explains the measurement approach taken to impact estimation and presents the detailed results.

4.1 Impact identification strategy

A rigorous identification of programme impact in quantitative studies relies on the idea that such impact can be defined as the difference in the outcomes measured among individuals that participate in a programme compared to the outcomes measured among the same individuals in a theoretical state of the world where the programme is not implemented but where everything else, except the programme, stays the same. This is normally referred to as the counterfactual and, because it is purely hypothetical, the key challenge that impact evaluations face is to find alternative observed counterfactual measures that can credibly be used to approximate this hypothetical counterfactual as closely as possible and thus infer programme impact.

A Randomised Controlled Trial (RCT), whereby subjects are randomly assigned to a treatment and control group, is commonly considered as one of the most robust designs to deal with the problem of the counterfactual. Because treatment assignment is implemented randomly in these trials, individuals from control and treatment groups are, on average, the same. This means that after the implementation of the programme, averages of outcomes measured among participants and non-participants can be compared directly and differences can be attributed to the programme, rather than any other confounding factors. Sometimes, however, implementing an RCT is not feasible or not appropriate. Alternative identification strategies use econometric modelling techniques to try to come as close as possible to replicating the situation of such an experimental design.

This was the case in the present evaluation, where an RCT was not feasible and schools were assigned to participate in the programme based on programme management decisions and some pre-defined characteristics. Control schools were selected to match those characteristics.⁸ Specifically, the quantitative impact identification methodology used in this study, both at midline and endline, follows a quasi-experimental design that combines two approaches: Propensity Score Matching (PSM) and Difference in Difference (DID) analysis. This is applied in the context of an evaluation design based on a panel of schools and on repeated cross-sections of pupils and teachers. Combining PSM and DID takes advantage of the strengths of both of these methods in order to robustly estimate the difference in key impact indicators across treatment and control schools that can be attributed with statistical confidence to EQUIP-T. The following sections describe how PSM and DID were combined in the current endline evaluation to obtain our impact estimates. Please refer to Sections 6.2 and 6.3 of the midline impact evaluation volume II (OPM 2017b) for a detailed discussion of the assumptions behind PSM and DID as well as a detailed explanation of the approach used to implement the estimation models.

4.2 Combining DID and PSM

In this study, two different strategies have been used to combine PSM with DID:

1. Directly comparing ATT estimates at endline and baseline across time.
2. Matching treatment observations across time to construct a pseudo panel⁹ of treatment observations and to construct an overall ATT estimate using this pseudo panel only.

⁸ Note that the term 'control group' is used throughout this document to refer to the quasi-experimental comparison group.

⁹ We refer to this as a 'pseudo' panel since it is constructed by matching repeated cross-sections of pupils and teachers.

ATT refers to the Average Treatment Effect on the Treated. In a PSM estimation, outcome indicators from treatment units (that is, programme school teachers and pupils) are compared to outcome indicators from specific control units based on the propensity score, which is a metric of similarity of treatment and control units. This implies that the estimated average treatment effect is valid for the group of treatment observations only, which, in turn, means that the ATT obtained with PSM cannot be extrapolated to observations (i.e. pupils and teachers) outside the sample.

The **first strategy** was to take a direct difference of baseline and endline estimations of ATTs derived from PSM at baseline and endline. Essentially, this amounts to comparing two estimated treatment coefficients with each other. In theory, ATT estimates at baseline should be close to zero – because EQUIP-T had not started at that time yet. However, this was not always the case, despite good balancing performance of models at baseline. This means that the overall impact of EQUIP-T is defined as the difference that EQUIP-T made in the estimated ATT at endline, compared to the baseline estimate:

$$ATT_{overall} = ATT_{endline} - ATT_{baseline} \cdot (1)$$

Of course, the main goal is to see whether the overall ATT estimate is different from zero or not. Test statistics for the estimate defined in (1) are calculated using the formula for comparing coefficient estimates presented in Paternoster et al. (1998). Using this test statistic, this study then calculates whether the estimated ATT is significantly different from zero or not from a statistical point of view. Note that all standard errors for the endline and baseline ATT used are based on bootstrapping procedures for PSM estimates. (See section 4.4 on why standard errors for PSM are bootstrapped.)

The **second strategy** is a robustness check where additional matching is used to create a ‘pseudo panel’ of treatment observations (that is, teachers and pupils in EQUIP-T schools) across time, given that these have not been panelled and were surveyed as repeated cross-sections. Figure 1 depicts this process graphically.

In a first step, treatment observations from teacher and pupil samples are uniquely matched across the two time periods¹⁰. This is done using a Nearest Neighbour PSM approach without replacement. This means that for each treatment observation at baseline a unique comparator is found at endline. In this way an artificial panel of teachers and pupils is constructed, as each individual teacher and pupil at baseline is associated with a single matched teacher and pupil at endline.

For this ‘pseudo panel’ of treatment observations, values obtained for their respective matched comparisons at baseline and endline are then used to calculate differences between estimated control group and treatment group individuals at baseline and at endline separately, using the same PSM models as in the main estimations. Note that kernel matching at baseline and endline provides, for each treatment observation, an appropriate estimated counterfactual value based on the PSM estimation. This value is used to calculate the first difference between treatment observations and counterfactuals, as part of the double differencing approach underpinning the DID analysis.

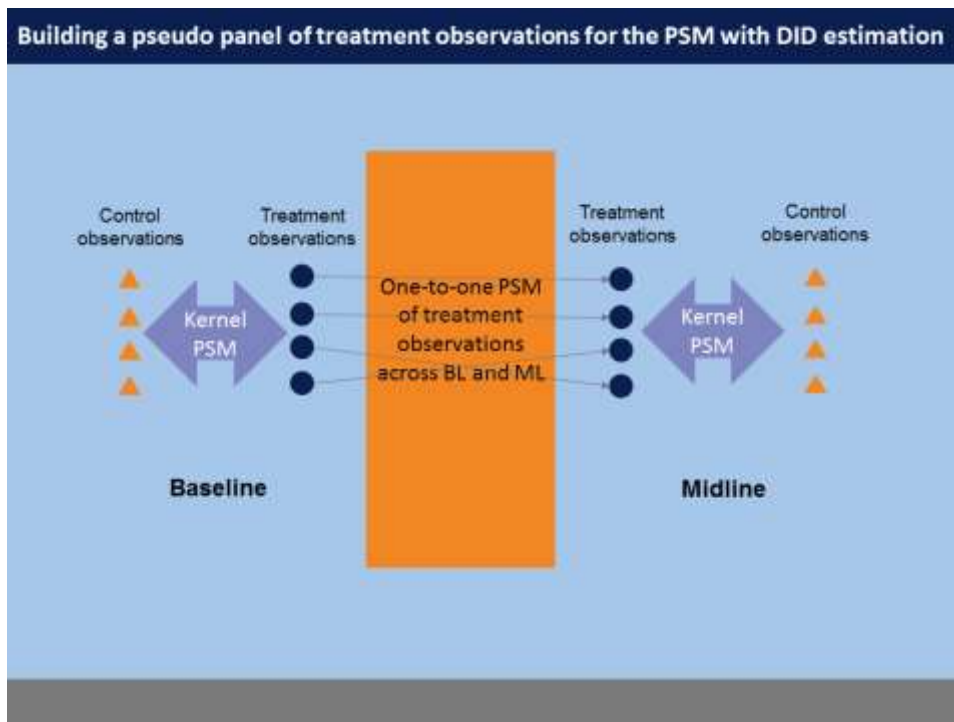
In a final step, those differences are then compared across baseline and endline for the ‘pseudo panel’. The average of this double difference for the pseudo panel is the estimated overall ATT. By implementing this approach, this study follows a suggestion by Blundell and Costa Dias (2000, p. 451). This study is likely to represent the first practical application of this PSM with DID procedure for

¹⁰ For more detailed information on how relevant teacher and pupil characteristics were selected to match treatment and control groups (in both the first and second strategy) please refer to Section 6.2 of the midline impact evaluation volume II (OPM 2017b)

a repeated cross-section, in an education evaluation of teachers and pupils. It was developed at midline and it is replicated here at endline.

The key difference to the first strategy is that this double differencing is implemented only across treatment observations that are similar to each other, as they have been matched one-to-one in the first step. One potentially adverse effect of this strategy is that the sample size used in the impact estimation can be reduced due to the common support requirements of the additional matching (that is, bad treatment matches are dropped from the sample).

Figure 1: Visual representation of second PSM with DID combination



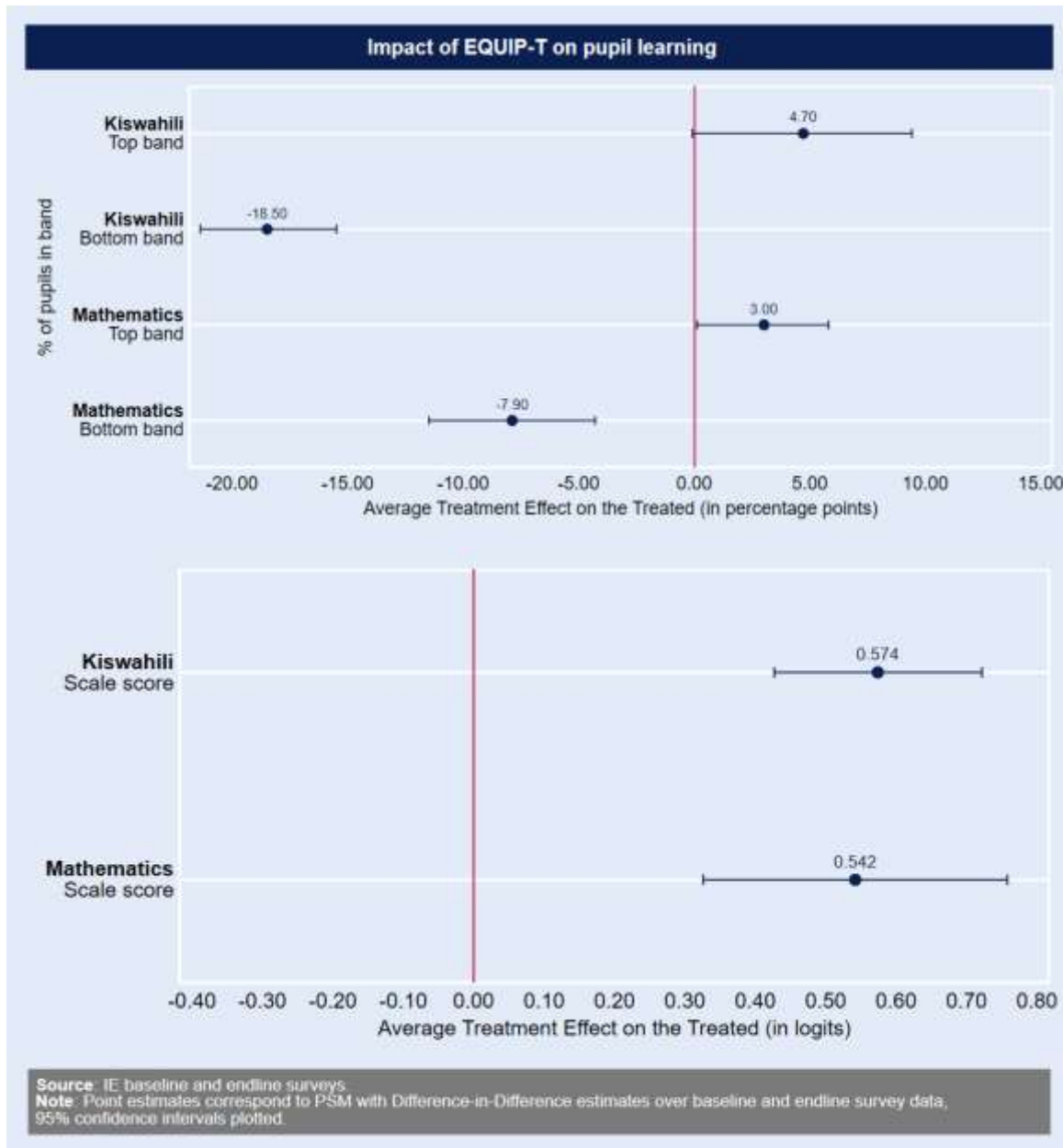
4.3 How results are presented in Volume I

In Volume I, headline results are presented in a visual form, with an explanation underneath each graph. These headline results are the results of the first PSM DID combination strategy. See Figure 2 below for an example. Each graph shows point estimates for treatment effects (ATT) on outcome indicators and 95% confidence intervals for these effects. This means that the probability for the true treatment estimate to fall within this area is 95%.

Outcome indicators used in this evaluation are mostly proportions. When that is the case, estimates of treatment effects are given in percentage point changes of these proportions. For example, if the ATT estimate on the proportion of pupils in the bottom performance band of Kiswahili in treatment schools is -0.183 , this means that EQUIP-T has reduced this proportion by approximately 18 percentage points, compared to a counterfactual of no EQUIP-T package and some alternative teacher training. Equivalently, this can be expressed as a decrease of 18 percentage points in the probability of pupils from treatment schools to fall in this bottom performance band. The only exception is represented by the impact estimates on the Kiswahili and maths test score (scaled as Rasch scores—see Annex E for details on the measurement of pupil learning outcomes using an interval scale). In that case, estimates of treatment effects are given in logits. For example, if the ATT estimate on the pupil Kiswahili Rasch score is 0.574 , this means that EQUIP-T has increased pupils' average maths Rasch score by 0.574 logits, compared to a counterfactual of no EQUIP-T package and some alternative

teacher training. When confidence intervals of such estimates do not overlap with zero, then this is an indication that this treatment effect is truly different from zero. This zero value is indicated using a red line in the graphs.

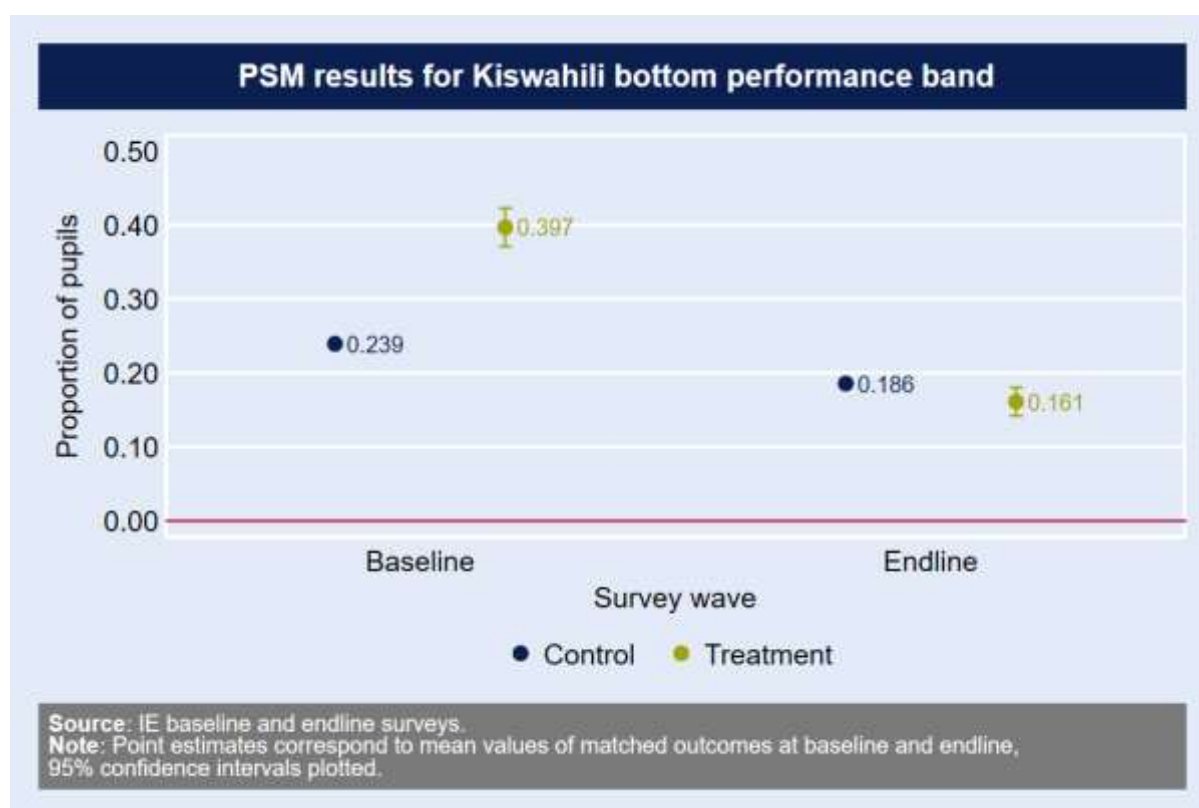
Figure 2: Impact of EQUIP-T on pupil learning



As explained above, the ATT estimates shown here are the results of the first strategy to implement PSM DID estimations, and thus take into account information both from the baseline and the endline data. Comparing Figure 2 to Figure 3 can help to understand this. Figure 3 below shows estimates of averages of the treatment group and of matched counterfactuals at baseline and at endline. Note that the control estimates are not simple descriptive statistics – they are the averages of counterfactual observations constructed using PSM. The PSM DID estimates presented in Figure 2 correspond to the double difference of the averages presented below. The first difference at endline is $0.161 - 0.186 = -0.025$. The same difference at baseline is $0.397 - 0.239 = 0.158$. The double difference estimate is $-0.025 - 0.158 = -0.183$. This corresponds to the ATT estimate presented in Figure 2 (rounding off decimals). When looking at the graph below, one can see that the difference between EQUIP-T and

control schools has effectively decreased over time – this decrease in difference is the ATT estimate and is attributable to EQUIP-T.

Figure 3: Example PSM comparisons



4.4 Caveats - Addressing weaknesses in the analysis

Four key caveats related to the present estimation strategy need to be mentioned here. First, PSM only controls for observable characteristics that cause selection bias. This is a problem for any impact identification strategy that relies on controlling only for factors (variables) that can be observed in the data, not only PSM. PSM helps addressing this by allowing for extensive balancing checks after matching, which can provide substantial evidence for the fact that balance is achieved across a large variety of characteristics and, by implication, is likely to also extend to unobservables. In this study, such extensive balancing checks were implemented. Results are presented in Section 4.5 below. In addition, the DID strategy implemented in the present case helps to control for remaining imbalances that may be due to time-invariant unobservable variables.

Second, DID helps to deal with time-invariant imbalances, but not time variant ones. This means that only time-invariant imbalances that remain after PSM would be controlled for, in contrast to imbalances that vary over time. In the present case, this is addressed by extensive balancing tests, which show little remaining covariate imbalance in general after PSM, by showing that results are robust to a variety of different PSM specifications, and by showing that results are robust to two separate DID strategies used. Together, this evidence implies that results are robust, remaining imbalances are small, and results are unlikely to be sensitive to or to be driven by such imbalances, even if they were time variant.

Third, as already discussed in Section 3.5.1 on contamination risks in Chapter 3, over the course of the evaluation period in-service teacher training and SLM training for head teachers (as well as some other initiatives to improve education quality) have been implemented not only in EQUIP-T schools,

but also in control schools. In the two years following the midline analysis (2016 and 2017), the LANES programme has continued and a new large-scale programme, Tusome Pamoja, has begun. It seems reasonable to assume that additional contamination risk from LANES activities that have taken place since the midline impact evaluation is minimal, because the vast majority of activities have affected all schools, including EQUIP-T schools and schools in the control group. However, Tusome Pamoja's interventions include a sub-set of similar activities to the EQUIP-T and the new programme operates in Ruvuma, which contains one of the impact evaluation's control districts. One of the main reasons for conducting the quantitative survey part of the impact in 2018, rather than waiting until 2019, was to try to minimise any contamination effects of Tusome Pamoja. In any case, the presence of these competing initiatives in the evaluation areas means that the impact identified is the effect that EQUIP-T as a package has had on the outcome indicators compared to a counterfactual situation where in the same schools the alternative training from control schools would have been implemented. The PSM DID approach still ensures that the treatment and control groups are comparable and allows identification of the marginal impact attributed to EQUIP-T and thus its added value. Box 4 in Section 3.5.1 (Chapter 3) explains more about the assumptions under-pinning the interpretation of the impact estimates in light of the contamination from other programmes.

Finally, calculating standard errors of estimated treatment effects using PSM methods is not straightforward. As Caliendo and Kopeinig (2005, p. 18) put it, 'The problem is that the estimated variance of the treatment effect should also include the variance due to the estimation of the propensity score, the imputation of the common support, and possibly also the order in which treated individuals are matched'. These estimations increase the variation of the treatment effect estimates over and above normal sampling variation. In the literature, there is no consensus on how to take this into account.

A popular approach to solve this problem is to bootstrap standard errors for the estimated treatment effect (see Lechner 2002). Each bootstrap draw re-estimates both the first and second stages of the estimation. This produces N bootstrap samples for which the ATT is estimated. The distribution of these means approximates the true sampling distribution, and therefore the standard errors of the population mean (Caliendo and Kopeinig 2005, p.18). This study followed this approach both at midline and at endline and implemented bootstrapping, using 200 repetitions, to estimate the standard errors of the estimated treatment effects. Note that, for the sake of completeness, this report shows both the bootstrapped and the non-bootstrapped standard errors below.

It is also important to note that there is no clear direction in which estimated standard errors should change due to bootstrapping. On the one hand, the additional variation taken into account should increase standard errors. On the other, bootstrapping generally makes estimates more precise, which tends to decrease standard errors. Overall, the direction of the change is not uniform. In fact, the results show that, with bootstrapping, standard errors in some instances are smaller and in some larger than without bootstrapping.

4.5 Results

This section presents the results obtained from applying PSM to EQUIP-T baseline and endline data. In the following paragraphs, the balancing results, the ATT estimates and the PSM-DID estimates described for all impact indicators for the main strategy and the robustness check¹¹. The following indicators were analysed in the context of this evaluation:

¹¹ It is important to highlight the fact that a large range of results were produced in the course of the analysis across a range of different models, including varying levels of trimming and bandwidth size for the kernel matching algorithm. This extensive investigation of alternative specifications provided the opportunity to select the most appropriate and robust estimation

Table 7: Impact indicators for PSM-DID estimation

Impact area	Impact indicators	Sample for the impact evaluation
Pupil learning	Proportion of pupils in the top performance band of the interval scale ¹ for Kiswahili	Standard 3 pupils who were assessed
	Proportion of pupils in the bottom performance band of the interval scale for Kiswahili	
	Proportion of pupils in the top performance band of the interval scale for Mathematics	
	Proportion of pupils in the bottom performance band of the interval scale for Mathematics	
	Rasch Scores for Kiswahili	
	Rasch Scores for Mathematics	
Teacher absenteeism	Proportion of teachers who were absent on the day of the survey	All teachers (from roster)
	Proportion of teachers present on the day of the survey, timetabled to teach before lunch and absent from the classroom	
School leadership and management	Proportion of teachers who report participation in performance appraisals	Interviewed teachers of Standards 1-3

Note: (1) Annex E explains how the scales for Kiswahili skills and maths skills were developed. The method applies the Rasch model of item response (the simplest item response theory model) to the pupil test data, and generates estimates of pupil ability (performance) and item difficulty on a common interval scale (under assumptions that the data satisfies the key properties of the model).

For each of the outcome variables, this study implemented two PSM DID strategies, one main strategy and a robustness check outlined in Section 4.2.

Presentation of results

For each outcome variable, three sets of results are presented in this volume:

- the second stage results;
- the propensity score matched outcomes at baseline and endline; and
- the PSM-DID estimates.

The following paragraphs use the example of Figure 13 to explain the interpretation of the results in detail. The rest of the results are then presented in graphical form.

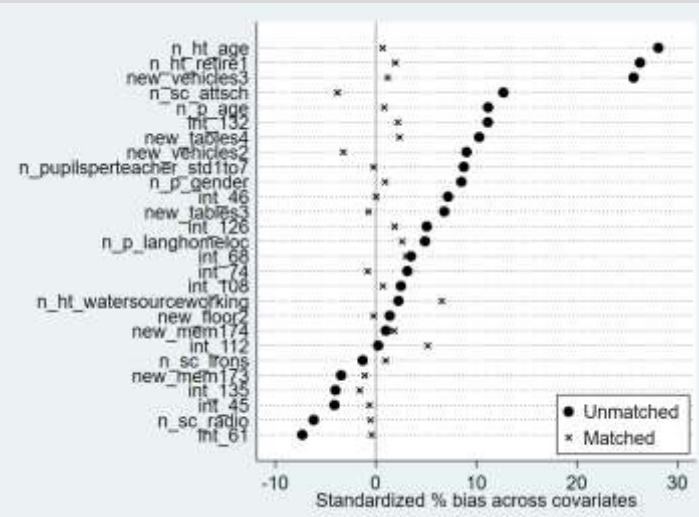
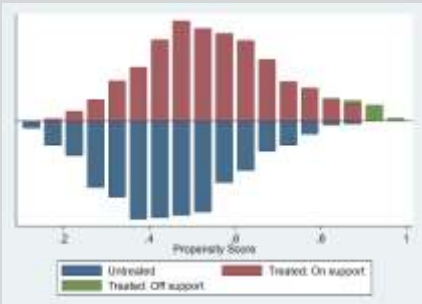
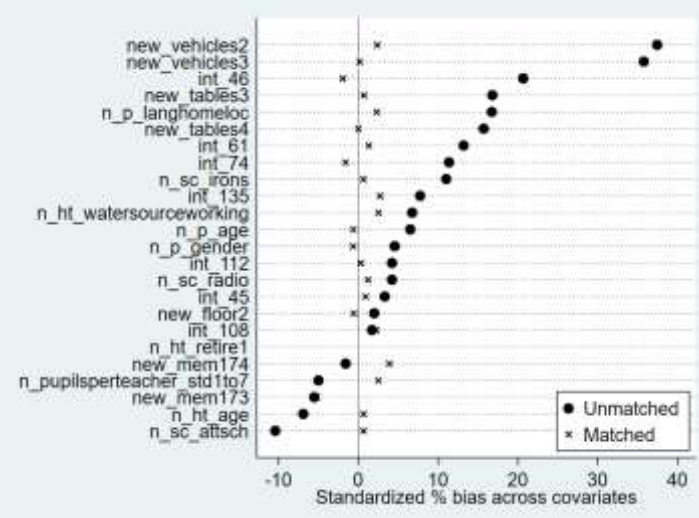
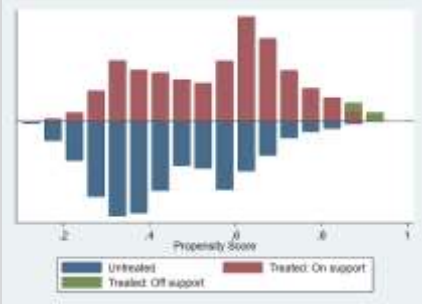
First, the second stage results for the main strategy are presented, as illustrated in Figure 13 for the indicator on top performance band for Kiswahili. The figure is divided into two panels; the top panel and the bottom panel show baseline and endline results respectively. The format for each panel is as follows:

strategies for which results are presented in this report. At the same time, consistency or inconsistency in the direction and significance of results emerging from this range of models help determine whether any findings on impact (or lack of thereof) can be considered conclusive or yet inconclusive.

- The first graph on the left-hand side indicates how individual variables balance before and after matching. The x-axis displays the standardised bias, which is the percentage difference of the sample means in the treated and non-treated (unmatched or matched) subsamples as a percentage of the square root of the average of the sample variances in the treated and non-treated groups (Rosenbaum and Rubin, 1985). In Figure 13 below, for example, the unmatched samples display large imbalances with standardised bias being present across many of the covariates of interest. However, once matching takes place, the standardised imbalances are diminished.
- The second graph, on the right-hand side, shows the distribution of propensity scores across treatment and control groups. This graph visually confirms that, after dropping observations that are off common support, both treatment and control groups contain observations with propensity scores across the full range of the distribution, which is an indication for overall balance. Although the distributions of propensity scores across treatment and control groups would ideally be symmetric, the presence of some level of skewness does not put at risk the estimation procedure, as indicated by the balance achieved for each covariate and the values of Rubin's R and B, which are tests on the overall balance achieved between treatment and control groups, after matching.
- The remaining rows on the right-hand side display information related to the PSM model. The bandwidth and level of trimming for the optimal PSM model can be found in the first two rows. For example, the optimal model has a bandwidth of 2 and a trimming value of 3 for the baseline sample in Figure 13. This is then followed by the number of observations on common support in the next row, and then the Rubin's R and Rubin's B values both before and after matching. Generally, a Rubin's B score under 25 after matching is desirable, whilst a Rubin's R score between 0.8 and 1.25 is the preferred range after matching (Rubin 2001). The unmatched samples are particularly unbalanced; for instance, the Rubin's B for the baseline sample and the endline sample is 67.97 and 74.06 respectively. However, the Rubin's B scores after matching, which are all below 25, show how matching removes the previous imbalances.
- Finally, the remaining rows on the left-hand side indicate the ATT for each corresponding survey wave and the associated standard errors. Given that it is not definitively clear how to produce standard errors for PSM, both bootstrapped and non-bootstrapped standard errors are presented for robustness purposes. (See Section 4.4 for more detail on this.)

Proportion of pupils in the top performance band for Kiswahili

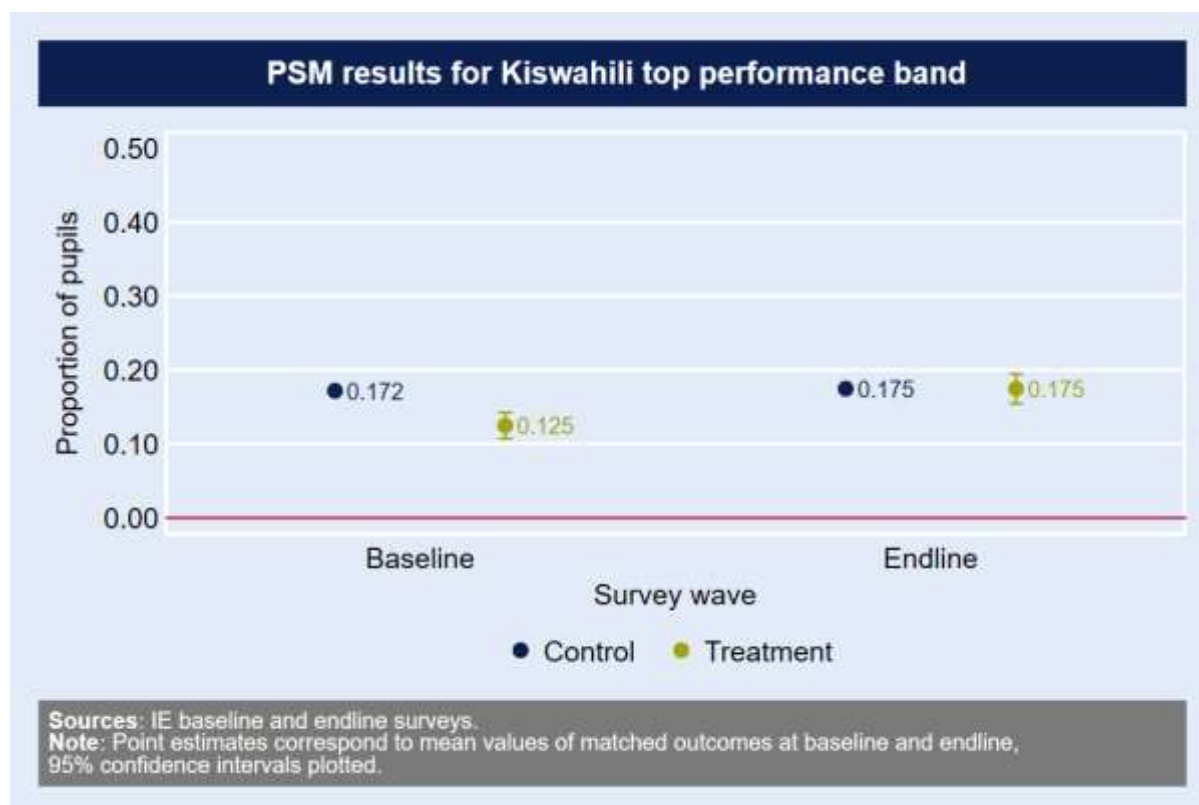
Figure 4: Kiswahili top band: Second stage results (Main strategy)

Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
 <p>A scatter plot showing standardized % bias across covariates for baseline. The x-axis ranges from -10 to 30. The y-axis lists 25 covariates. Matched points (marked with 'x') are clustered around zero bias, while unmatched points (marked with '•') show a clear positive bias, increasing from left to right across the covariates.</p>		 <p>A propensity score plot for baseline. The x-axis is Propensity Score from -2 to 1. The y-axis represents density. The plot shows three distributions: Unreated (blue), Treated: On support (red), and Treated: Off support (green). The red distribution is centered around 0, overlapping with the blue distribution.</p>	
		Bandwidth	2
		Trimming	3
		N on common support	2780
		Rubin's B	[before matching] 67.97
		Rubin's R	[before matching] 1.33
ATT	-0.047	Rubin's B	[after matching] 13.04
SE (bootstrapping)	0.017	Rubin's R	[after matching] 1.15
SE (no bootstrapping)	0.015		
Endline			
 <p>A scatter plot showing standardized % bias across covariates for endline. The x-axis ranges from -10 to 40. The y-axis lists 18 covariates. Matched points (marked with 'x') are clustered around zero bias, while unmatched points (marked with '•') show a clear positive bias, increasing from left to right across the covariates.</p>		 <p>A propensity score plot for endline. The x-axis is Propensity Score from -2 to 1. The y-axis represents density. The plot shows three distributions: Unreated (blue), Treated: On support (red), and Treated: Off support (green). The red distribution is centered around 0, overlapping with the blue distribution.</p>	
		Bandwidth	2
		Trimming	3
		N on common support	2813
		Rubin's B	[before matching] 74.06
		Rubin's R	[before matching] 1.28
ATT	0.00	Rubin's B	[after matching] 9.89
SE (bootstrapping)	0.018	Rubin's R	[after matching] 0.99
SE (no bootstrapping)	0.016		
DID Estimate	0.047		
p-value (bootstrapping)	0.05		
p-value (no bootstrapping)	0.03		

Second, the mean values of the matched outcome and associated confidence intervals at baseline and endline for the treatment group and the control group are plotted. An example can be seen in

Figure 5 for top performance band of Kiswahili. For the treatment group, the mean of the outcome variable is plotted for observations on common support. For the control group, the mean of the counterfactual outcome estimated by the matching algorithm is plotted here.

Figure 5: Kiswahili top band: Matched outcomes at baseline and endline



Finally, the PSM DID estimates for both the first and second strategies are presented, along with the associated bootstrapped and non-bootstrapped p-values. See Table 8 below as an example of how the overall impact results should be interpreted across the two strategies. In that table, the PSM DID estimate for strategy one shows a statistically significant positive marginal impact of EQUIP-T (that is, the proportion of pupils in the Kiswahili top performance band increases by around five percentage points due to EQUIP-T). The PSM DID estimate from strategy two (robustness check) confirms the positive impact of EQUIP-T, showing a slightly larger and more significant increase in the proportion of pupils in the top band.

Table 8: Kiswahili top band: PSM-DID estimate

	Strategy 1	Strategy 1
PSM-DID estimate	0.047	0.07
P-value (bootstrapping)	0.05	0.00
P-value (no bootstrapping)	0.03	0.00

The balancing results for strategy two, whereby treatment observations across the two survey waves are matched (Treatment vs treatment), are also summarized at the end for each outcome indicator, as illustrated in Figure 6 below. This figure shows that the balancing properties for this matching process concerning this particular indicator were ideal, note that Rubin's B moves from 81.21 to 9.61. This robustness check strengthens the finding from our main strategy, which drives our impact narrative, that EQUIP-T has a significant impact on this outcome indicator.

Figure 6: Kiswahili top band- Second stage results (Strategy 2)

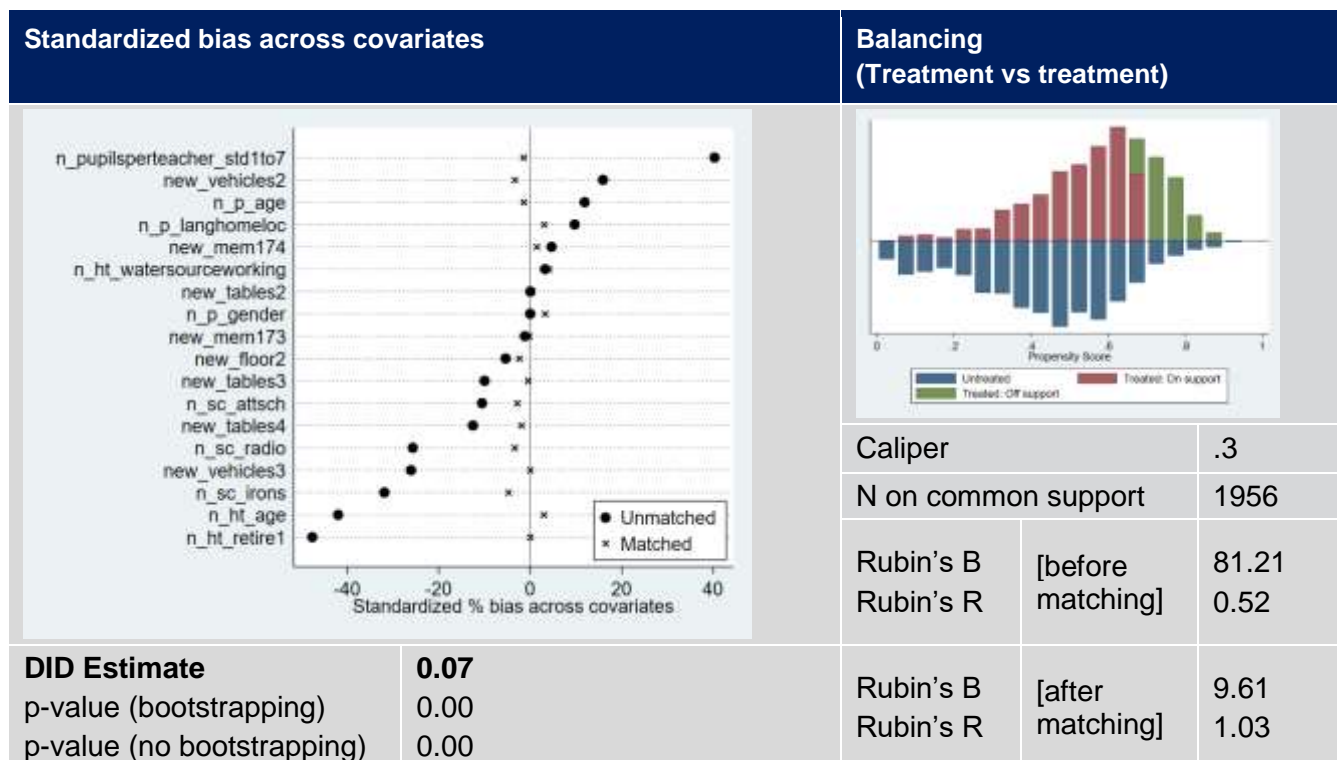


Figure 6 above also shows that the number of treatment pupils on common support (that is, pupils for which an equivalent match was found and can therefore be used in the impact estimation) after applying strategy two is considerably lower than for strategy one. The 1,956 treatment pupils on common support represent the sum of 978 baseline pupils and 978 endline pupils matched one-to-one. This is equivalent to 31.51% of pupils being off support at endline for this indicator. Similar proportions of pupils off support are recorded across all other pupil indicators for strategy two. For strategy one, the proportion of pupils on common support is always over 90%, thus not problematic.

However, given the sizeable proportion of off common support for strategy two (treatment-to-treatment matching), we have undertaken an additional sensitivity analysis to compare the characteristics of treatment pupils on support with those of treatment pupils off support. We have specifically focused on the socioeconomic characteristics of the pupils in the two categories. All results from strategy two across pupil indicators reinforce (from a magnitude as well as significance perspective) the results from strategy one. The aim of the additional sensitivity analysis is to determine whether this is due to the fact that pupils on common support are systematically less poor than those off common support.

When controlling for other variables, including those used for matching, we find no statistically significant correlation between poverty (as measured by poverty score) and common support status for the indicator in Figure 6. Generally, no strong correlation is found between poverty and common support status across any of the other pupil indicators either. We only detect some weak correlation (10% significance) for the maths bottom band indicator sample at baseline and for the Kiswahili bottom band indicator sample at endline. In both cases, pupils on common support are found to be slightly poorer than those off support. However, these are weak correlations in statistical terms.

It seems reasonable to conclude from these results that the positive impact of EQUIP-T is not driven by the socioeconomic conditions of the pupils analysed. On the one hand, socioeconomic variables are used for matching and so controlled for in the estimation of programme impact; on the other hand, pupils on common support used as part of the analysis for strategy two, which tends to reinforce the positive estimates of impact emerging from strategy one, are not found to be systematically wealthier than those off support. The more detailed findings of this sensitivity analysis are available on request.

Proportion of pupils in the bottom performance band for Kiswahili

Figure 7: Kiswahili bottom band: Second stage results (Strategy 1)

Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
		Bandwidth	2
		Trimming	3
		N on common support	2780
		Rubin's B [before matching]	67.61
		Rubin's R [before matching]	1.364
ATT	0.158	Rubin's B [after matching]	13.35
SE (bootstrapping)	0.009	Rubin's R [after matching]	1.25
SE (no bootstrapping)	0.02		
Endline			
		Bandwidth	2
		Trimming	3
		N on common support	2843
		Rubin's B [before matching]	70.83
		Rubin's R [before matching]	1.31
ATT	-0.027	Rubin's B [after matching]	7.62
SE (bootstrapping)	0.012	Rubin's R [after matching]	1.25
SE (no bootstrapping)	0.016		
DID Estimate	-0.185		
p-value (bootstrapping)	0.00		
p-value (no bootstrapping)	0.00		

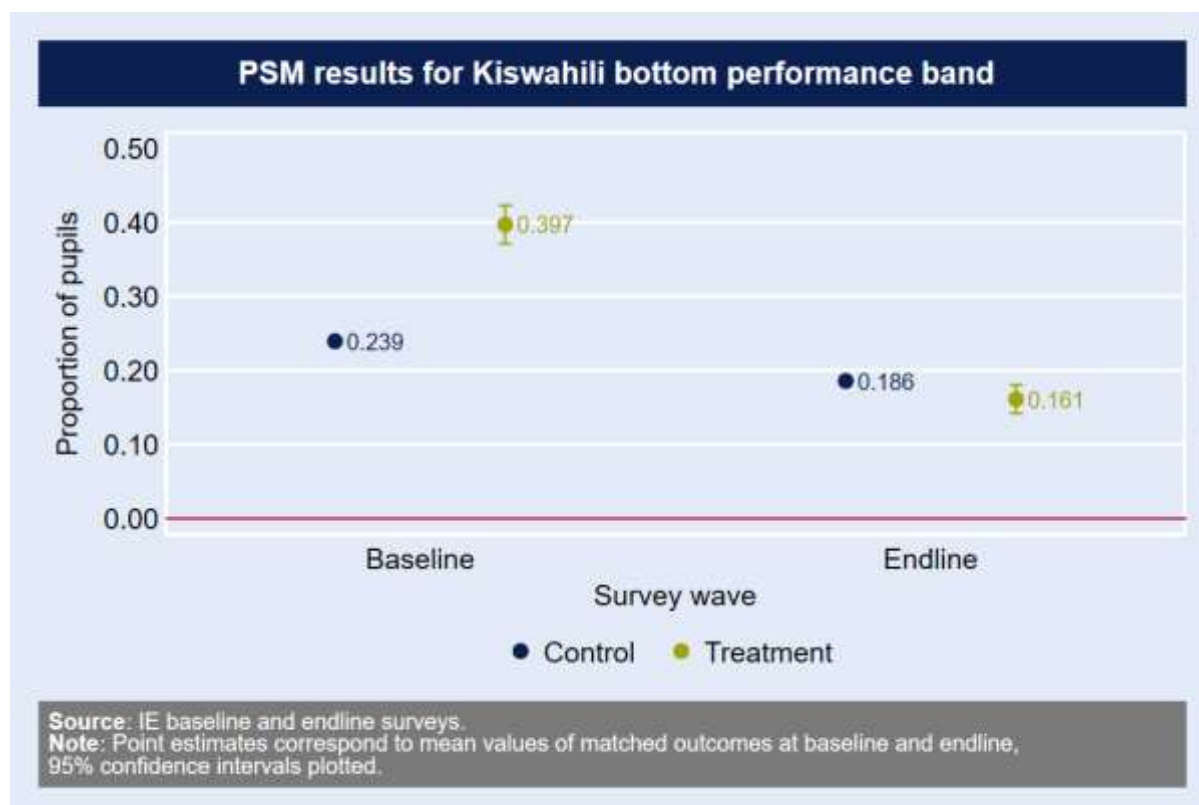
Figure 8: Kiswahili bottom band: Matched outcomes at baseline and endline

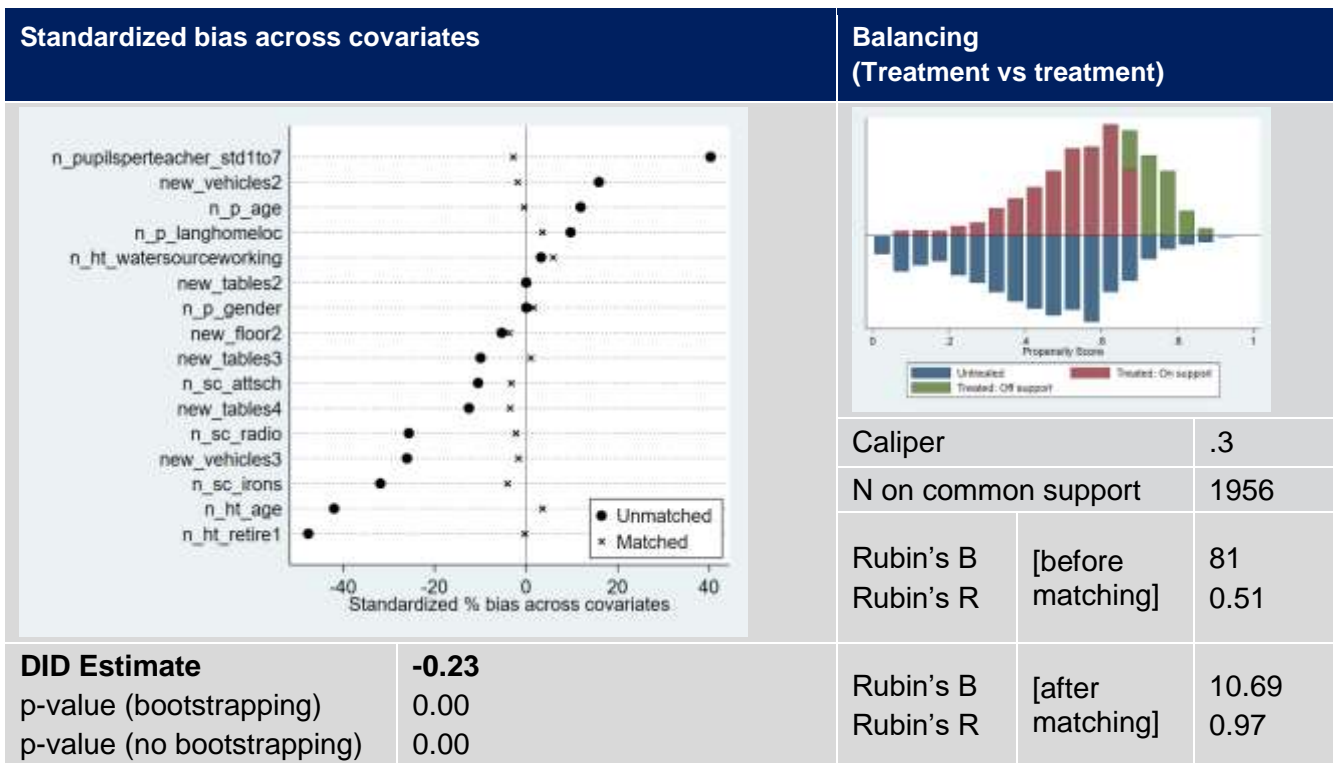
Figure 8 above shows that the PSM analyses at baseline and endline point to a decreasing gap between treatment and comparison schools in terms of pupils who are in the bottom performance band of Kiswahili. Indeed the gap is inverted over time, as the proportion of pupils in the bottom performance band in programme schools becomes smaller than the proportion in control schools at endline. This means that the overall PSM DID analysis finds strong evidence that EQUIP-T has reduced the proportion of pupils in the bottom performance band for Kiswahili in programme schools. See Table 9 below for this. These results remain strong and highly significant across both the strategies.

Table 9: Kiswahili bottom band: PSM-DID estimate

	Strategy 1	Strategy 2
PSM-DID estimate	-0.185	-0.23
P-value (bootstrapping)	0.00	0.00
P-value (no bootstrapping)	0.00	0.00

The balancing results for the robustness check matching across time for treatment observations, presented below, show that for this outcome indicator balancing after matching performs very well. This further strengthens the findings presented above, that EQUIP-T has significantly reduced the proportion of pupils in the bottom performance band for Kiswahili in treatment schools, compared to a counterfactual situation without EQUIP-T.

Figure 9: Kiswahili bottom band- Second stage results (Strategy 2)



Kiswahili Rasch Scale

Figure 10: Kiswahili Rasch scale: Second stage results (Strategy 1)

Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
		Bandwidth	2
		Trimming	3
		N on common support	2794
		Rubin's B [before matching]	55.62
		Rubin's R [before matching]	1.11
ATT	-0.417	Rubin's B [after matching]	6.72
SE (bootstrapping)	0.055	Rubin's R [after matching]	1.32
SE (no bootstrapping)	0.052		
Endline			
		Bandwidth	6
		Trimming	3
		N on common support	2814
		Rubin's B [before matching]	69.69
		Rubin's R [before matching]	1.33
ATT	0.157	Rubin's B [after matching]	11.54
SE (bootstrapping)	0.052	Rubin's R [after matching]	1.23
SE (no bootstrapping)	0.051		
DID Estimate	0.574		
p-value (bootstrapping)	0.00		
p-value (no bootstrapping)	0.00		

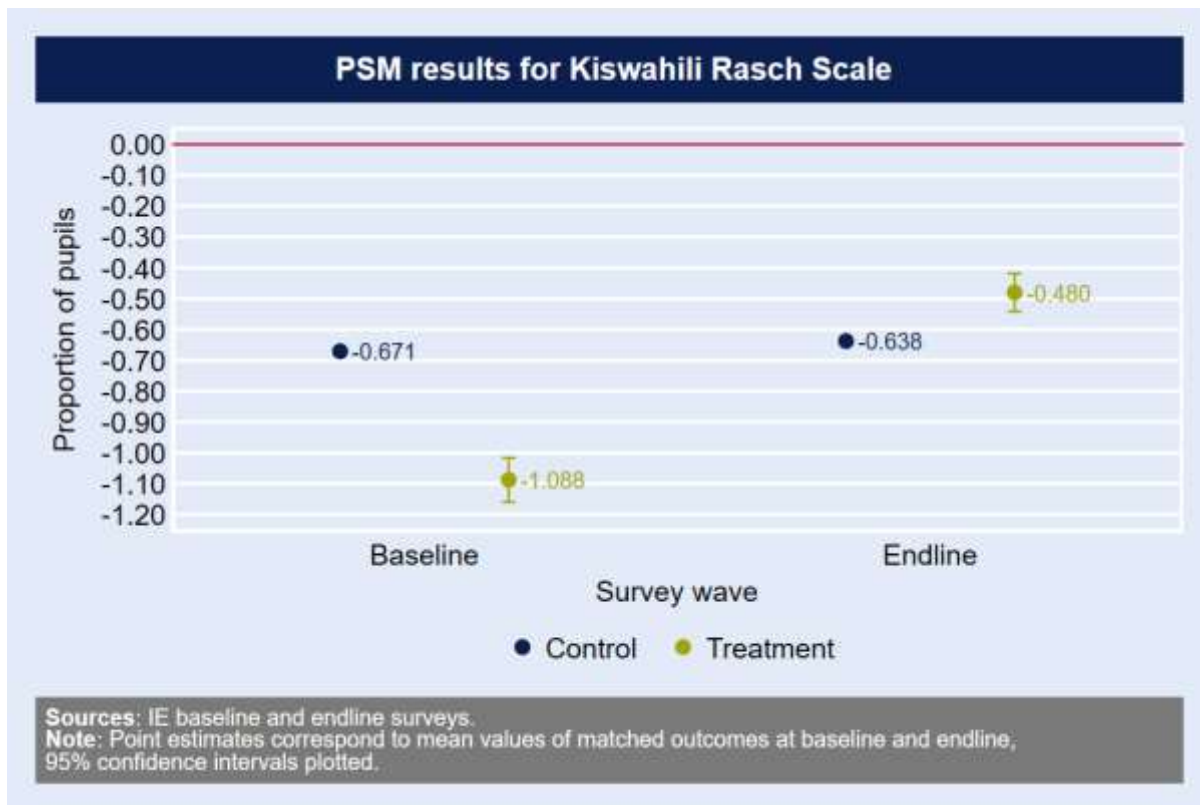
Figure 11: Kiswahili Rasch scale: Matched outcome at baseline and endline

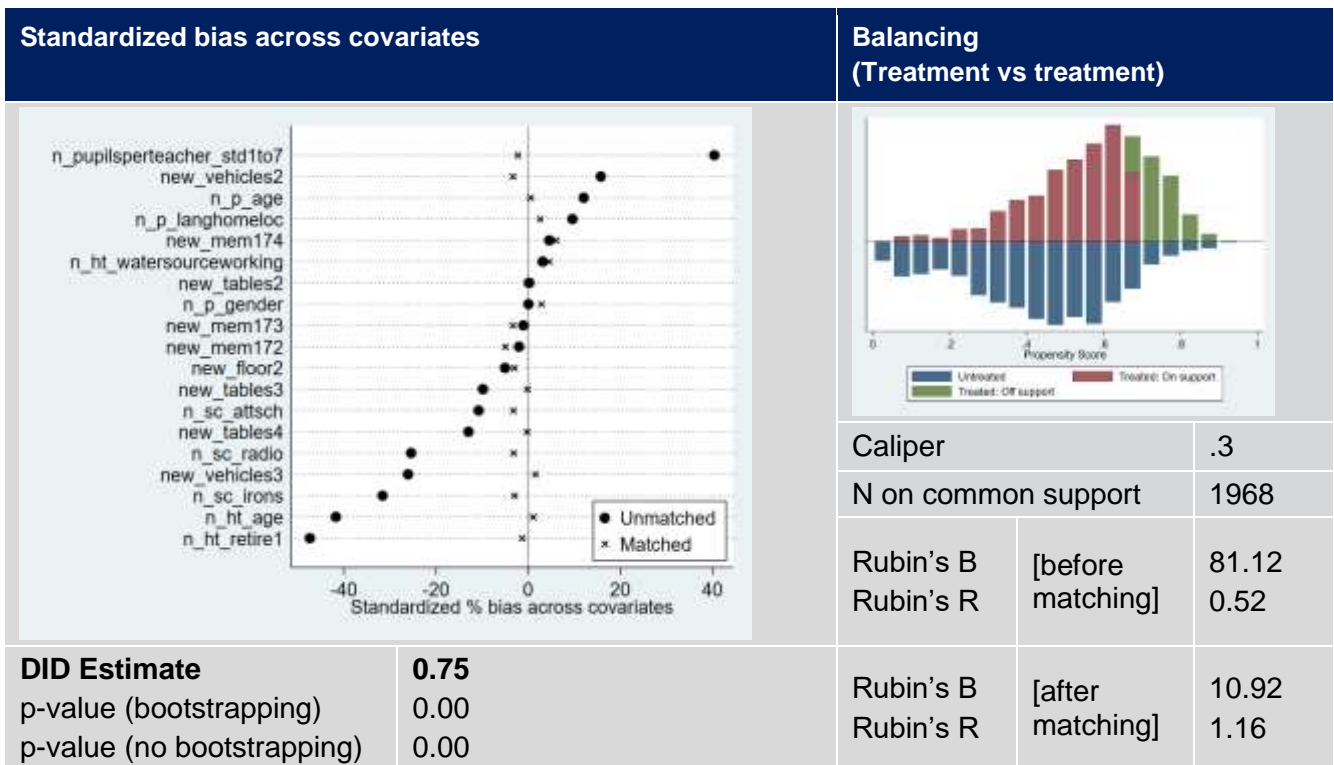
Figure 11 above confirms the Kiswahili top and bottom band findings. There is a statistically significant increase in the average Kiswahili Rasch scores between baseline and endline in programme schools that is attributable to the marginal impact of EQUIP-T. As can be seen in Table 10 below, both the strategies consistently point towards this result.

Table 10: Kiswahili Rasch Score: PSM-DID estimate

	Strategy 1	Strategy 2
PSM-DID estimate	0.574	0.75
P-value (bootstrapping)	0.00	0.00
P-value (no bootstrapping)	0.00	0.00

Figure 12 below presents the results on the balancing properties of strategy two. The model performs well, and confirms results obtained through strategy one.

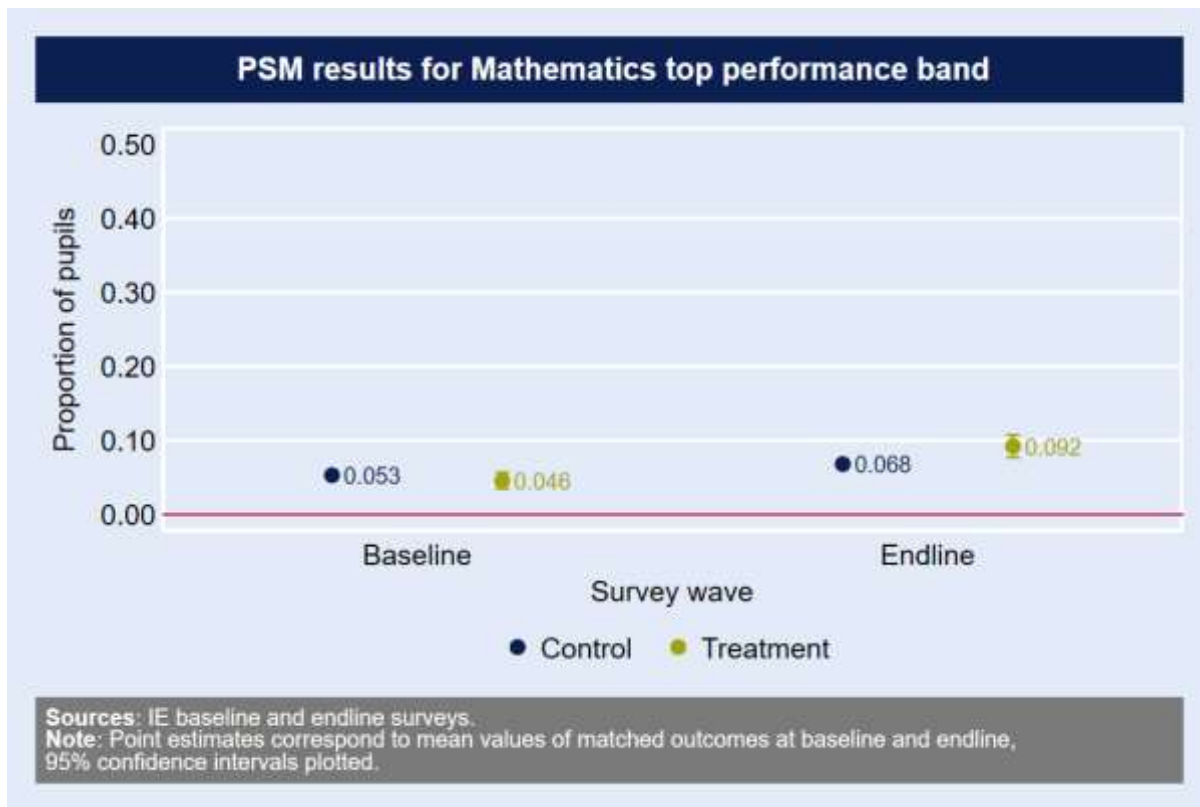
Figure 12: Kiswahili Rasch Scale- Second stage results (Strategy 2)



Proportion of pupils in the top performance band for Mathematics

Figure 13: Mathematics top band: Second stage results (Strategy 1)

Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
		Bandwidth	2
		Trimming	3
		N on common support	2794
		Rubin's B	[before matching] 64.51
		Rubin's R	[before matching] 1.32
ATT	-0.007	Rubin's B	[after matching] 9.27
SE (bootstrapping)	0.009	Rubin's R	[after matching] 1.25
SE (no bootstrapping)	0.009		
Endline			
		Bandwidth	4
		Trimming	3
		N on common support	2870
		Rubin's B	[before matching] 70.42
		Rubin's R	[before matching] 1.26
ATT	0.022	Rubin's B	[after matching] 8.4
SE (bootstrapping)	0.011	Rubin's R	[after matching] 1.05
SE (no bootstrapping)	0.011		
DID Estimate	0.03		
p-value (bootstrapping)	0.04		
p-value (no bootstrapping)	0.03		

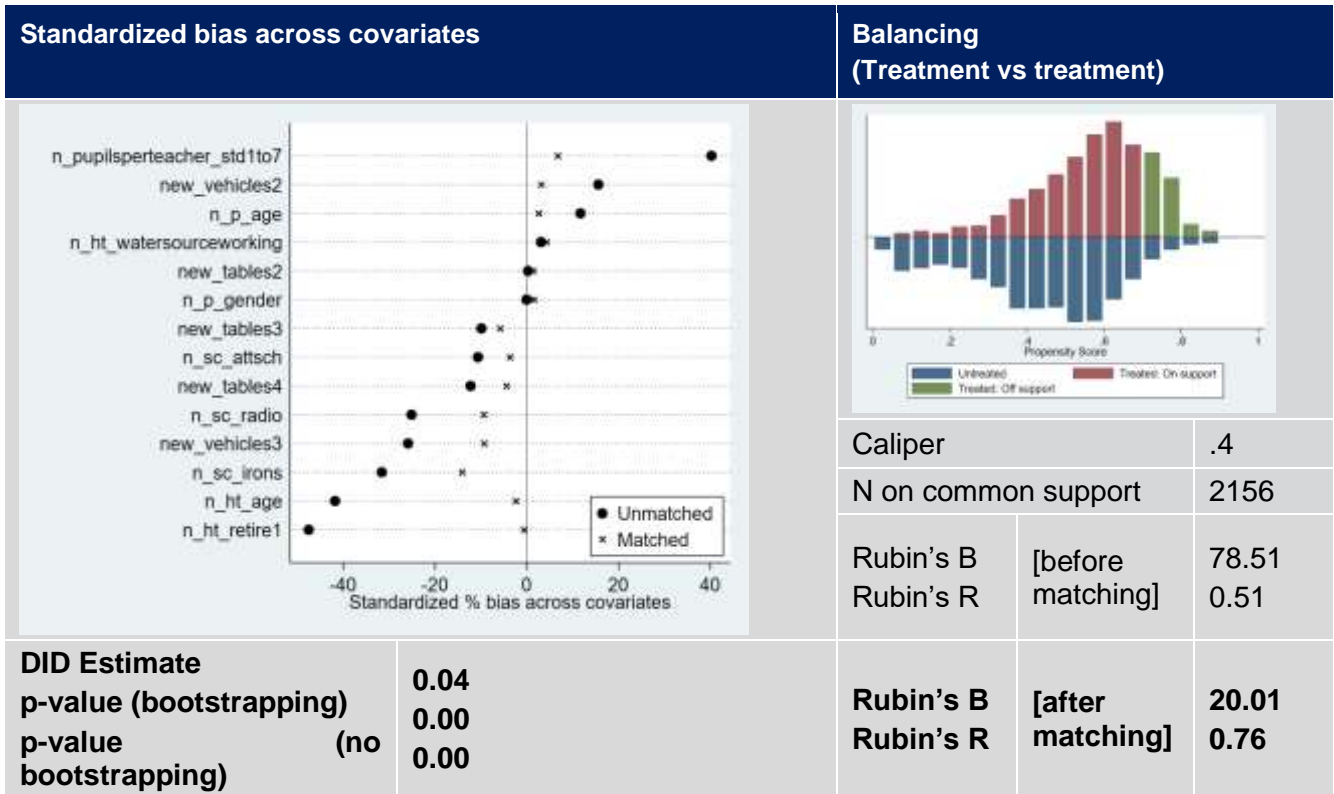
Figure 14: Mathematics top band: Matched outcome at baseline and endline

Both strategy one and strategy two show a positive and statistically significant change in the proportion of pupils in the top performance band for maths. The analysis, which achieves optimal balance, is thus able to provide a positive assessment of the impact of EQUIP-T on this indicator. As shown in Table 11 below, the proportion of pupils in the maths top performance band increases by three percentage points due to EQUIP-T. The PSM DID estimate from strategy two (robustness check) confirms the positive impact of EQUIP-T, showing a slightly larger and more significant increase in the proportion of pupils in the top band.

Table 11: Mathematics top band: PSM-DID estimate

	Strategy 1	Strategy 2
PSM-DID estimate	0.03	0.04
P-value (bootstrapping)	0.04	0.00
P-value (no bootstrapping)	0.03	0.00

Figure 15: Mathematics top band- Second stage results (Strategy 2)



Proportion of pupils in the bottom performance band for Mathematics

Figure 16: Mathematics bottom band: Second stage results (Strategy 1)

Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
		Bandwidth	2
		Trimming	8
		N on common support	2722
		Rubin's B [before matching]	65.36
		Rubin's R [before matching]	1.36
ATT	0.042	Rubin's B [after matching]	9.32
SE (bootstrapping)	0.014	Rubin's R [after matching]	0.97
SE (no bootstrapping)	0.013		
Endline			
		Bandwidth	4
		Trimming	5
		N on common support	2832
		Rubin's B [before matching]	69.26
		Rubin's R [before matching]	1.24
ATT	-0.036	Rubin's B [after matching]	10.54
SE (bootstrapping)	0.012	Rubin's R [after matching]	1.01
SE (no bootstrapping)	0.013		
DID Estimate	-0.079		
p-value (bootstrapping)	0.00		
p-value (no bootstrapping)	0.00		

Figure 17: Mathematics bottom band: Matched outcome at baseline and endline

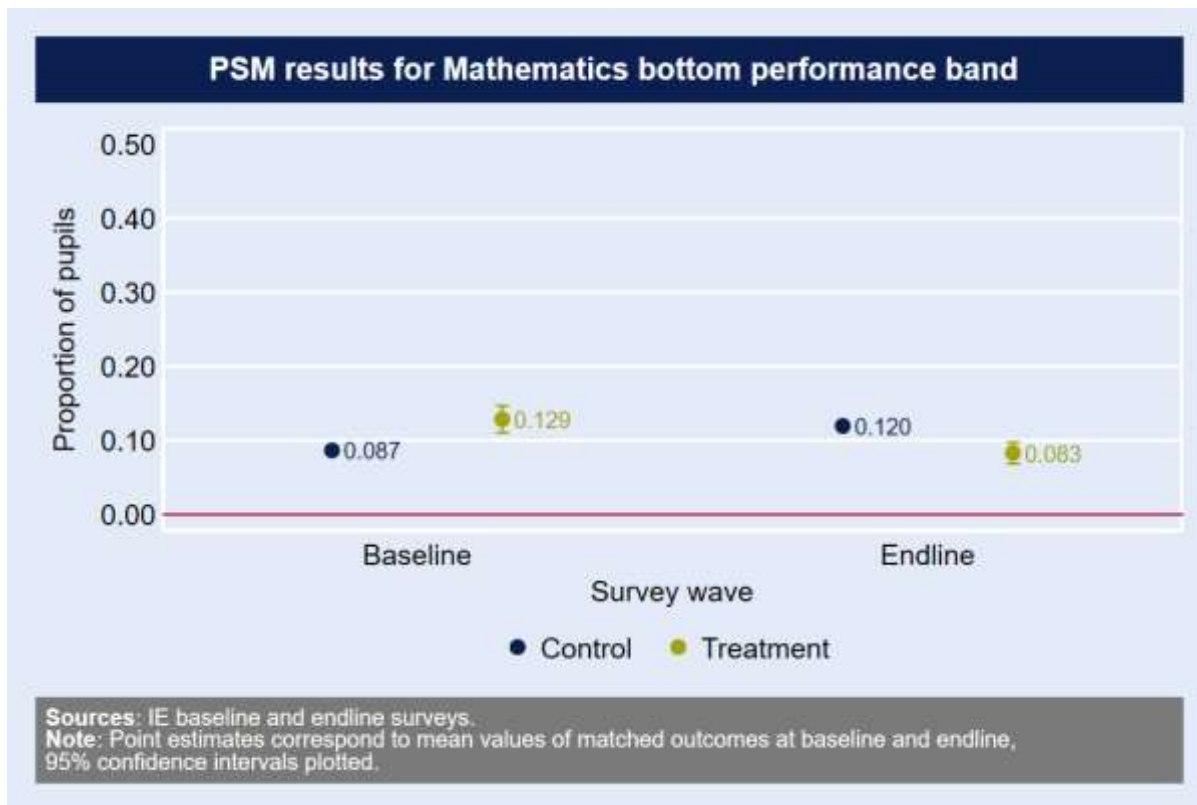


Figure 17 above shows that the PSM estimates point to an overall decrease in the proportion of pupils in the bottom performance band for Mathematics, but also indicates a difference in this trend across treatment and comparison schools. There seems to be an overall increase in the proportion of pupils in the bottom performance band in control schools between baseline and endline.

As can be seen in Table 12 below, this means that the study finds evidence of a statistically significant impact of EQUIP-T (over and above the potential effects of the other training initiatives) on the proportion of pupils in the bottom performance band for Mathematics. Both the strategies are consistent with each other with regards to this assessment.

Table 12: Mathematics bottom band: PSM-DID estimate

	Strategy 1	Strategy 2
PSM-DID estimate	-0.079	-0.10
P-value (bootstrapping)	0.00	0.00
P-value (no bootstrapping)	0.00	0.00

Figure 18 below shows results on the balancing properties of the robustness check strategy. As can be seen in the 'after matching' row, balancing performs very well also for treatment-to-treatment observations across time.

Mathematics Rasch Scale

Figure 19: Mathematics Rasch scale: Second stage results (Strategy 1)

Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
		Bandwidth	6
		Trimming	5
		N on common support	2755
		Rubin's B [before matching]	44.12
		Rubin's R [before matching]	1.16
ATT	-0.285	Rubin's B [after matching]	11.01
SE (bootstrapping)	0.071	Rubin's R [after matching]	1.12
SE (no bootstrapping)	0.07		
Endline			
		Bandwidth	6
		Trimming	3
		N on common support	2832
		Rubin's B [before matching]	69.58
		Rubin's R [before matching]	1.32
ATT	0.257	Rubin's B [after matching]	11.5
SE (bootstrapping)	0.084	Rubin's R [after matching]	1.22
SE (no bootstrapping)	0.079		
DID Estimate	0.542		
p-value (bootstrapping)	0.00		
p-value (no bootstrapping)	0.00		

Figure 20: Mathematics Rasch scale : Matched outcome at baseline and endline

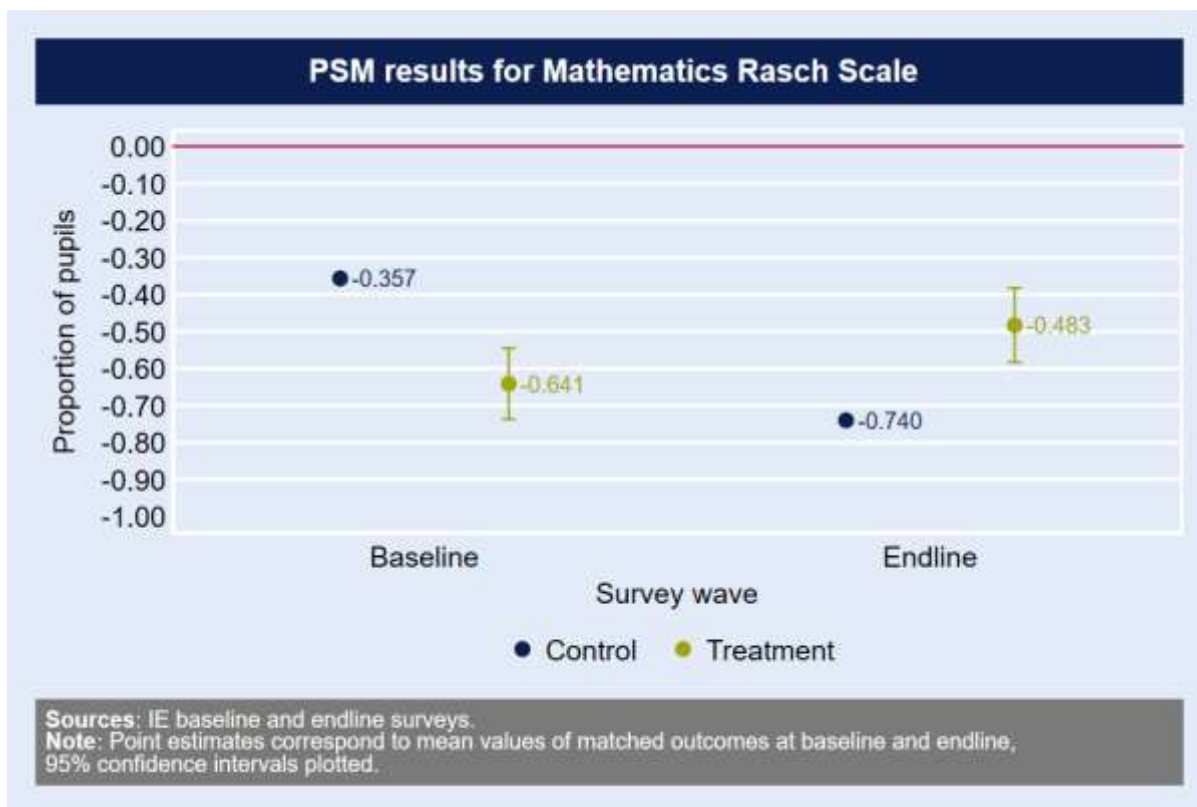


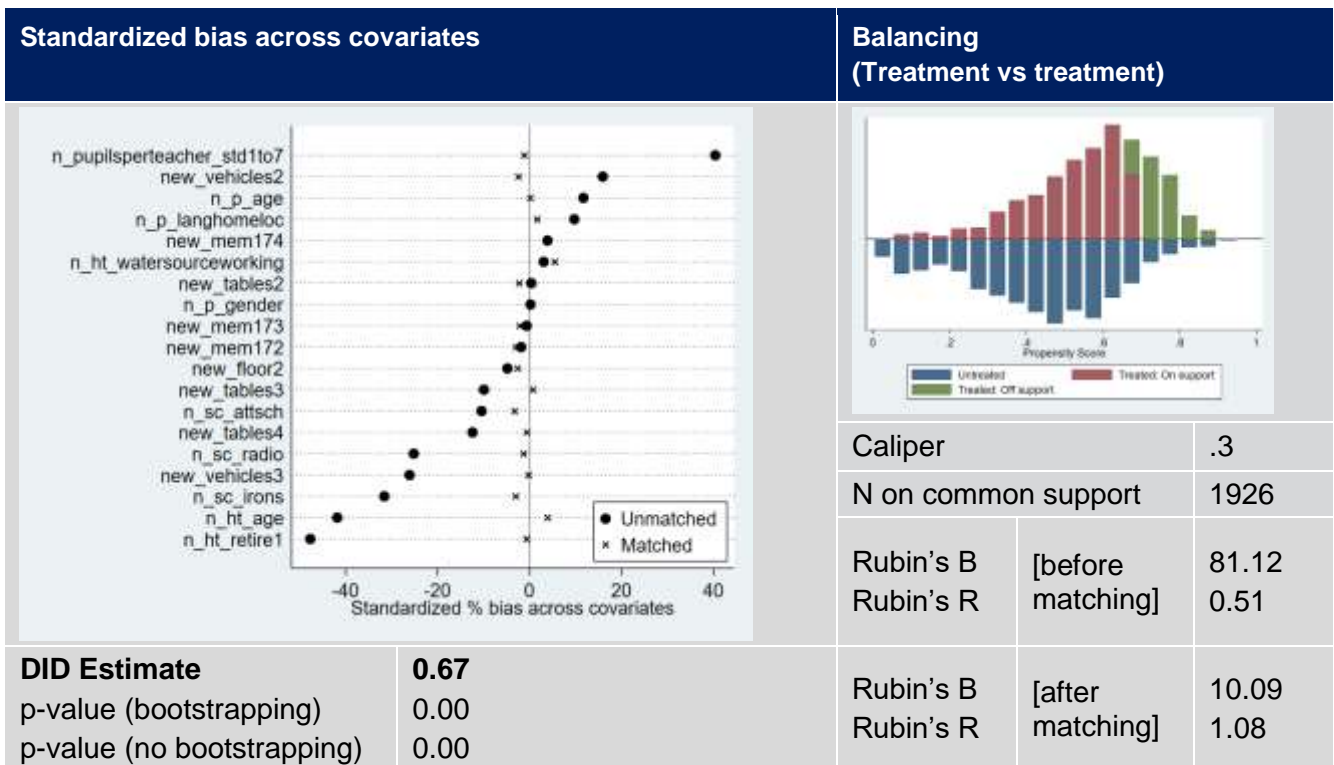
Figure 20 above confirms the Mathematics top and bottom band findings. There is a statistically significant improvement in the Rasch scores between baseline and endline that is attributable to the marginal impact of EQUIP-T. This points to an overall positive picture of the impact of the programme on maths skills. As can be seen in Table 13 below, the two strategies are consistent with each other with regards to this assessment in terms of direction and statistical significance, with strategy two showing a slightly larger impact.

Table 13: Mathematics Rasch Scale: PSM-DID estimate

	Strategy 1	Strategy 2
PSM-DID estimate	0.542	0.67
P-value (bootstrapping)	0.00	0.00
P-value (no bootstrapping)	0.00	0.00

Figure 21 below presents results on the balancing properties of the robustness check strategy. Also in this case, balancing is ideal for treatment observations across time.

Figure 21: Mathematics Rasch Scale- Second stage results (Strategy 2)

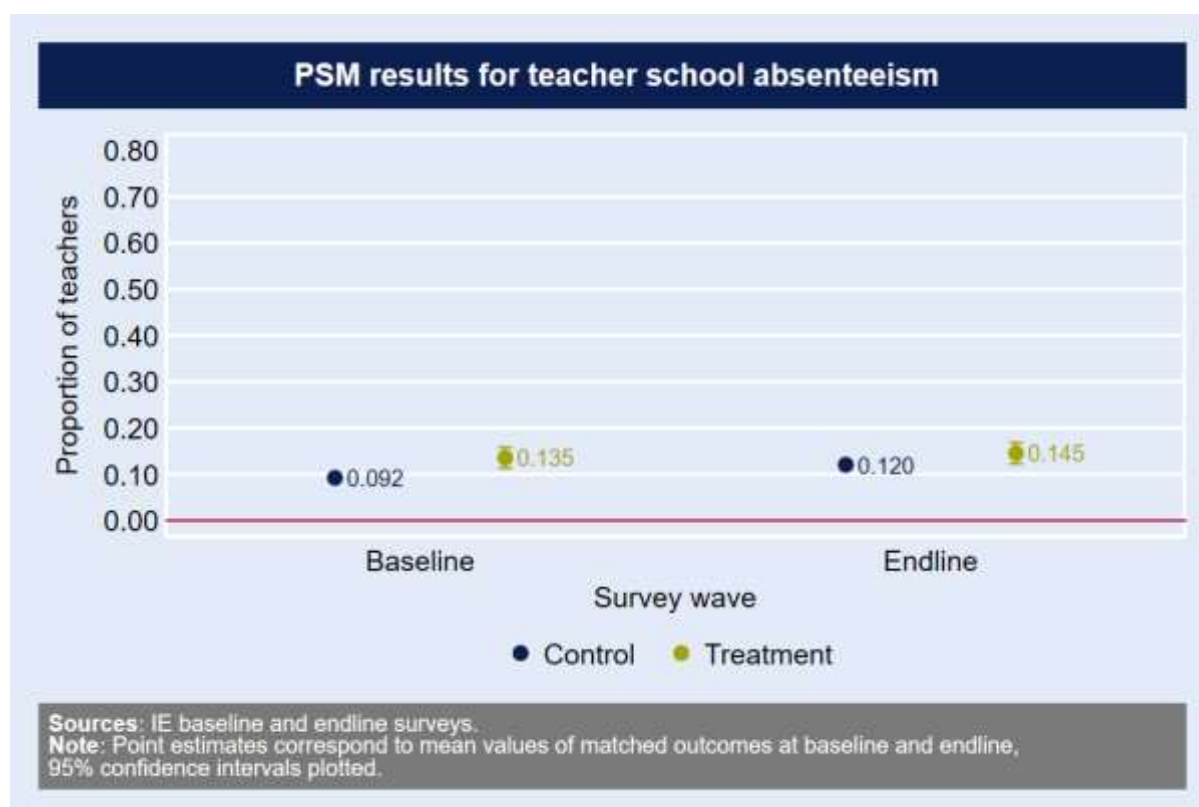


Teacher school absenteeism

Figure 22: Teacher school absenteeism: Second stage results (Strategy 1)

Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
		Bandwidth	4
		Trimming	5
		N on common support	1814
		Rubin's B [before matching]	77.91
		Rubin's R [before matching]	0.73
ATT	0.05	Rubin's B [after matching]	20.25
SE (bootstrapping)	0.016	Rubin's R [after matching]	1.04
SE (no bootstrapping)	0.016		
Endline			
		Bandwidth	2
		Trimming	3
		N on common support	1831
		Rubin's B [before matching]	68.53
		Rubin's R [before matching]	1.32
ATT	0.01	Rubin's B [after matching]	18.93
SE (bootstrapping)	0.018	Rubin's R [after matching]	1.2
SE (no bootstrapping)	0.017		
DID Estimate	-0.039		
p-value (bootstrapping)	0.11		
p-value (no bootstrapping)	0.10		

Figure 23: Teacher school absenteeism: Matched outcomes at baseline and endline

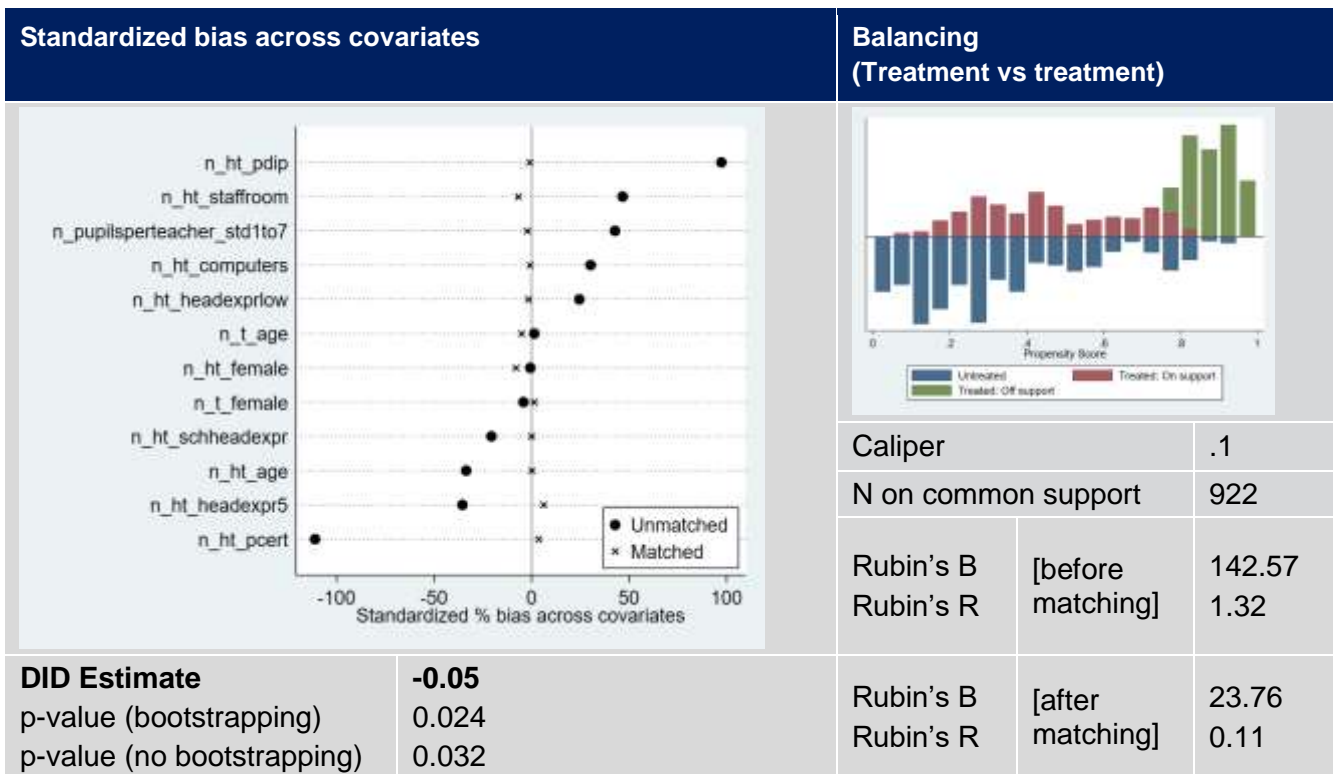


The analysis does not find strong evidence of a programme impact in terms of the proportion of teachers who are absent from school on the day of the survey. As shown in Table 14, both models display positive trends on this indicator, but the results are weakly significant in the main strategy. The impact estimate is more significant in the robustness check result. However, as shown in Figure 24 below, the balancing properties of strategy two are not ideal (that is, drastic reduction of analytical sample on common support and disproportionately low Rubin's R value) and its indications are not as reliable. Therefore, these results indicate that the evidence of programme impact on school absenteeism cannot be considered strong.

Table 14: Teacher school absenteeism: PSM-DID estimate

	Strategy 1	Strategy 2
PSM-DID estimate	-0.039	-0.05
P-value (bootstrapping)	0.11	0.024
P-value (no bootstrapping)	0.10	0.032

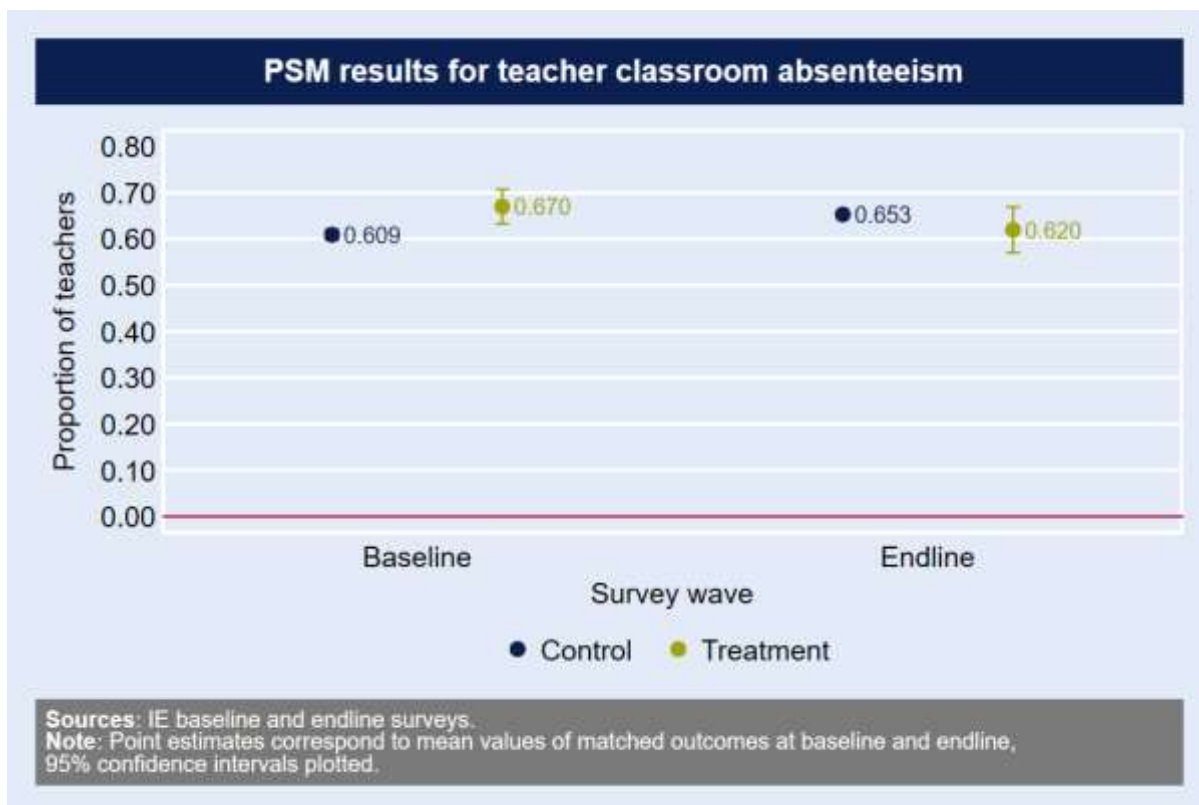
Figure 24: Teacher School Absenteeism- Second stage results (Strategy 2)



Teacher classroom absenteeism

Figure 25: Teacher classroom absenteeism: Second stage results (Strategy 1)

Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
ATT		Bandwidth	
SE (bootstrapping)		Trimming	
SE (no bootstrapping)		N on common support	
		Rubin's B	Rubin's R
		[before matching]	[before matching]
		[after matching]	[after matching]
0.061		6	1326
0.35		5	94.75
0.31		1.05	23.39
		1.69	
Endline			
ATT		Bandwidth	
SE (bootstrapping)		Trimming	
SE (no bootstrapping)		N on common support	
		Rubin's B	Rubin's R
		[before matching]	[before matching]
		[after matching]	[after matching]
-0.027		2	758
0.04		3	54.67
0.038		0.94	14.01
DID Estimate		1.15	
p-value (bootstrapping)			
p-value (no bootstrapping)			
0.10			
0.07			

Figure 26: Teacher classroom absenteeism: Matched outcome at baseline and endline

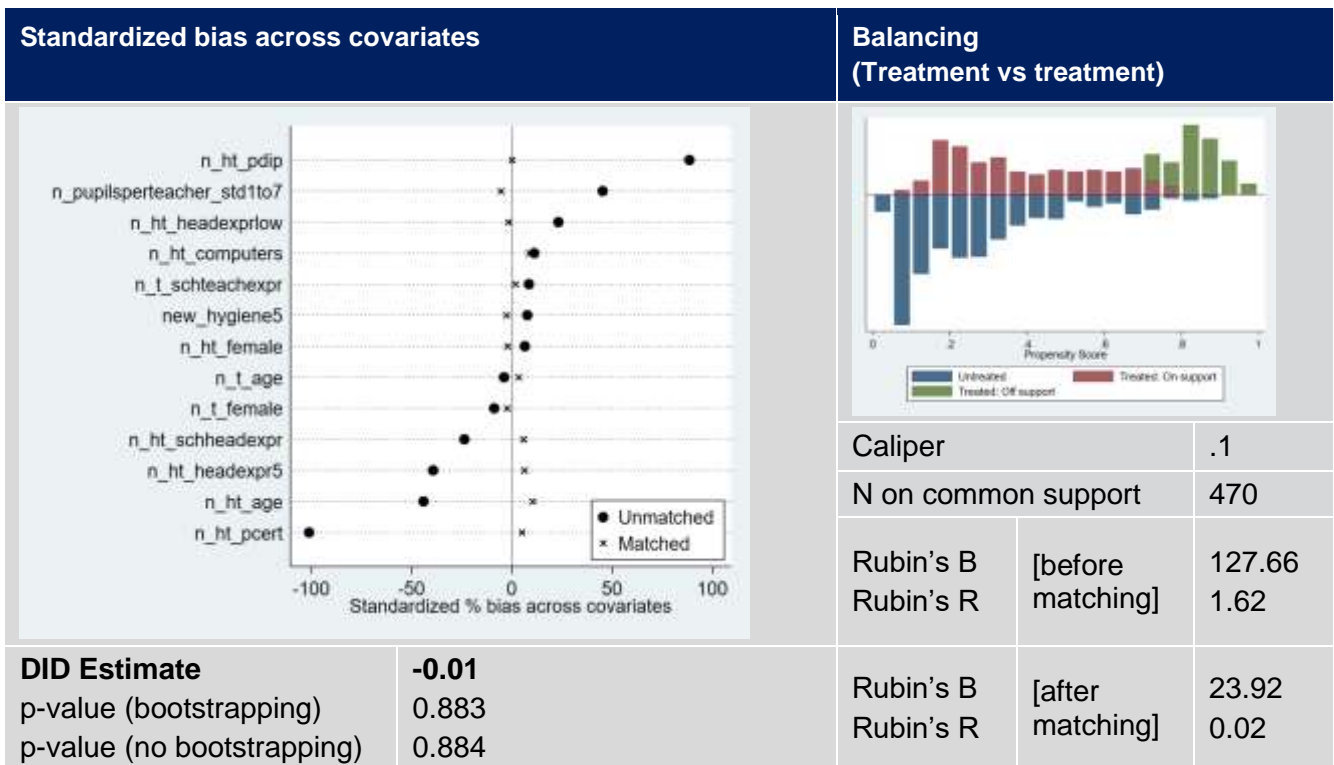
As shown in Figure 26 above, the PSM analyses indicate that at baseline the proportion of teachers that were absent from classes was marginally higher in treatment schools than in comparison schools, whereas the opposite was true at endline. Consequently, the PSM DID strategy one estimates indicates a reduction in the proportion of teachers absent from classes they are timetabled to teach before lunch as a result of EQUIP-T. As shown in Table 15 below, these results are, however, weak from a statistical significance point of view and are not at all confirmed by our second strategy robustness check, which shows a statistically insignificant result (though the magnitude of the estimate is in line with strategy one). Also in this case, the evidence of programme impact on this indicator cannot be considered strong.

Table 15: Teacher classroom absenteeism: PSM-DID estimate

	Strategy 1	Strategy 2
PSM-DID estimate	-0.088	-0.01
P-value (bootstrapping)	0.10	0.883
P-value (no bootstrapping)	0.07	0.884

Figure 27 below shows that the balancing properties of the strategy two matching are not ideal due to a reduced sample in common support and a low value for Rubin's R test.

Figure 27: Teacher Classroom Absenteeism- Second stage results (Strategy 2)



Proportion of teachers who report participation in performance appraisal

Figure 28: Teacher performance appraisal: Second stage results (Strategy 1)

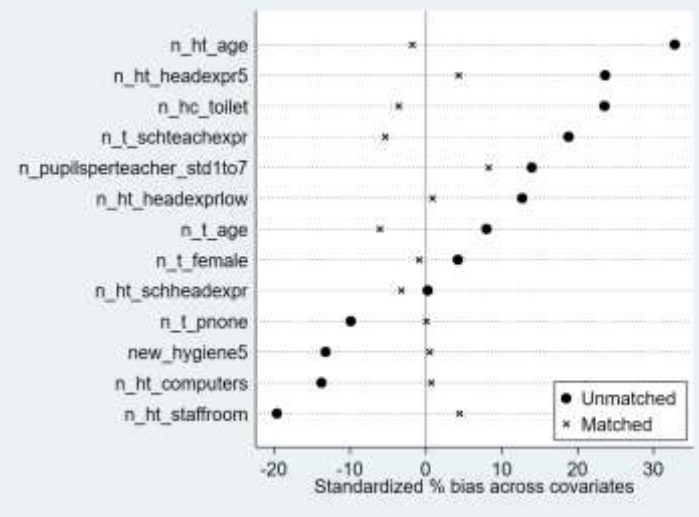
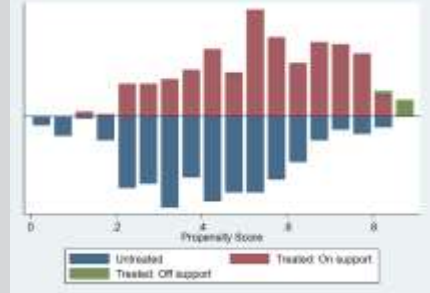
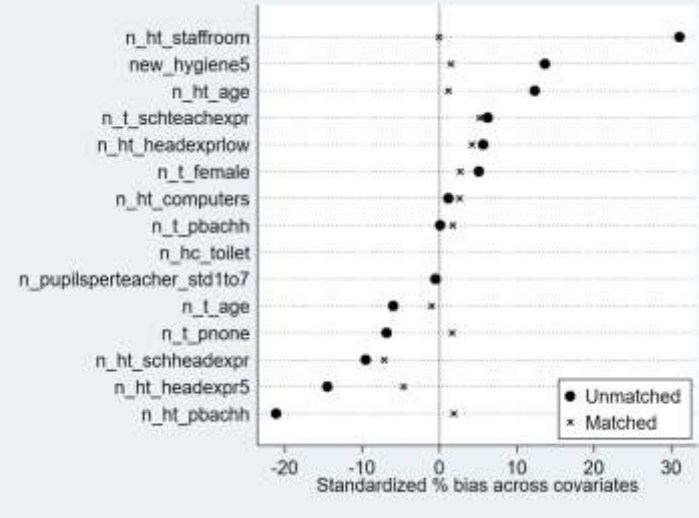
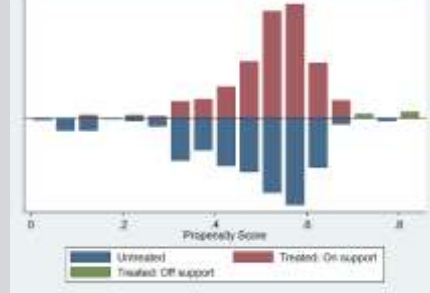
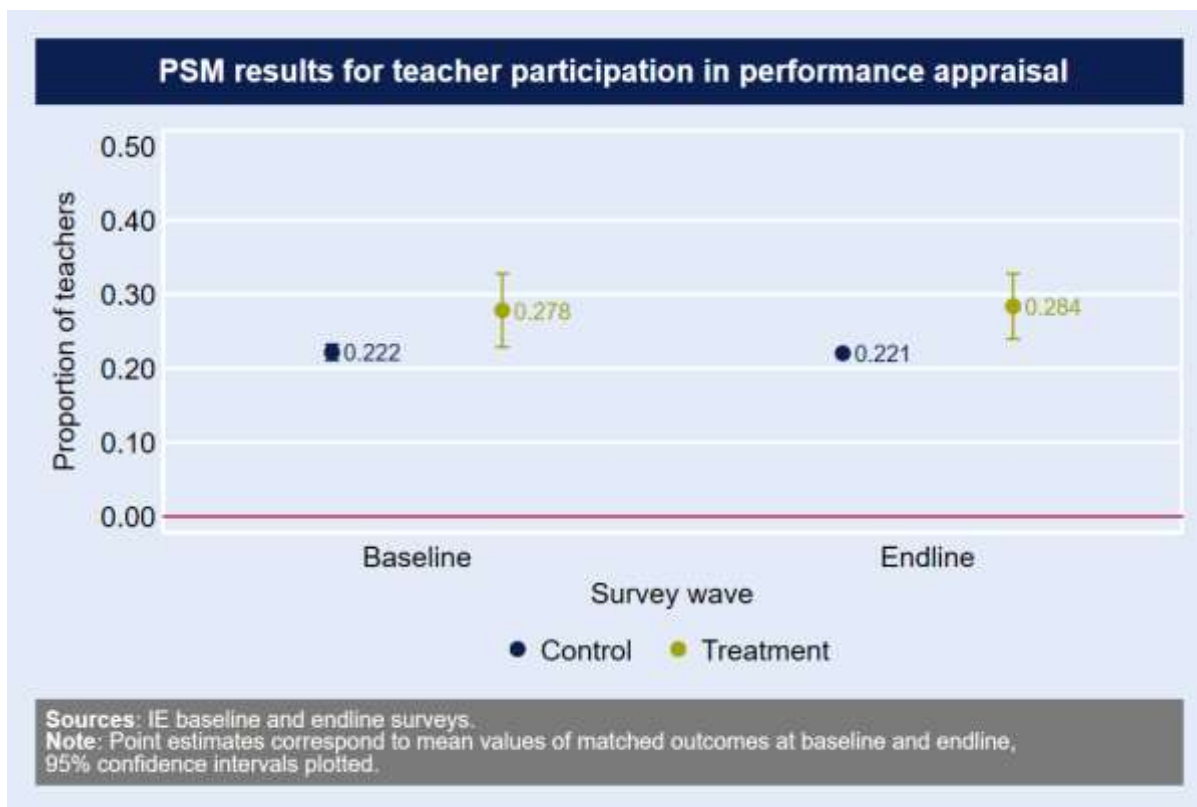
Standardized bias across covariates and ATT		Balancing (Control vs Treatment)	
Baseline			
 <p>A dot plot showing standardized bias across covariates for baseline. The x-axis ranges from -20 to 30. The y-axis lists covariates: n_ht_age, n_ht_headexpr5, n_hc_toilet, n_t_schteacherexpr, n_pupilsperteacher_std1to7, n_ht_headexprlow, n_t_age, n_t_female, n_ht_schheadexpr, n_t_pnone, new_hygiene5, n_ht_computers, n_ht_staffroom. Legend: ● Unmatched, * Matched.</p>		 <p>A histogram of propensity scores for baseline. The x-axis is Propensity Score (0 to 8). The y-axis shows density. Legend: ● Unmatched (blue), * Matched (red), * Off support (green).</p>	
ATT		Bandwidth	
SE (bootstrapping)		Trimming	
SE (no bootstrapping)		N on common support	
		Rubin's B	[before matching]
		Rubin's R	0.89
		Rubin's B	[after matching]
		Rubin's R	1.28
Endline			
 <p>A dot plot showing standardized bias across covariates for endline. The x-axis ranges from -20 to 30. The y-axis lists covariates: n_ht_staffroom, new_hygiene5, n_ht_age, n_t_schteacherexpr, n_ht_headexprlow, n_t_female, n_ht_computers, n_t_pbachh, n_hc_toilet, n_pupilsperteacher_std1to7, n_t_age, n_t_pnone, n_ht_schheadexpr, n_ht_headexpr5, n_ht_pbachh. Legend: ● Unmatched, * Matched.</p>		 <p>A histogram of propensity scores for endline. The x-axis is Propensity Score (0 to 8). The y-axis shows density. Legend: ● Unmatched (blue), * Matched (red), * Off support (green).</p>	
ATT		Bandwidth	
SE (bootstrapping)		Trimming	
SE (no bootstrapping)		N on common support	
		Rubin's B	[before matching]
		Rubin's R	0.4
		Rubin's B	[after matching]
		Rubin's R	1.26
DID Estimate			
p-value (bootstrapping)			
p-value (no bootstrapping)			

Figure 29: Teacher performance appraisal: Matched outcome at baseline and endline

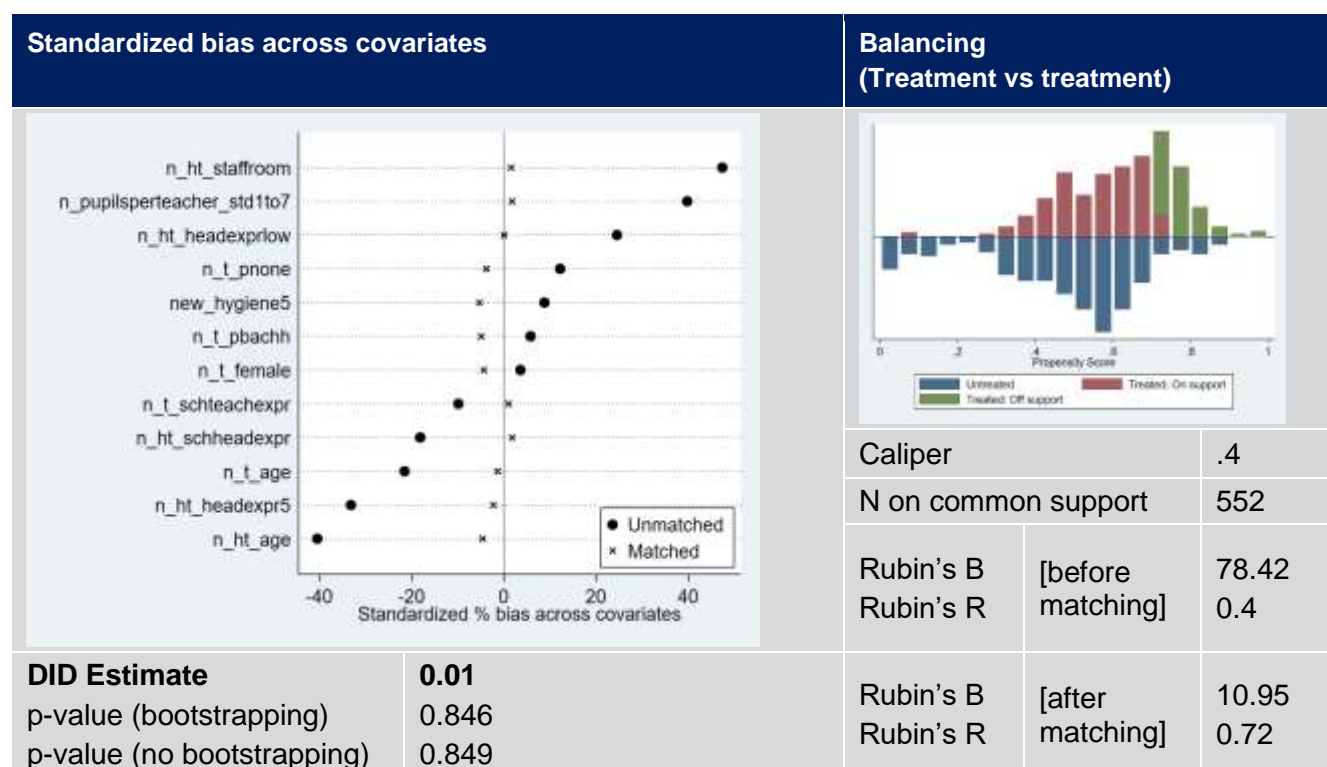


As shown by Figure 28 and Figure 30, the balance achieved by the matching protocols for both strategy one and strategy two is acceptable, even if the sample of teachers is small. Neither strategy one nor strategy two find any statistically significant estimate of impact on teacher participation in performance appraisal, as reported in Table 16 below. Hence, there is strong evidence that the EQUIP-T programme has had no impact on this indicator.

Table 16: Teacher performance appraisal: PSM-DID estimate

	Strategy 1	Strategy 2
PSM-DID estimate	0.01	0.01
P-value (bootstrapping)	0.85	0.846
P-value (no bootstrapping)	0.85	0.849

Figure 30: Teacher Performance Appraisal- Second stage results (Strategy 2)



4.6 Comment on effect size

A logical question which flows from the analysis of programme impact on early grade learning discussed above, is whether a 0.5 SD impact on Kiswahili scale scores, and a 0.3 SD impact on maths scale scores over four years, can be considered a low, moderate or large impact? It is very common in the education impact literature to cite impact sizes in SD units and to compare these across different impact evaluations. Hence, one approach is to apply one of the commonly used rules-of-thumb to put the EQUIP-T impact into context. For example, Hattie (2009) synthesised findings from a large range of analyses into what works to improve student achievement and found that the average effect size was 0.4 SD—he proposed this as a useful comparator point for judging the success of other interventions.¹² JPAL (2014) describes an effect size of 0.5 SD or more on student learning as ‘very large’.

Another approach is to be more specific, and to find comparators from impact evaluations of programmes that are targeting early grade literacy and numeracy in similar contexts. Gove et al (2017) summarise the impact of a number of recent early learning programmes in Africa. Cluster randomised control trials were used to assess two of the early primary grade learning programmes covered in the paper, and so these studies had counterfactuals for estimating impact. The impact estimate for Liberia’s Teacher Training Program II (operating in 6 out of 15 counties, targeted at grades 1 to 3 pupils, assessed after 2 years), on oral reading fluency¹³ was 0.3 SD. Uganda’s School Health and Reading Program (operating in 30 out of 111 districts, targeted at grades 1 to 4 pupils, assessed after 3 or 4 years) had an estimated impact on oral reading fluency of between 0.2 SD and 1.2 SD for

¹² Hattie’s more recent review from 2017 is based on an even larger number of studies but his conclusions remain similar: see <https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/>.

¹³ Number of correct words read per minute.

different districts. Put in this context, the EQUIP-T impact results for Kiswahili are not dissimilar. But there is a question of how valid such comparisons are.

Singh (2015a) highlighted a number of flaws with using SDs to compare effect sizes across impact evaluations in education. One key problem is that a SD is a measure of dispersion, and this is not the same in different samples. Singh cites the example of the PISA maths assessment where the SD of test scores for the same age group was 75% lower in one country compared with another. In this situation, an intervention which delivers the same absolute gain in learning in both countries will look less effective (i.e. have a lower effect size in SDs) in the country with the high spread of test scores. Singh also argues that differences in test design, scoring, and analysis methods can also greatly affect the level of impact expressed in SD units. One illustration cited by Singh of the difference that analysis methods can make to effect sizes in SD is based on English test scores for a sample of private school students in India—one method yielded an effect size of 0.28 SD while the other method, using the same data, gave a result of more than 0.6 SD (Singh, 2015b). In short, comparing effect sizes in SDs across different studies is not a very reliable approach.

The impact of EQUIP-T is perhaps best understood by reflecting on the local context, on the increased share of pupils that are achieving at higher band levels as a result of EQUIP-T, and the estimated additional number of children that are reaching the required curriculum standards.

5 Supplementary descriptive trends in programme areas

This chapter provides supplementary descriptive trend analysis of indicators of pupil learning (Section 5.1), teacher performance (Section 5.2), SLM (Section 5.3), and turnover of staff in education posts (Section 5.4) in programme treatment schools. Staff turnover is discussed in a cross-cutting section, as it covers teachers, INCOs, head teachers and WEOs, and thus gives some insight into the combined effects of high turnover on schools. These support the findings presented in Chapters 3, 4 and 5 in Volume I. For ease of reference, the SLM section in this chapter also includes the box which summarises the impact estimates on teachers' participation in performance appraisal, as this was not included in Volume I.

5.1 Pupil learning and background characteristics

There are three main types of supplementary descriptive analysis of pupil learning discussed in this section.

The first two subsections (5.1.1 and 5.1.2) show **trends in raw score indicators of pupil learning achievement** in Kiswahili (literacy) skills and maths (numeracy) skills between baseline, midline and endline. Similar types of raw score based indicators are reported in the monitoring reports of other large-scale education quality improvement programmes in Tanzania (see, for example, RTI, 2016). These include indicators such as 'number of words correctly read per minute' and 'percentage of addition questions answered correctly'.

The next three sub-sections (5.1.3, 5.1.4 and 5.1.5) focus on **learning trends for different subpopulations of Standard 3 pupils**. In Volume I Chapter 3, trends in learning disparities by gender, pupils' home language and poverty status are discussed, while in this section the focus is on changes in levels of learning outcomes for each beneficiary group separately. In addition, this section reports the prevalence of disabilities among Standard 3 pupils (based on pupil self-reporting), as well as changes in learning outcomes for pupils with and without physical disabilities between midline and endline.

The final sub-section (5.1.6) presents estimates of **trends in the absolute number of pupils achieving at the level of the different curriculum-linked performance bands**, in the 17 districts represented by the evaluation survey sample.

5.1.1 Pupil Kiswahili raw test score indicators

Pupil's Kiswahili skills: There have been large improvements in reading speeds, reading and listening comprehension and writing skills between baseline and endline (Table 17, Table 18 and

Table 19). On average, pupils read significantly faster at endline than at baseline, and this holds across the four different subtests of syllables, familiar words, invented words and a story passage. The size of the change in each case is fairly large—for example, pupils read a story passage at 21 words per minute at baseline and this increased by 9 words per minute to reach 30 words per minute by endline. However, the skills improvement happened between baseline and midline (see Annex G for results of significance tests between baseline and midline), while there was no significant change between midline and endline in reading speed on any of the subtests (Table 17).

Table 17: Pupils' oral reading speed at baseline, midline and endline in programme schools (trends in programme areas)

Skill areas	Indicator	Baseline (BL)		Midline (ML)		Endline (EL)		Difference	
		Estimate	N	Estimate	N	Estimate	N	BL-EL	ML-EL
Syllable sounds	Mean # of correct syllables read per minute	20.86	1491	30.39	1477	30.19	1488	9.33***	-0.19
Familiar words	Mean # of correct words read per minute	13.74	1496	19.87	1481	19.6	1495	5.85***	-0.27
Invented words	Mean # of words read per minute	9.33	1493	13.28	1477	12.9	1496	3.57***	-0.38
Story passage	Mean # of words read per minute	21.33	1496	29.97	1477	30.16	1495	8.82***	0.18

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil Kiswahili test).
Note: (1) Asterisks indicate statistical significance levels ***p<0.01, **p<0.05, *p<0.1

Pupils' reading and listening comprehension have also improved significantly on average between baseline and endline (Table 18), and the gains in average comprehension scores are large. For reading comprehension, the significant improvement happened between baseline and midline (Annex G), but for listening comprehension there was significant improvement between baseline and midline, and between midline and endline (Table 18).

Table 18: Pupils' reading and listening comprehension skills at baseline, midline and endline in programme schools (trends in programme areas)

Skill areas	Indicator	Baseline (BL)		Midline (ML)		Endline (EL)		Difference	
		Estimate	N	Estimate	N	Estimate	N	BL-EL	ML-EL
Reading comprehension	Mean test score (%)	19.04	1496	26.82	1477	29.19	1494	10.15***	2.37
	Percentage of pupils who scored more than 80%	1.18	1497	1.73	1477	2.7	1494	1.51**	0.97
	Percentage of pupils who scored 0%	55.88	1496	40.82	1477	35.39	1494	-20.49***	-5.44
Listening comprehension ²	Mean test score (%)	31.85	1496	40.94	1483	46.33	1499	14.48***	5.39***

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil Kiswahili test).
Note: (1) Asterisks indicate statistical significance levels ***p<0.01, **p<0.05, *p<0.1 (2) One of the five listening comprehension questions was changed between baseline and midline because the baseline question had not been translated correctly. Hence the baseline and midline test scores cannot be strictly compared.

Writing skills have also improved significantly between baseline and endline (**Error! Not a valid bookmark self-reference.**), with the average scores on spelling and punctuation rising by about 50% over this period. For spelling the improvement happened between baseline and midline (Annex G), but for punctuation there was significant improvement between baseline and midline, and between midline and endline (**Error! Not a valid bookmark self-reference.**).

Table 19 Pupils' writing skills at baseline, midline and endline in programme schools (trends in programme areas)

Skill area	Indicator	Baseline (BL)		Midline (ML)		Endline (EL)		Difference	
		Estimate	N	Estimate	N	Estimate	N	BL-EL	ML-EL
Spelling	Mean test score (%)	39.08	1496	55.47	1483	57.7	1499	18.61***	2.23
Punctuation	Mean test score (%)	30.03	1496	42.68	1483	46.65	1499	16.62***	3.97*

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil Kiswahili test).

Note: (1) Asterisks indicate statistical significance levels ***p<0.01, **p<0.05, *p<0.1

5.1.2 Pupil maths raw test score indicators

Pupils' maths skills: On the simplest mathematical task on the test, number comparison, pupils' skills have not changed significantly between baseline and endline, whereas skills in filling missing numbers in sequences, a more complex task, have improved significantly over the same period on average (Table 20).

Table 20: Pupils' skills in number comparison and missing numbers at baseline, midline and endline in programme schools (trends in programme areas)

Skill area	Indicator	Baseline (BL)		Midline (ML)		Endline (EL)		Difference	
		Estimate	N	Estimate	N	Estimate	N	BL-EL	ML-EL
Number comparison (discrimination)	Mean test score (%)	64.58	1495	64.39	1483	63.7	1499	-0.87	-0.69
Missing numbers in sequences	Mean test score (%)	28.47	1495	33.8	1483	35.71	1499	7.24***	1.91
	Percentage of pupils who scored more than 60%	7.29	1495	9.74	1483	14.11	1499	6.82***	4.38**
	Percentage of pupils who scored 0%	13.07	1495	7.7	1483	5.79	1499	-7.28***	-1.91

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil maths test).

Note: (1) Asterisks indicate statistical significance levels ***p<0.01, **p<0.05, *p<0.1

Pupils' skills in addition and subtraction have improved significantly on average between baseline and endline, with the exception of the harder (level 2) addition questions where average scores did not change significantly over this period (Table 21). For the other addition and subtraction subtests where skills improved on average, this happened between baseline and midline (see Annex G), and there was no significant change in average scores between midline and endline (Table 21). The share of pupils who scored zero on the harder (level 2) addition and subtraction questions has fallen significantly between baseline and endline, while there was no significant change in the share of pupils achieving over 80% on the same questions. This suggests that pupils with weaker skills have been supported to strengthen their skills over the period.

Table 21: Pupils' skills in addition and subtraction at baseline, midline and endline in programme schools (trends in programme areas)

Skill area	Indicator	Baseline (BL)		Midline (ML)		Endline (EL)		Difference	
		Estimate	N	Estimate	N	Estimate	N	BL-EL	ML-EL
Addition level 1	Mean score (%)	61.34	1495	68.53	1483	70.74	1499	9.4***	2.21
Additional level 2	Mean score (%)	30.03	1495	36.35	1483	33.43	1499	3.4	-2.92
Subtraction level 1	Mean score (%)	45.63	1495	53.87	1483	53.54	1499	7.91***	-0.33
Subtraction level 2	Mean score (%)	19.6	1495	24.42	1483	24.81	1499	5.22**	0.4
Addition and subtraction level 2	Mean score (%)	24.81	1495	30.38	1483	29.12	1499	4.31*	-1.26
	Percentage of pupils who scored more than 80%	7.87	1495	12.32	1483	11.27	1499	3.39	-1.05
	Percentage of pupils who scored 0%	37.85	1495	29.59	1483	29.68	1499	-8.17**	0.09

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil maths test).
Note: (1) Asterisks indicate statistical significance levels ***p<0.01, **p<0.05, *p<0.1

There have also been significant gains in the skills required to solve word problems between baseline and endline (Table 22). These problems require pupils to apply maths to real life scenarios. There has been persistent improvement in these skills on average, as shown by the significant gain in average score between baseline and midline (Annex G), and between midline and endline (Table 22).

Pupils' average score on multiplication problems fell significantly from 19% at baseline to 16% at endline, and there was a marked and significant fall between midline and endline (Table 22). This weakening of multiplication skills is likely to be related to changes in the Standards 1 and 2 curriculum which took place in 2015 (for more details, see Annex E).

Table 22: Pupils' skills in multiplication and word problems at baseline, midline and endline in programme schools (trends in programme areas)

Skill areas	Indicator	Baseline (BL)		Midline (ML)		Endline (EL)		Difference	
		Estimate	N	Estimate	N	Estimate	N	BL-EL	ML-EL
Word problems	Mean score (%)	28.8	1495	37.28	1483	41.8	1499	13.01***	4.53**
Multiplication	Mean score (%)	19.37	1495	24.4	1483	16.22	1499	-3.15*	-8.17***

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil maths test).
Note: (1) Asterisks indicate statistical significance levels ***p<0.01, **p<0.05, *p<0.1

5.1.3 Pupils' background characteristics, including disability

Standard 3 pupils' background characteristics, including gender, home language, poverty and disability status, are set out in Table 23. There is near even balance between boys and girls among the pupils and this has not changed significantly over time. Also largely unchanged over time, is the share of pupils whose main language at home is not Kiswahili—at endline 81% of pupils come from this linguistic background compared to 77% at baseline which is not a significant difference. There

has, however, been some change in pupils' economic circumstances since baseline. The share of pupils who live in households that fall below the poverty line increased by 6%, from 33% at baseline (a weakly significant change).

Turning to pupils' self-reported disability status which was captured at midline and endline only (see notes under Table 23 for measurement details). About 5-6% of pupils reported separately having difficulties with seeing and hearing at both midline and endline. The prevalence of pupils with movement difficulties was also 5% at midline, but this dropped significantly to 3% at endline. Taking these physical impairments together, 13% of pupils reported having either seeing or hearing or movement difficulties at midline and this fell to 10% at endline, although the change is only weakly significant. A much larger and significant drop of 11 percentage points occurred in the share of pupils saying that they have difficulties concentrating (from 18% to 7% between midline and endline). This is an unexpectedly large change and may reflect difficulties in measuring this type of disability (poor concentration related to a health problem), particularly using pupil self-reporting, rather than a real change.¹⁴ For this reason, learning outcome results are disaggregated by physical disability only in the subsections which follow.

Table 23 Pupils' background characteristics at baseline, midline and endline in programme schools (trends in programme areas)

	Baseline (BL)		Midline (ML)		Endline (EL)		Differences	
	Estimate	N	Estimate	N	Estimate	N	BL to EL	ML to EL
Pupil is female (% Std 3 pupils)	52.37	1497	50.53	1483	52.07	1499	-0.29	1.54
Main language spoken at home is local (% Std 3 pupils)	76.6	1497	76.23	1478	80.97	1498	4.37	4.75
Pupil from household below poverty line (% Std 3 pupils)	33.12	1443	36	1476	38.79	1495	5.67*	2.79
Pupil has difficulties in (% Std 3 pupils): ²								
seeing			5.9	1483	4.85	1499		-1.04
hearing			5.42	1483	4.73	1499		-.69
movement			4.99	1483	2.56	1499		-2.42**
concentration			17.87	1483	7.33	1499		-10.54***
seeing or hearing or movement (physical)			13.41	1483	10.32	1499		-3.09*
seeing or hearing or movement or concentration			26.87	1483	15.54	1499		-11.34***

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil background questions).

Note: (1) Asterisks indicate statistical significance levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. (2) Pupils self-reported their disability status by answering four of the Washington Group's short set of questions on disability: 'Do you have difficulties seeing, even if wearing glasses?'; 'Do you have difficulties hearing, even if using a hearing aid?'; 'Do you have difficulties walking or climbing steps?'; 'Do you have difficulties remembering or concentrating?'. These questions were taken from DFID's guide to disaggregating programme data by disability (undated) that was shared with the evaluation team in early 2016, just prior to the midline survey.

¹⁴ Pupils are required to identify if they have serious concentration difficulties that are related to a health problem rather than typical concentration problems most children experience in their learning. It may be very difficult for young children to make this distinction.

5.1.4 Pupil Kiswahili scale scores for different subpopulations¹⁵

Standard 3 pupils' Kiswahili skills improved significantly on average between baseline and endline for each of the subpopulations of pupils shown in Table 24. Average Kiswahili scale scores increased by 1.0 logit (0.8 SD) for boys over the four years, while the gain was 1.3 logits (1.1 SD) for girls. For both genders the largest improvement happened between baseline and midline, while in the final two years, average scale scores only improved significantly for girls not for boys.

For pupils whose mother tongue is not Kiswahili, average Kiswahili scale scores went up significantly by 1.2 logits (1.0 SD) between baseline and endline, outstripping the gain of 0.9 logits (1.0 SD) for the group of pupils who come from Kiswahili speaking homes. Indeed, the pupils whose home language is not Kiswahili continued to improve their Kiswahili skills on average between midline and endline, while for the Kiswahili home language group there was no significant change over the same period.

The improvement between baseline and endline in average Kiswahili scale scores for pupils from both poorer and from richer homes is very similar: 1.1 logit (0.9 SD) for pupils from poorer homes, and 1.1 logit (1.0 SD) for pupils from richer homes.

For pupils with physical disabilities, there was no significant change in average Kiswahili scale scores between midline and endline, while for pupils without seeing, hearing or movement impairments there was a modest, but only weakly significant, improvement in scale scores of 0.2 logits (0.2 SD).

Table 24 Trends in average Kiswahili scale scores by gender, home language, poverty status and disability in programme schools (trends in programme areas)

	Baseline (BL)		Midline (ML)		Endline (EL)		Difference	
	Estimate (logits)	N	Estimate (logits)	N	Estimate (logits)	N	BL-EL	ML-EL
All Standard 3 pupils	-1.6	1,487	-0.7	1,463	-0.5	1,487	1.1***	0.2*
Boys	-1.6	717	-0.8	723	-0.7	709	1.0***	0.2
Girls	-1.6	770	-0.6	740	-0.3	778	1.3***	0.3**
Kiswahili	-1.0	329	-0.3	317	-0.1	278	0.9***	0.1
Local language	-1.8	1,158	-0.8	1,141	-0.6	1,208	1.2***	0.3**
Poorer	-1.8	477	-0.9	531	-0.7	579	1.1***	0.2
Richer	-1.4	957	-0.6	925	-0.4	904	1.1***	0.2*
No physical disability			-0.7	1,268	-0.5	1,335		0.2*
Physical disability			-0.6	195	-0.5	152		0.1

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil maths test).
 Note: (1) Asterisks indicate statistical significance levels ***p<0.01, **p<0.05, *p<0.1

5.1.5 Pupil maths scale scores for different subpopulations¹⁶

Standard 3 pupils' maths skills improved significantly on average between baseline and endline for almost all of the subpopulations of pupils shown in Table 25. This strengthening of maths skills

¹⁵ Changes in average scale scores are expressed in logits and in SD units. The basis of the SD units are the endline distributions of pupil Kiswahili scale scores for each subpopulation. For example, the change in average Kiswahili scale score for girls in logits is expressed in SD units based on the SD of the endline distribution of girls' Kiswahili scale scores.

¹⁶ Changes in average scale scores are expressed in logits and in SD units. The basis of the SD units are the endline distributions of pupil maths scale scores for each subpopulation. For example, the change in average maths scale score for girls in logits is expressed in SD units based on the SD of the endline distribution of girls' maths scale scores.

occurred between baseline and midline, while in the final two years there was no significant change in skills level for any of the subpopulations.

Average maths scale scores increased by 0.4 logits (0.2 SD) for boys over the four years, while the gain was 0.5 logits (0.3 SD) for girls. Improvements in maths skills of a similar magnitude (0.4 to 0.6 logits) took place for pupils from poorer homes (0.2 SD) and from richer backgrounds (0.3 SD), and for pupils whose home language is not Kiswahili (0.3 SD) over the four years. Average maths scale scores improved by 0.4 logits (0.2 SD) for pupils who speak Kiswahili at home over the same period, but this change is not significant.

In line with midline to endline trends for the other subpopulations, pupils with (and without) physical disabilities showed no significant change in average maths scale scores over this two year period.

Table 25 Trends in average maths scale scores by gender, home language, poverty status and disability in programme schools (trends in programme areas)

	Baseline (BL)		Midline (ML)		Endline (EL)		Difference	
	Estimate (logits)	N	Estimate (logits)	N	Estimate (logits)	N	BL-EL	ML-EL
All Standard 3 pupils	-1.0	1,495	-0.6	1,483	-0.6	1,499	0.5***	0.1
Boys	-0.8	721	-0.5	734	-0.4	716	0.4***	0.1
Girls	-1.2	774	-0.8	749	-0.7	783	0.5***	0.1
Kiswahili	-0.4	330	0.2	320	0.0	280	0.4	-0.2
Local language	-1.2	1,165	-0.9	1,158	-0.7	1,218	0.6***	0.2
Poorer	-1.2	480	-0.9	536	-0.8	581	0.4**	0.1
Richer	-0.9	961	-0.5	940	-0.4	914	0.5***	0.1
No physical disability			-0.6	1,283	-0.5	1,345		0.1
Physical disability			-0.5	200	-0.6	154		-0.2

Source: Evaluation baseline, midline and endline surveys (Standard 3 pupil maths test).
 Note: (1) Asterisks indicate statistical significance levels ***p<0.01, **p<0.05, *p<0.1

5.1.6 Trends in absolute numbers of pupils achieving at different performance bands in 17 programme districts

Between the baseline and the endline, there was a positive shift in the distribution of Standard 3 pupils in treatment schools achieving at different performance bands in both Kiswahili and maths—the share of pupils achieving in the top two bands increased, while the share of pupils falling into the bottom two bands decreased (Table 26 and Table 27). Over the same period, the dramatic growth in Standard 3 enrolment (nationally and in the treatment areas) means that the change in the absolute number of pupils achieving at higher band levels in treatment areas is even starker.

Table 26 and Table 27 present estimates of the absolute number of Standard 3 pupils in each of the Kiswahili and maths performance bands alongside the share of Standard 3 pupils in each performance band. The former was calculated by multiplying the proportion of Standard 3 pupils in each performance band (estimated from the evaluation's baseline, midline and endline samples of Standard

3 pupils) with the total number of Standard 3 pupils enrolled in government primary schools in the 17 programme districts that are covered by the impact evaluation (obtained from secondary sources).¹⁷

Table 26: Distribution and estimates (in absolute terms) of Standard 3 pupils by Kiswahili performance band in treatment areas, baseline, midline, and endline (trends in programme areas)

	Baseline		Midline		Endline	
	Distribution (%)	Estimates	Distribution (%)	Estimates	Distribution (%)	Estimates
Kiswahili performance bands						
Band 0 below emerging std 1	39.37	53,851	23.16	31,610	16.41	33,341
Band 1E emerging std 1	8.02	10,970	6.39	8,721	7.26	14,750
Band 1A achieving std 1	16.78	22,952	19.57	26,710	26.87	54,593
Band 2E emerging std 2	23.76	32,499	28.46	38,843	31.94	64,893
Band 2A achieving std 2 or above	12.07	16,510	22.43	30,613	17.52	35,596
Total Standard 3 enrolment		136,782		136,484		203,173

Sources: Evaluation baseline, midline, and endline surveys (pupil test) for the distributions; EMIS data uploaded on to the Government of Tanzania's open data website (<http://opendata.go.tz/group/education-group>) for enrolment of Standard 3 pupils in 2016 and 2018; and EMIS data for 2014 (access database shared by the EQUIP-T MA in January 2015).

Notes: (1) The distribution of Standard 3 pupils by performance band was estimated from the evaluation's baseline, midline, and endline samples. (2) The total Standard 3 enrolment is equal to the number of Standard 3 pupils enrolled in government primary schools in the 17 programme districts that are covered by the impact evaluation. This is the pool of which the evaluation's sample of Standard 3 pupils is representative of.

Table 27: Distribution and estimates (in absolute terms) of Standard 3 pupils by maths performance band in treatment areas, baseline, midline, and endline (trends in programme areas)

	Baseline		Midline		Endline	
	Distribution (%)	Estimates	Distribution (%)	Estimates	Distribution (%)	Estimates
Maths performance bands						
Band 0 below emerging std 1	13.22	18,083	11.27	15,382	8.52	17,310
Band 1E emerging std 1	27.82	38,053	19.51	26,628	22.82	46,364
Band 1A achieving std 1	30.66	41,937	31.86	43,484	31.3	63,593
Band 2E emerging std 2	23.88	32,664	30.36	41,437	28.14	57,173
Band 2A achieving std 2 or above	4.42	6,046	7.01	9,568	9.23	18,753
Total Standard 3 enrolment		136,782		136,484		203,173

Sources: Evaluation baseline, midline, and endline surveys (pupil test) for the distributions; EMIS data uploaded on to the Government of Tanzania's open data website (<http://opendata.go.tz/group/education-group>) for enrolment of Standard 3 pupils in 2016 and 2018; and EMIS data for 2014 (access database shared by the EQUIP-T MA in January 2015).

¹⁷ Annex A in this Volume lists the 17 programme districts (located in the five regions where EQUIP-T implementation started in 2014) that are covered by the impact evaluation.

Notes: (1) The distribution of Standard 3 pupils by performance band was estimated from the evaluation's baseline, midline, and endline samples. (2) The total Standard 3 enrolment is equal to the number of Standard 3 pupils enrolled in government primary schools in the 17 programme districts that are covered by the impact evaluation. This is the pool of which the evaluation's sample of Standard 3 pupils is representative of.

Despite the shift in the distribution of Standard 3 pupils by Kiswahili performance band over time, the number of Standard 3 pupils in each performance band at endline is higher than the number of Standard 3 pupils in each respective performance band at both midline and baseline - with the exception of the number of pupils in Band 0. Furthermore, in the cases where the proportion of pupils in a band has increased over time, the increase in absolute terms is greater than the increase in proportions (for example, the proportion of pupils in Band 1A for Kiswahili increased from 17% at baseline to 27% at endline – an increase of 60% - while the number of pupils in that band increased from 22,952 at baseline to 54,593 at endline – an increase of 138%). This reflects the spike in enrolment of Standard 3 pupils at endline: 203,173 pupils at endline, an increase of 49% since both baseline and midline.¹⁸

However, the significant and dramatic decrease in the proportion of Standard 3 pupils in the bottom performance band for Kiswahili between baseline (39%) and endline (16%) has resulted in a substantial decrease in the number of pupils in that band (52,851 at baseline to 33,341 at endline), despite the increase in enrolment.

The results are similar for maths learning outcomes. With the exception of the bottom performance band, the number of pupils in each of the other four performance bands at endline is higher than the number of pupils in each respective performance band at both midline and baseline despite the changes in the distribution over time.

On the other hand, the significant decrease in the proportion of Standard 3 pupils in the bottom maths performance band between baseline (13%) and endline (9%) has resulted in a slight decrease in the number of pupils in that band (18,083 at baseline to 17,310 at endline), despite the increase in enrolment.

It is important to note some potential limitations to the method used to estimate absolute trends in the number of pupils achieving at different performance levels discussed above. The impact evaluation sample of Standard 3 pupils is representative of Standard 3 pupils in all government primary schools in 17 programme districts, which were drawn from the five regions where EQUIP-T started implementation in 2014. As a result, the total enrolment of Standard 3 pupils in government primary schools in the 17 programme districts was used to estimate the number of Standard 3 pupils in each performance band. However, it is worth noting two changes to the list of government primary schools in these 17 districts since baseline (that is, since the construction of the sampling frame at baseline from which the 100 sample treatment schools were drawn): (i) government primary schools in these 17 districts have opened and closed since baseline and the characteristics of new schools may be different, and (ii) some district boundaries have changed which means that some schools that were part of the sampling frame at baseline are no longer located in these 17 programme districts while new schools that were not part of the sampling frame at baseline are now located in the 17 districts.

¹⁸ Enrolment figures for Standard 3 at baseline and midline are almost identical and that is likely due to the fact that the rise in enrolment since baseline was largely the result of the fees-free education policy which was introduced in 2016 and led to a rise in enrolment for Standard 1 which has translated into an increased enrolment for Standard 3 by endline.

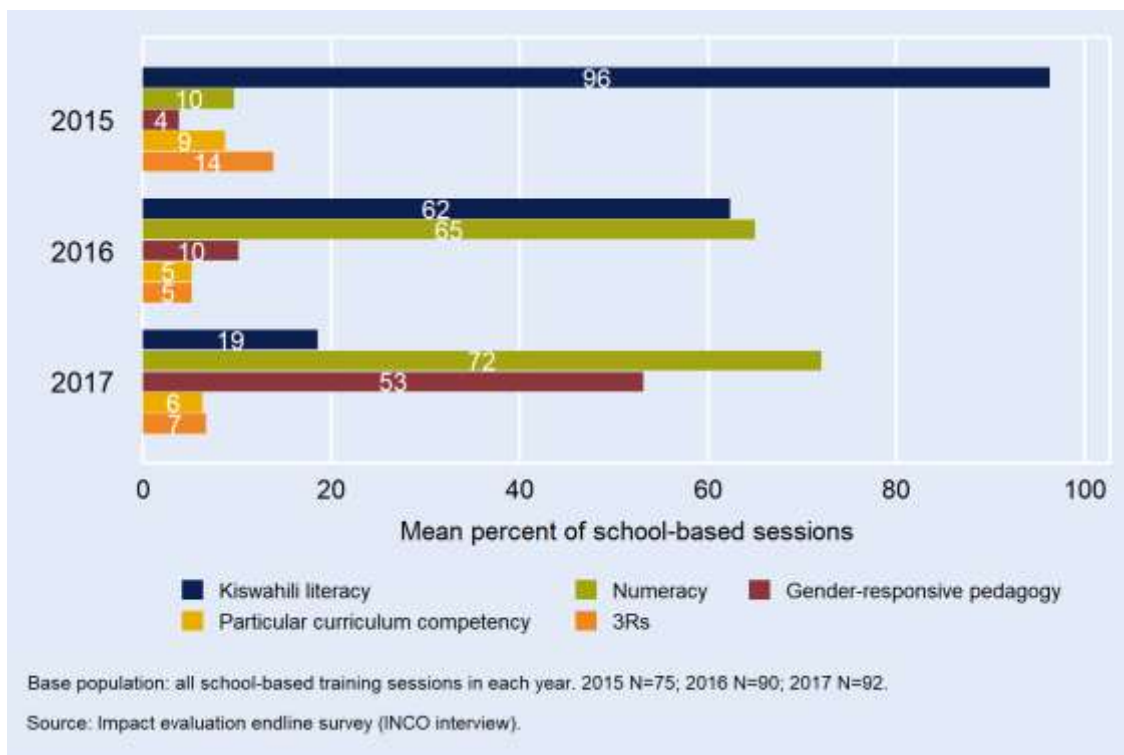
5.2 Teacher performance

5.2.1 EQUIP-T school-based in-service teacher training

While all schools are holding school-based in-service training sessions, the majority are not implementing it as intensively as required. Schools are expected to hold training sessions for three hours twice a month, while school is in session, which accounting for term breaks and holidays, would translate into about 18 sessions per year. In 2017, 45% of schools held 15 or more school-based training sessions, while 7% held from ten to 14 sessions, 29% held from five to nine sessions, and 19% held zero to four sessions. This is an improvement from 2015 and 2016 where 30% and 33% of schools respectively held 15 or more school-based training sessions, and 38% and 25% of schools respectively held zero to four sessions (Figure 9 in Volume I Chapter 4).

Since baseline, schools have held, on average, 44 school-based training sessions in total. This has gradually increased over time from an average of less than one session in 2014 to 12 sessions in 2015, 15 sessions in 2016 and 16 sessions in 2017. In the first quarter of 2018, the average number of sessions held was less than one, which may suggest that the school-based training component had started to wane by 2018.

Figure 31: Topics covered in school-based training sessions from 2015 to 2017 (trends in programme areas)



The topics covered by school-based training over time have been in line with the sequence of residential training rolled-out since baseline (Figure 31). Almost all sessions in 2015 (96%) covered Kiswahili literacy modules, while some covered 3Rs (14%), numeracy modules (10%)¹⁹, and particular

¹⁹ The most likely reason some sessions in 2015 were reported to have covered early grade numeracy despite this training only being rolled-out by EQUIP-T in 2016, is that some teachers confuse between the training on 3Rs and the training on Kiswahili and numeracy modules, given that both cover Kiswahili and numeracy topics under the new curriculum. As a result, some of the trainings reported on numeracy or Kiswahili modules might have in reality been on 3Rs while some of the trainings reported on 3Rs might have been on Kiswahili or numeracy modules.

curriculum competency (9%).²⁰ In 2016, the focus of the sessions was on numeracy and Kiswahili literacy modules. The majority of sessions (65%) covered numeracy modules, followed closely by Kiswahili modules (62%), and then gender-responsive pedagogy (10%) and particular curriculum competency (5%). In 2017, the focus was on numeracy (72% of sessions) and gender-responsive pedagogy (53%) and much less on Kiswahili literacy (19%). Only 6% of sessions in 2017 were on particular curriculum competencies.

School-based sessions last on average, two hours, which is less than the stipulated three. They are mostly held on school days after teaching hours (86%), while some 11% were held on school days during teaching hours, and 2% were held outside school days.

Less than a third (27%) of all school-based training sessions had any written record of that session taking place, and only 18% had session minutes. This has two implications: firstly, it means that all other training sessions that were reported by schools were based on participants' memory which introduces recall bias, particularly for sessions that happened some three or two years ago, and consequently poses some limitations to the data presented in this section. Secondly, it might reflect a lack of seriousness of schools' attitudes towards these sessions.

Facilitators of school-based training sessions

School-based training sessions are expected to be facilitated by the INCO as well as potentially some other teachers who had attended the training away from school. The majority of school-based training sessions have more than one teacher facilitating the sessions (Table 28). Almost half of all sessions (45%) have two facilitators, while 34% have three or more facilitators and 21% have only one facilitator. While the majority of sessions had the INCO as one of the facilitators, a sizeable minority (21%) were not facilitated by the INCO.

On average, facilitators are 38 years old and have 14 years of teaching work experience. Over three-quarters (76%) are Standards 1 to 3 teachers, 41% are the INCO and 9% are the head teacher. While the majority of the facilitators had attended the residential training, a sizeable minority had not. Of all facilitators of school-based training sessions that covered gender-responsive pedagogy, 21% had not attended the gender-responsive pedagogy training away from school, while 15% of facilitators of numeracy school-based sessions had not attended any numeracy training away from school and 12% of Kiswahili school-based session facilitators had not attended any Kiswahili training away from school.

Table 28: Profile of facilitators of EQUIP-T school-based training sessions

	Endline	
	Estimate	N
INCO was one of the facilitators of school-based training session (mean % school-based training sessions)	78.93	98
Number of facilitators in school-based training sessions (mean % school-based training sessions)		
One facilitator	21.39	98
Two facilitators	44.76	98
Three or more facilitators	33.85	98
Age (mean years)	38.02	94
Time working as a teacher (mean years)	14.35	94
Facilitators of school-based training sessions are (mean % facilitators)		
Standards 1-3 teachers	76.17	94
In-service training coordinator	40.58	94

²⁰ Note that one training session in the programme schools can cover more than one topic.

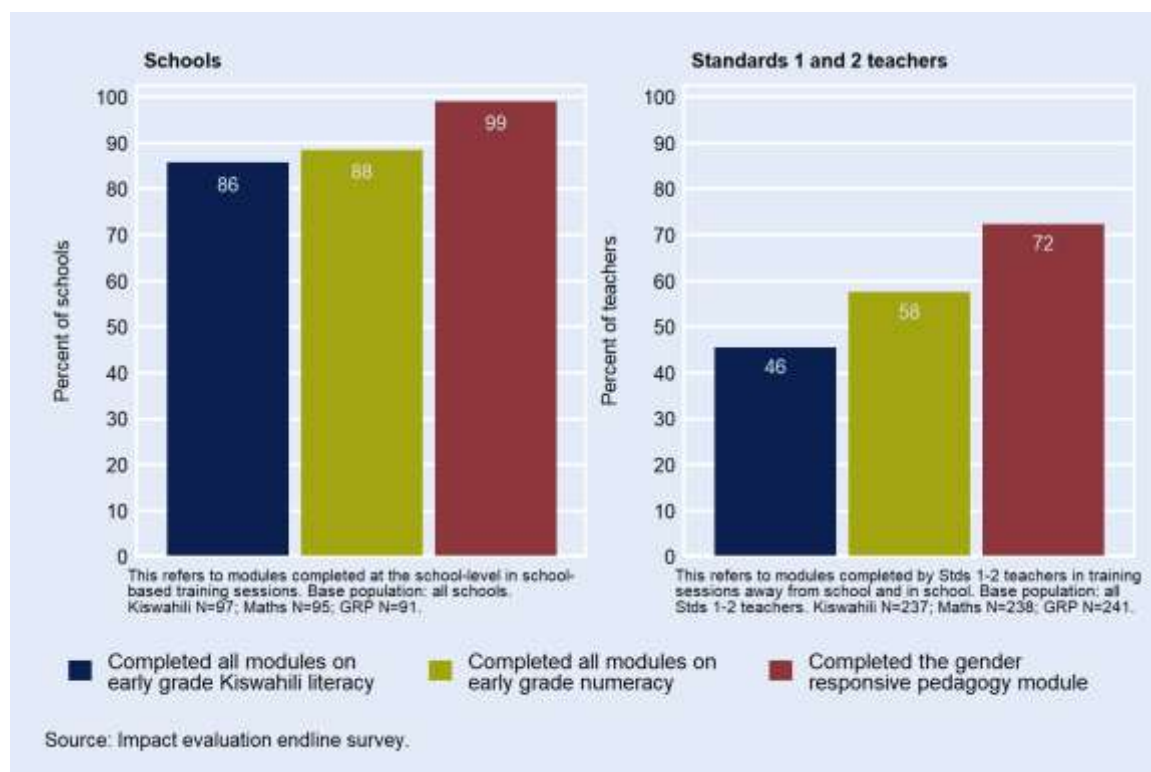
Head teacher	9	94
Facilitators of Kiswahili school-based training sessions (mean % facilitators of Kiswahili sessions)		
Attended at least one Kiswahili literacy training session away from school	88.2	91
Facilitators of numeracy school-based training sessions (mean % facilitators of numeracy sessions)		
Attended at least one numeracy training session away from school	85.12	90
Facilitators of gender-responsive pedagogy school-based training sessions (mean % facilitators of GRP sessions)		
Attended at least one gender-responsive pedagogy training session away from school	79.23	86

Sources: Impact evaluation endline survey (INCO interview).
 Notes: (1) The percent of school-based training sessions is reported as the mean percent of school-based training sessions taken over all schools. (2) The percent of facilitators is reported as the mean percent of facilitators taken over all schools. These estimates are weighted by the proportion of school-based sessions that a given facilitator has facilitated since baseline.

Module completion

The majority of schools (86%) report that they have covered and completed all the 13 Kiswahili literacy modules in their school-based training sessions over the years. Similarly, 88% report completing all the nine numeracy training modules and nearly all schools (99%) report completing the gender-responsive pedagogy module. On the other hand, completion rates of these modules by Standard 1 and 2 teachers, which is the target group of teachers for the EQUIP-T in-service training, is much lower (Figure 32). Only 46% of Standard 1 and 2 teachers report completing all Kiswahili modules as part of their training away from school and in school, 58% have completed all numeracy modules, and 72% have completed the gender-responsive pedagogy module. On average, teachers have completed nine out of the 13 Kiswahili modules and seven out of the nine numeracy modules.

Figure 32: Completion of EQUIP-T training modules by schools and early grade teachers



5.2.2 Profile of INCO

Almost all programme schools (97%) have an INCO (Table 29). INCOs have been in post, on average, for 2.6 years and only 43% of schools have had the same INCO in post since January 2015 when the school-based component of the EQUIP-T in-service training was expected to start in full capacity. This is a high turnover in the INCO post (see section 5.4.2 for more detail on INCO turnover).

Table 29: INCO post at school

	Endline	
	Estimate	N
School has an INCO (% schools)	96.74	99
Number of years current INCO has been in post at school (mean years)	2.58	95
Current INCO has been in post since January 2015 or earlier (% INCOs)	42.65	95
Sources: Impact evaluation endline survey (INCO interview).		
Notes: (1) This is for all schools.		

INCOs are mostly male (68%) and are 35 years old on average. They have 11 years of teaching working experience, on average, and 8 years of teaching experience at their current school. The vast majority (95%) have a certificate in education as the highest professional qualification. On the other hand, the majority (90%) have completed Form 4 as their highest academic qualification apart from their professional qualification, and about 6% have attained primary education as their highest academic qualification (Table 30).

INCOs have, on average, 27 teaching periods in the current school term. The majority of INCOs also hold other positions in the school: 28% are also the assistant head teachers and 34% are the academic masters. On top of their teaching and sometimes other administrative and SLM duties, INCOs are expected to attend all in-service training sessions away from school and the quarterly ward cluster reflection meetings, as well as required to facilitate all school-based in-service training sessions on a biweekly basis.

Table 30: Profile of INCO

	Endline	
	Estimate	N
Female (% INCOs)	32.37	95
Age (mean years)	35.14	95
Time working as a teacher (mean years)	11.25	95
Time teaching at current school (mean years)	7.86	95
Highest professional education qualification (% INCOs)		
Bachelors of Education or higher	0	95
Diploma or advanced diploma	4.82	95
Certificate in education	95.18	95
Other professional qualification	0	95
No professional qualification	0	95
Highest academic qualification apart from professional education qualification (% INCOs)		
Bachelors or higher	0	94
Diploma or advanced diploma	0	94
Certificate	0	94
Form 6	4.19	94
Form 4	89.98	94

Primary	5.83	94
Other	0	94
Teaches maths to any standard this school term (% INCOs)	81.17	92
Teaches Kiswahili to any standard this school term (% INCOs)	74.83	92
Teaches maths or Kiswahili to any standard this school term (% INCOs)	90.85	92
Teaches Standards 1-3 this school term (% INCOs)	57.86	95
Number of teaching periods per week (mean periods)	27.4	91
Holds other positions at school (% INCOs)		
Head teacher	0.76	94
Assistant head teacher	28.22	94
Academic master	33.72	94
Attended at least one EQUIP-T training session away from school on (% INCOs)		
Early grade Swahili literacy	71.84	95
Early grade numeracy	91.77	95
Attended all EQUIP-T training sessions away from school on (% INCOs)		
Early grade Swahili literacy	42.86	94
Early grade numeracy	70.87	94
Gender-responsive pedagogy	75.97	94
Sources: Impact evaluation endline survey (INCO interview).		
Notes: (1) This is for all INCOs.		

Of all INCOs, 58% are teaching Standards 1 to 3 in the current school term, while 81% are teaching maths to any standard in the current term and 75% are teaching Kiswahili to any standard. There are 9% of INCOs who are not teaching maths or Kiswahili to any standard in the current school term, despite INCOs being required to attend all residential in-service training at the district level and to facilitate all school-based in-service training including on Kiswahili literacy and numeracy topics.

INCOs are expected to have attended all EQUIP-T in-service teacher training away from school since baseline. However, less than half of INCOs (43%) have attended all training sessions away from school on early grade Kiswahili literacy while 28% have not attended any session on early grade Kiswahili literacy. Attendance of early grade numeracy training away from school is higher but still low: 71% have attended all training sessions away from school on early grade numeracy while 8% have not attended any session. About a quarter of INCOs (24%) have not attended the training session away from school on gender-responsive pedagogy.²¹ This shortfall in coverage may in part be explained by the high INCO turnover discussed above. Regardless of the reason, this could negatively affect the quality of school-based training as the INCOs are expected to facilitate all these sessions and provide support and mentoring to their peers.

5.2.3 Difficulties with EQUIP-T training

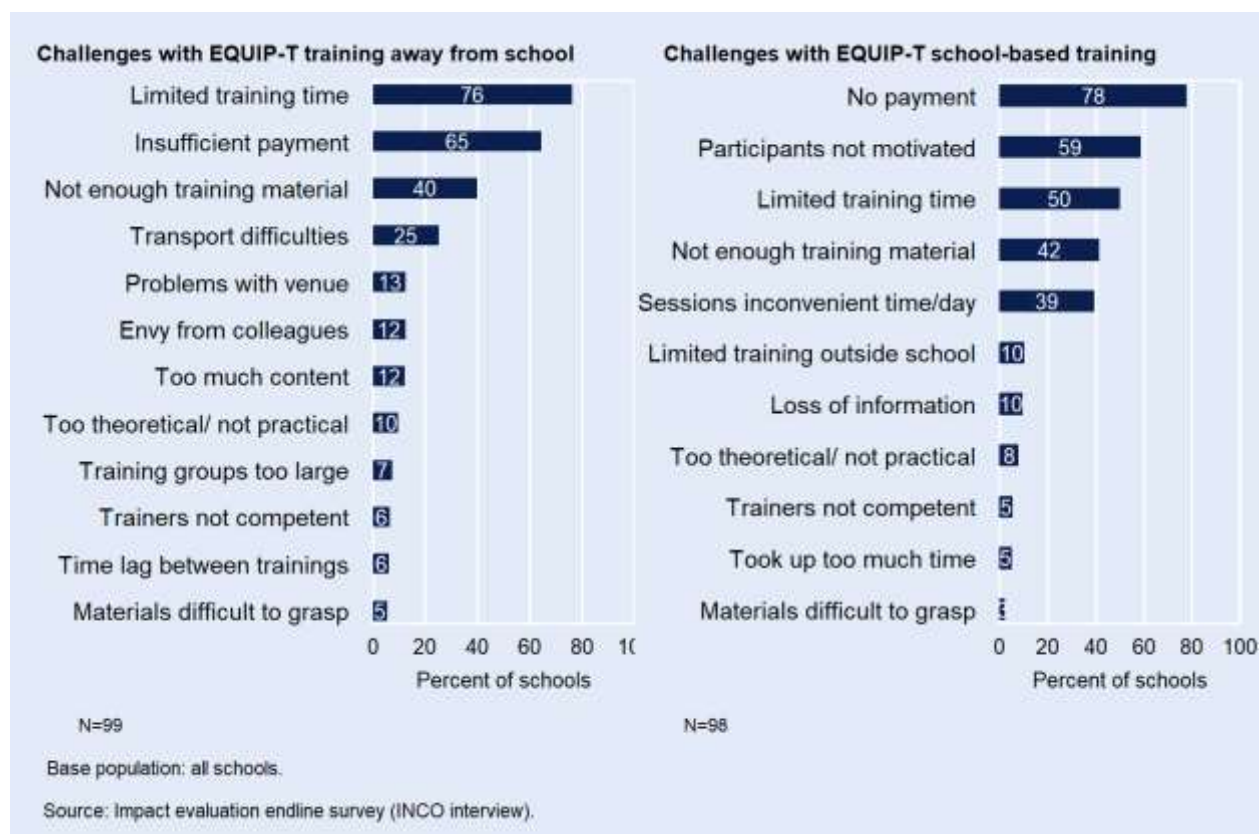
Information on the difficulties teachers face with the EQUIP-T training was collected at school-level and teacher-level. The former was collected in the interview with the INCO and group of other teachers, while the latter was collected in the individual interviews with Standards 1 to 3 teachers.

Figure 33 presents the challenges that were reported at the school-level with the EQUIP-T training away from school and in school. The most common challenge reported with training away from school

²¹ This is based on school-level data of participants who attended all training sessions away from school since 2014. Therefore, it does not take into account that some of the INCOs could have attended some of the earlier training sessions as part of their previous schools. However, this will be a small share given that 87% of INCOs had been teaching at their current school since baseline.

is the limited training time (76%), followed by insufficient payment (65%), not enough training material (40%), transport difficulties (25%), and problems with the venue (13%). Most common challenges with school-based training are no payment (78%), participants not motivated (59%), limited training time (50%), not enough training material (42%), and sessions scheduled at inconvenient times (39%).

Figure 33: Challenges reported by schools with EQUIP-T training



When asked what EQUIP-T could do differently to improve the training, the majority of schools (64%) suggest supplying more training materials, followed by having an allowance for the school-based training (63%), training all teachers away from school (63%), increasing the allowance for the training away from school (30%), and training when the school is closed (22%) (Table 31).

Table 31: Suggested improvements by schools to the EQUIP-T in-service training

	Endline	
	Estimate	N
Improvements to the in-service training for teachers (% schools)		
Supply more training materials	64.15	98
Allowance for school based training	62.75	98
Train all teachers away from school	62.65	98
More allowance for residential training	30.3	98
Train when school closed	21.97	98
Less content / more time	12.24	98
More training for inspectors / WEOs / DEOs	8.87	98
Reduce other teacher tasks	7.07	98
Other	21.13	98

Sources: Impact evaluation endline survey (INCO interview).
Notes: (1) This is for all schools.

Challenges reported by Standards 1 to 3 teachers mirror those that were reported at the school-level (Table 32) with the five most common challenges being limited training time (28%), no or insufficient allowance (24%), not enough training material (19%), sessions scheduled at inconvenient times (16%) and transport difficulties (15%). There are a few notable changes since midline. Significantly more teachers at endline report facing difficulties with the EQUIP-T training than at midline, and significantly more teachers find limited training time as a challenge; while significantly less teachers at endline report that materials being difficult to understand is a challenge.

Table 32: Challenges reported by Standards 1 to 3 teachers with EQUIP-T training (trends in programme areas)

	Midline		Endline		Difference
	Estimate	N	Estimate	N	ML-EL
Difficulties with EQUIP-T training (% Stds 1-3 teachers who found EQUIP-T training useful)					
None	44.88	303	24.28	384	-20.6***
Limited training time	10.9	303	28.18	384	17.27***
No / insufficient allowance	17.42	303	24.43	384	7
Not enough training material			19.35	384	
Sessions inconvenient time/day			16.28	384	
Transport difficult/venue too far away			15.47	384	
Too much content	10.48	303	11.9	384	1.42
Time lag between trainings			6.19	384	
Envy from colleagues			4.9	384	
Materials difficult to understand	10.7	303	3.81	384	-6.89***
No / insufficient direct training outside school			3.34	384	
Content not completed			2.62	384	
Too theoretical/ not practical	5.05	303	2.16	384	-2.89
Other	21.12	303	4.76	384	-16.36***

Sources: Impact evaluation midline and endline surveys (teacher interview).

Notes: (1) Asterisks indicate statistical significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. (2) This is for all interviewed teachers who teach maths or Kiswahili to Standards 1-3 and who attended EQUIP-T training in 2016-17 and found it useful. (3) The missing cells correspond to data that was not collected at midline.

5.2.4 Teacher access to curriculum, syllabi and teacher guides

Teachers' access to essential material related to teaching the new 3Rs curriculum is inadequate and has deteriorated since midline (Table 33). Of all Standard 1 and 2 teachers, 26% have none or limited access to the new Standard 1 and 2 curriculum (compared to 17% at midline), while 15% of Standard 1 teachers have none or limited access to Standard 1 syllabi (11% at midline) and 16% of Standard 2 teachers have none or limited access to the Standard 2 syllabi (5% at midline). Access to teacher guides is also poor or non-existent for a sizeable minority of teachers. Of all Standard 1 and 2 teachers, about 12-14% have limited access to teachers' guides for reading, writing and arithmetic compared to 6-8% at midline, while about 14-18% have no access at all to these materials compared to 7-10% at midline.

Table 33: Teacher access to curriculum, syllabi and teacher guides, self-reported by teachers (trends in programme areas)

	Midline		Endline		Difference
	Estimate	N	Estimate	N	ML-EL
Teacher has access to Standards 1 and 2 curriculum (% Standard 1 and 2 teachers)					
Yes, good access	82.71	213	73.98	242	-8.74
Yes, limited access	12.59	213	16.92	242	4.33
No access	4.7	213	9.11	242	4.41
Teacher has access to syllabi for Standard 1 (% Standard 1 teachers)					
Yes, good access	88.95	123	84.23	139	-4.72
Yes, limited access	8.26	123	12.7	139	4.44
No access	2.79	123	3.07	139	.28
Teacher has access to syllabi for Standard 2 (% Standard 2 teachers)					
Yes, good access	94.72	123	83.92	138	-10.81**
Yes, limited access	3.25	123	12.1	138	8.85**
No access	2.02	123	3.98	138	1.96
Teacher has access to teachers' guide for reading (% Standard 1 and 1 Kiswahili teachers)					
Yes, good access	84.55	201	70.84	228	-13.71**
Yes, limited access	7.03	201	13.34	228	6.31
No access	8.42	201	15.82	228	7.4
Teacher has access to teachers' guide for writing (% Standard 1 and 2 Kiswahili teachers)					
Yes, good access	82.95	201	69.99	227	-12.96**
Yes, limited access	8.4	201	12.37	227	3.98
No access	8.66	201	17.64	227	8.98*
Teacher has access to teachers' guide for arithmetic (% Standard 1 and 2 maths teachers)					
Yes, good access	83.35	200	71.73	224	-11.63*
Yes, limited access	6.17	200	14.25	224	8.08**
No access	10.48	200	14.02	224	3.54

Sources: Impact evaluation midline and endline surveys (teacher interview).
Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1. (2) This is for all interviewed teachers who teach maths or Kiswahili to Standards 1-3.

5.2.5 Early grade teacher background characteristics

The personal characteristics, work experience and tenure of teachers of Standards 1 to 3 have to a large extent not significantly changed since baseline or midline (Table 34). At endline, 58% of teachers are female similar to baseline and midline levels. On average, teachers are slightly younger at endline and midline (37 years old) than at baseline (40 years). The average time worked as a teacher is 13 years at endline (lower than the 16 years at baseline but with weak significance) and as a teacher at the current school 8 years.

Almost all teachers (93%) have a certificate in education as their highest professional qualification, and that has not changed since baseline or midline. There is, however, a higher share of teachers at endline with a diploma or advanced diploma in education (weak significance). Besides their professional education qualification, the majority of teachers (88%) have attained Form 4 as their highest academic qualification and that has increased significantly since baseline. Almost one tenth of all teachers (9%) do not have an academic qualification, apart from their professional qualification, above primary school.

Table 34: Background characteristics and qualifications of Standards 1 to 3 teachers (trends in programme areas)

	Baseline		Midline		Endline		Difference	Difference
	Estimate	N	Estimate	N	Estimate	N	BL-EL	ML-EL
Female (% Stds 1-3 teachers)	55.61	327	58.43	384	57.67	418	2.06	-0.75
Age (mean years)	39.62	327	37.47	384	36.77	418	-2.85**	-0.7
Time working as a teacher (mean years)	15.79	327	13.82	384	13.22	418	-2.57*	-0.6
Time working as a teacher at current school (mean years)	8.39	327	7.76	384	7.76	418	-0.63	-0.01
Highest professional education qualification (% Stds1-3 teachers)								
Bachelors of Education or higher	0.42	326	0.18	384	0.55	418	0.12	0.37
Diploma or advanced diploma	1.52	326	1.83	384	4.35	418	2.83*	2.53*
Certificate in education	94.18	326	96	384	93.35	418	-0.84	-2.65
Other professional qualification	3.35	326	1.26	384	0	418	-3.35*	-1.26*
No professional qualification	0.53	326	0.73	384	1.76	418	1.23	1.02
Highest academic qualification apart from professional education qualification (% Stds1-3 teachers)								
Bachelors degree or higher	0.42	327			0	417	-0.42	
Diploma or advanced diploma	0.79	327			0.45	417	-0.34	
Certificate	6.1	327			0.72	417	-5.38***	
Form 6	2.47	327			1.97	417	-0.5	
Form 4	76.33	327			88.3	417	11.98***	
Primary school	13.72	327			8.56	417	-5.16*	
Sources: Impact evaluation baseline, midline and endline surveys (teacher interview).								
Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1. (2) This is for all interviewed teachers who teach maths or Kiswahili to Standards 1-3. (3) Academic qualifications of teachers at midline are not presented as it is not comparable to the baseline or endline data due to a change in administration of this question.								

5.2.6 Teacher performance indicators by gender and age

Differences in a range of teacher-level output and intermediate outcome indicators by the gender and age of teachers were examined. The indicators include completion of the EQUIP-T in-service training modules, attendance of school-based training sessions, use of inclusive and gender-responsive teaching practices in the classroom, and use of selected positive teaching practices during the introductory, middle and concluding stages of a lesson.

At endline, there are no statistically significant differences in teacher output indicators by the gender or age of teachers (Table 35). Similarly, with the exception of more male teachers holding a plenary to summarise and extend learning at the end of a lesson than female teachers, there are no significant differences in the rest of the intermediate outcomes by gender or age (Table 36 and Table 37).

Table 35: Teacher output indicators at endline by gender and age

	Endline			
	Males	Females	Aged <35	Aged ≥35
Completed all EQUIP-T in-service training modules on early grade Kiswahili literacy (% Standards 1 and 2 teachers)	48.6	44.0	40.6	50.2
<i>N</i>	(73)	(164)	(129)	(108)
Completed all EQUIP-T in-service training modules on early grade numeracy (% Standards 1 and 2 teachers)	54.4	59.1	55.0	59.9
<i>N</i>	(74)	(164)	(129)	(109)
Completed the EQUIP-T in-service training module on gender-responsive pedagogy (% Standards 1 and 2 teachers)	72.7	72.2	72.6	72.2
<i>N</i>	(76)	(165)	(132)	(109)
Attended all EQUIP-T school-based training sessions in 2016-2017 (% Standards 1 and 2 teachers who attended any EQUIP-T school-based training)	50.9	52.7	51.6	52.7
<i>N</i>	(63)	(150)	(117)	(96)

Source: Impact evaluation endline survey (teacher interview).
Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1. (2) This is for all interviewed teachers who teach maths or Kiswahili to Standards 1 and 2.

Table 36: Teacher intermediate outcome indicators at baseline, midline, and endline by gender (trends in programme areas)

	Baseline		Midline		Endline	
	Males	Females	Males	Females	Males	Females
Teachers' interactions with pupils in the classroom is gender-balanced (% lessons)	53.6	53.8	76.0	57.1***	72.2	65.5
<i>N</i>	(79)	(111)	(81)	(138)	(78)	(114)
Teacher interacts with at least one pupil from all six areas of the classroom (% lessons)	69.6	48.5**	82.9	76.8	73.7	77.2
<i>N</i>	(79)	(111)	(83)	(142)	(78)	(114)
Teacher demonstrates at least 7 positive teaching practices (% lessons)	69.8	66.3	56.2	56.3	39.2	40.5
<i>N</i>	(82)	(114)	(83)	(142)	(78)	(114)
Teacher states objectives of lesson during the introductory stage of a lesson (% lessons)	69.5	76.0	51.1	42.3	31.9	39.1
<i>N</i>	(82)	(114)	(83)	(142)	(78)	(115)
Teacher holds a plenary to summarise and extend learning during the concluding stage of a lesson (% lessons)	63.2	64.1	42.9	29.4**	14.4	5.3**
<i>N</i>	(82)	(114)	(83)	(142)	(78)	(114)
Teacher uses different instructional materials during the lesson (% lessons)	45.4	44.8	45.3	59.1*	49.1	61.7
<i>N</i>	(82)	(114)	(83)	(142)	(78)	(115)
Teacher provides feedback to pupils on their individual work during the lesson (% lessons)	59.4	58.6	60.4	70.5	73.6	70.2
<i>N</i>	(82)	(114)	(83)	(142)	(78)	(115)

Source: Impact evaluation baseline, midline and endline surveys (lesson observations).

Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1. (2) This is for all lessons observed disaggregated by the gender of the teacher.

Table 37: Teacher intermediate outcome indicators at baseline, midline, and endline by age (trends in programme areas)

	Baseline		Midline		Endline	
	Aged <35	Aged ≥35	Aged <35	Aged ≥35	Aged <35	Aged ≥35
Teachers' interactions with pupils in the classroom is gender-balanced (% lessons)	61.9	52.2	63.1	66.4	66.4	69.7
<i>N</i>	(61)	(111)	(109)	(104)	(97)	(91)
Teacher interacts with at least one pupil from all six areas of the classroom (% lessons)	56.9	61.0	78.7	78.2	78.2	73.2
<i>N</i>	(61)	(111)	(112)	(107)	(97)	(91)
Teacher demonstrates at least 7 positive teaching practices (% lessons)	84.9	66.1***	58.9	51.8	44.1	37.6
<i>N</i>	(62)	(114)	(112)	(107)	(97)	(91)
Teacher states objectives of lesson during the introductory stage of a lesson (% lessons)	93.5	65.1***	46.8	46.6	39.6	33.9
<i>N</i>	(62)	(114)	(112)	(107)	(97)	(92)
Teacher holds a plenary to summarise and extend learning during the concluding stage of a lesson (% lessons)	83.7	61.2***	34.8	33.2	9.3	9.3
<i>N</i>	(62)	(114)	(112)	(107)	(97)	(91)
Teacher uses different instructional materials during the lesson (% lessons)	50.9	46.2	55.9	50.0	60.4	52.6
<i>N</i>	(62)	(114)	(112)	(107)	(97)	(92)
Teacher provides feedback to pupils on their individual work during the lesson (% lessons)	69.3	57.2	64.2	65.5	66.4	77.3
<i>N</i>	(62)	(114)	(112)	(107)	(97)	(92)

Source: Impact evaluation baseline, midline and endline surveys (lesson observations).

Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1. (2) This is for all lessons observed disaggregated by the age of the teacher.

5.3 SLM

5.3.1 Teacher management: teacher absence

Most head teachers consider teacher attendance at their school 'good' or 'very good' (94% at midline and 92% at endline). At endline, the most common reasons for teachers being absent from school according to head teachers are: illness (86%); family reasons (58%); official education work / meeting (40%); transport problems (21%); collecting their salary (20%); other official government work (18%); lack of motivation (13%); and attending training (10%). There have been some significant changes since midline. The proportion of head teachers reporting transport as a reason for teacher absence has more than doubled to 21% at endline. The reason for this is not clear as the average time to school (18 minutes) for teachers has not changed significantly since midline. Over the same period,

collecting salaries has declined from 39% to 21%, and doing other private work has decreased from 11% to less than 1%. There have recently been changes in technology with regard to money transfers so that teachers can use their phones to transfer money from their bank, which means they do not physically need go to the bank to collect their salaries.²² Some possible reasons for the decline in other private work include public examples of mass disciplinary actions and the increase in WEO visits to schools (see section 5.8 in Volume I).

The majority of head teachers (71%) report that teachers are sometimes absent from the classroom, which is in line with the high teacher classroom absence in the programme schools (see Section 4.7 in Chapter 4 in Volume I). The main reasons reported by head teachers for teachers at their schools being absent from the classroom are a mix of school and individual ones. Large work load (54%); illness (18%); lack of motivation (17%); meeting with teachers (17%); meeting with head teacher (14%); and feeling tired / exhausted (8%). Two reasons have risen significantly in importance since midline: lack of motivation from 6% to 17%, and meeting with teachers from 4% to 17%. It is not clear what types of teacher meetings this refers to but it may include EQUIP-T school-based training sessions and SPMMs.

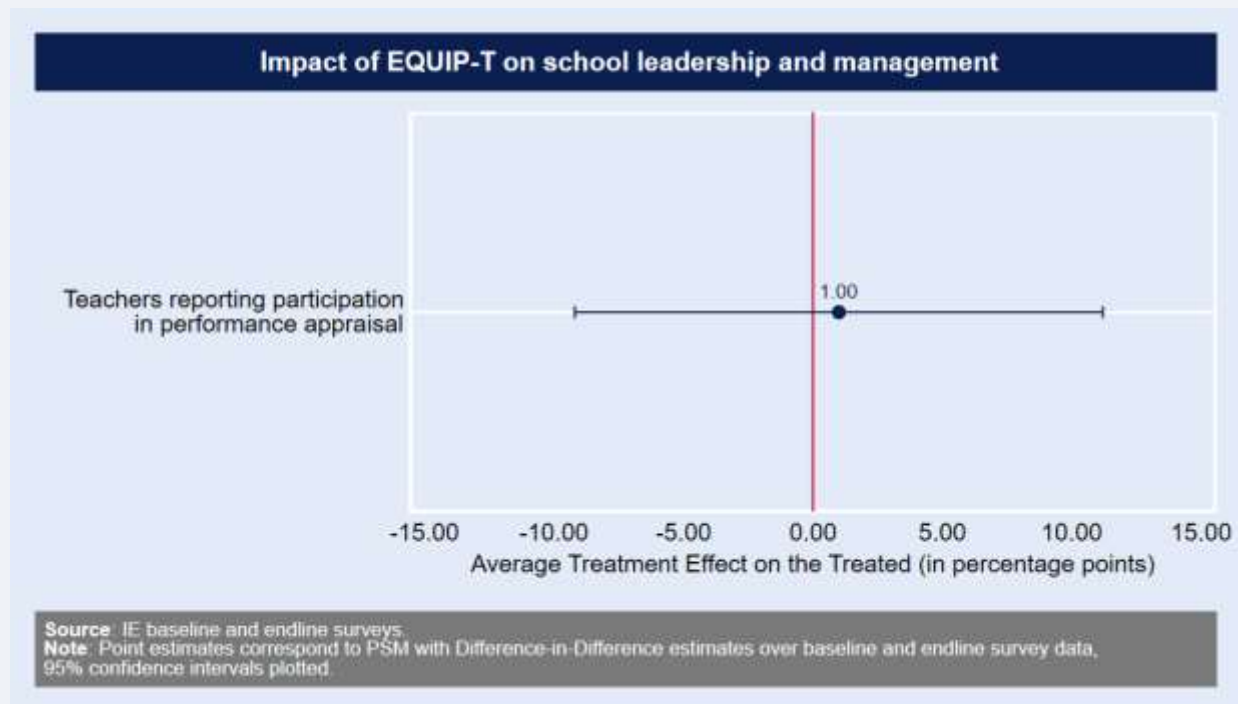
5.3.2 Teacher management: EQUIP-T impact on teachers receiving performance appraisals

Head teachers' use of performance appraisals did not change significantly since baseline. At baseline, midline and endline, only around 28-30% of Standards 1 to 3 teachers reported receiving at least one performance appraisal during the previous school year to discuss their performance and professional development needs. The lack of change for this particular practice is not unexpected, because although EQUIP-T developed a Leadership Competency Framework (see Volume I Annex B.3), implementation stalled, and the approach was later superseded by materials developed together with ADEM (these were at pilot phase at the time of the endline survey). Consistent with the pace and challenges of implementation, there is no evidence of any impact of EQUIP-T as a whole on the use of teacher performance appraisals (see Box 5).

²² Related to this, banks and phone companies have extended their financial services in rural Tanzania so that there is now much less need to travel for drawing money.

Box 5: EQUIP-T impact on teacher performance appraisals

The figure below shows the ATT detected on the proportion of teachers reporting participation in performance appraisals (in percentage points). It compares changes in EQUIP-T schools to changes in control schools between baseline and endline.



There is no evidence of a positive impact of EQUIP-T on teachers' participation in performance appraisals. This is demonstrated by the figure above, which shows that although the impact estimate is positive in sign, the 95% confidence interval largely overlaps with zero. This absence of programme impact is confirmed by an array of estimation and robustness checks. Although the sample of teachers used for this impact estimation is small, the matching procedure performs satisfactorily well and the impact results are thus reliable. This means that it can be conclusively inferred that EQUIP-T has had no impact on teacher participation in performance appraisals. This finding is in line with the midline result, which found no conclusive evidence of programme impact, and is consistent with the lack of significant change from baseline to endline in the descriptive analysis of the proportion of teachers reporting participation in performance appraisals.

5.4 Turnover in education posts at the school and ward level

This section provides supplementary descriptive trend analysis of indicators of turnover in the teacher, head teacher, INCO and WEO posts in the programme areas. Turnover in this report is defined as the rate in which staff leave their current posts that are defined by their role (that is: teacher, head teacher, INCO, WEO) and place of work. Therefore, staff who remain in the same role but change their place of work (e.g. teachers who transfer to another school between midline and endline) are included in the turnover rate. Similarly, staff who remain in their place of work but change roles (e.g. head teachers who are demoted to a teacher within the same school; or teachers who are promoted to head teacher within the same school) are also included in the turnover rate. Note that teachers who move from early grade teaching to upper grade teaching within the same school are not included in the turnover rate, as their role ('teacher') has not changed.

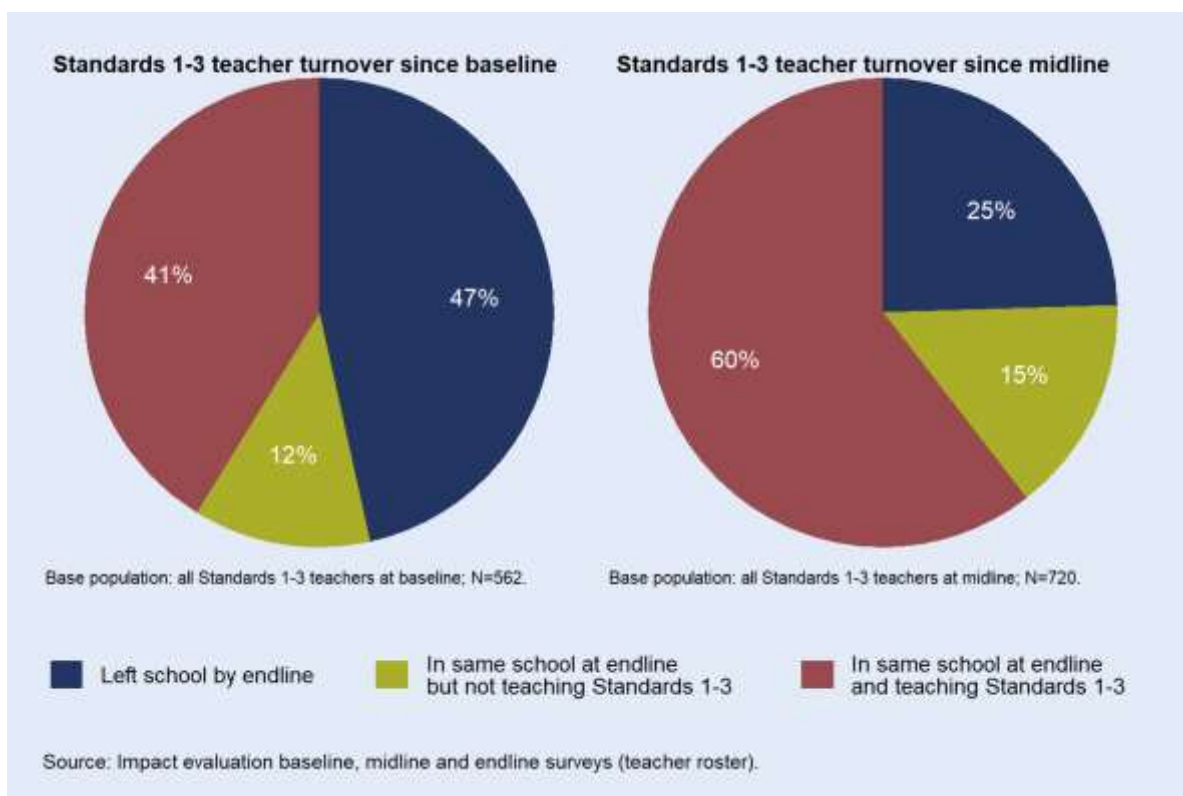
5.4.1 Teacher turnover

Turnover for Standards 1 to 3 teachers and movement from lower to upper grades

Teacher turnover since midline and particularly since baseline is very high (Figure 34). Of all Standards 1 to 3 teachers at midline, 25% are no longer teaching at the same school at endline. Looking over a four-year period, almost half (47%) of all Standards 1 to 3 teachers at baseline have left the school by endline.

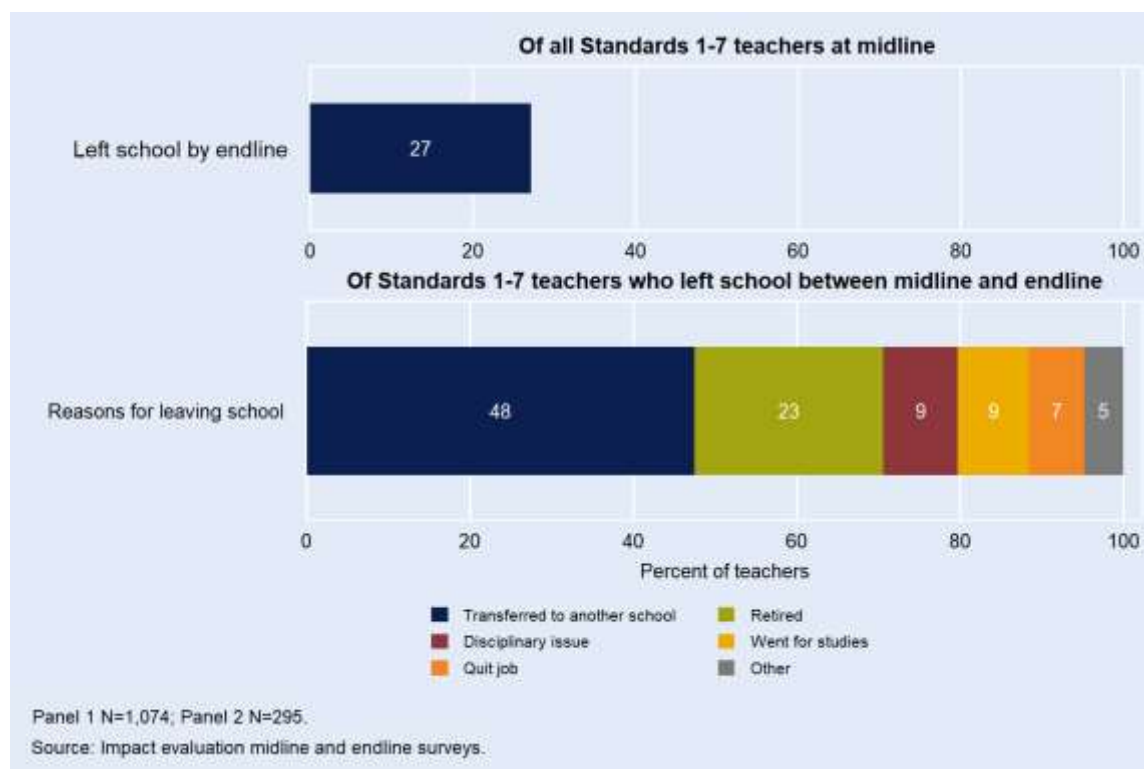
In addition to the large shares of teachers leaving their schools, there is also movement of teachers between the standards being taught. Of all Standards 1 to 3 teachers at midline, 15% are still at the same school at endline but are no longer teaching Standards 1 to 3. Similarly, between baseline and endline, 12% of Standards 1 to 3 teachers are no longer teaching early grade standards. Combining turnover with movement to upper standards leads to only 41% of Standards 1 to 3 teachers at baseline who are still teaching these standards in the same schools at endline.

Figure 34: Standards 1 to 3 teacher turnover since baseline and midline (trends in programme areas)



Turnover for Standards 1 to 7 teachers and reasons for leaving the school

High turnover is prevalent among all teachers at the school and not just early grade teachers. Of all Standards 1 to 7 teachers at midline, 27% had left their schools by endline (Figure 35).

Figure 35: Standards 1-7 teacher turnover between midline and endline

The main reasons for the teacher turnover between midline and endline are: transferring to another school (48%)²³, retiring (23%), disciplinary issues (9%), seeking further studies (9%), and quitting their job (7%). Put in another way, this means that of all teachers who had left the school between midline and endline, 44% of them had left the teaching profession altogether while 56% are still teaching or working in education.²⁴

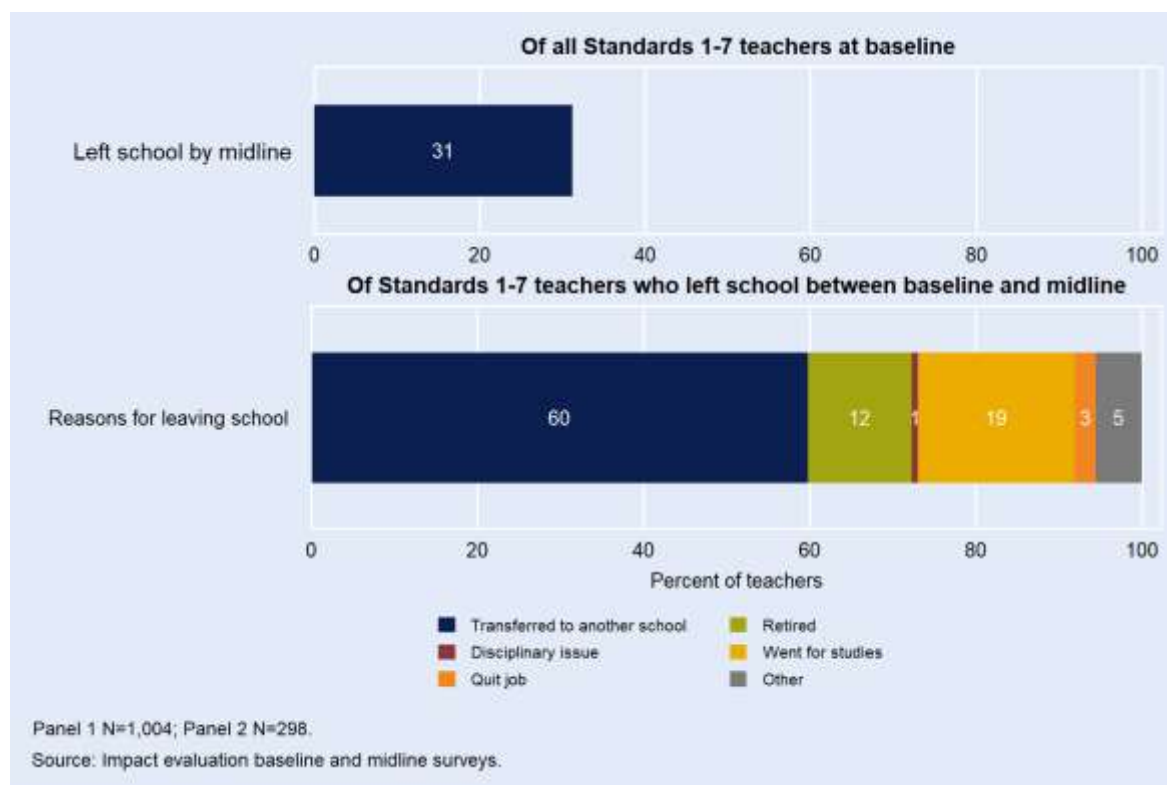
The teacher turnover rate between baseline and midline is 31% - not statistically different to the rate between midline and endline (Figure 36). There are, however, a few notable differences in the reasons for turnover. Significantly more teachers at baseline had left to seek further studies or to be transferred to another school, while more teachers at midline had left because of retirement or disciplinary issues. The latter is most likely associated with the mass dismissal of teachers that took place nationwide in 2017 due to the discovery of fake education certificates by some teachers.

Retirement is one of the main reasons for turnover. Of all Standards 1 to 3 teachers, 5% will reach the official retirement age of 60 years within the next year and therefore the knowledge and skills gained by these teachers from the EQUIP-T in-service training will only benefit pupils for a short period of time.

²³ The survey did not collect data on the districts or regions that teachers had transferred to.

²⁴ The 56% of teachers who are still working in the teaching profession may be overestimated as it assumes that all teachers who left to seek further studies will return to the teaching profession, which might not be the case.

Figure 36: Standards 1-7 teacher turnover between baseline and midline



Differences in teacher turnover across gender and age

There are differences in teacher turnover rates between males and females and across different age groups. The turnover rate for female teachers is significantly higher than that for male teachers (Table 38). Of all female teachers at midline, 30% had left the school by endline compared to 24% of male teachers at midline who had left by endline. The gap is bigger for Standard 1 to 3 teachers: 28% of female teachers at midline had left the school by endline compared to 20% of male teachers.

Table 38: Turnover of teachers between midline and endline disaggregated by gender

	Endline					
	Total	N	Female	N	Male	N
Standard 1 to 7 teacher no longer teaching at same school at endline (% Standard 1 to 7 midline teachers)	27.22	1074	30.21	561	24.23*	513
Standard 1 to 3 teacher no longer teaching at same school at endline (% Standard 1 to 3 midline teachers)	24.5	720	28.37	409	20.07**	311

Sources: Impact evaluation baseline, midline and endline surveys (teacher roster).
Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Teachers in the highest age group (50 and above) have the highest turnover rates, and this is the case for both Standard 1 to 7 teachers and Standard 1 to 3 teachers (Table 39). Of all midline teachers aged 50 and above, 40% had left the school by endline. This is significantly higher than the turnover rate for teachers in the lower age groups which ranges from 23% to 25%. Similarly, 39% of Standard 1 to 3 teachers aged 50 or above at midline had left the school by endline, compared to 18% to 25% for teachers in the lower age groups. This is most likely due to the fact that some teachers in the age group 50 and above retire and therefore leave the school.

Table 39: Turnover of teachers between midline and endline disaggregated by age

	Endline									
	Total	N	20-29	N	30-39	N	40-49	N	50+	N
Standard 1 to 7 teacher no longer teaching at same school at endline (% Standard 1 to 7 midline teachers)	27.22	1074	23.14	489	25.27	269	22.96	112	40.43	182
Standard 1 to 3 teacher no longer teaching at same school at endline (% Standard 1 to 3 midline teachers)	24.5	720	19.27	338	24.68	172	18.44	68	38.58	129

Sources: Impact evaluation baseline, midline and endline surveys (teacher roster).

New teachers and their previous teaching posting

While a large share of teachers have left their schools between baseline and endline, schools have also been recruiting new teachers. A fifth of all Standards 1 to 7 teachers at endline had joined the school since midline. This is significantly lower than the share of Standards 1 to 7 teachers at midline that had joined the school since baseline (30%). The rate of recruitment is similar for early grade teachers who teach Kiswahili or maths. The share of interviewed Standards 1 to 3 teachers at endline that had joined the school in the last two years is 15%, and also significantly lower than the share (29%) of interviewed Standards 1 to 3 teachers at midline that had joined the school since baseline (Table 40).

The vast majority (94%) of Standard 1 to 3 teachers that had joined their current school between midline and endline had been previously teaching in another school, while only 6% had joined the school on their first teaching post. This is significantly different than at midline, where the majority (64%) of new Standard 1 to 3 teachers between baseline and midline did not have any previous teaching experience.

Almost all new Standards 1 to 3 teachers who were teaching somewhere else prior to joining the school between midline and endline came from EQUIP-T regions, and the majority came from within the same district. Three quarters came from another school in the same district, while 15% came from another school in a different district but same region and 9% came from another school in another region. While the overall share of new teachers joining from other schools in the same region has increased between baseline and midline (79%), and midline and endline (91%), this change is not statistically significant. However, a notable change is the significant and large increase in new teachers coming from other schools in the same district (46% between baseline and midline, compared to 75% between midline and endline).

Table 40: New teachers employed at the school (trends in programme areas)

	Midline		Endline		Difference
	Estimate	N	Estimate	N	ML-EL
Teacher joined school since the previous round (% Stds 1-7 teachers)	29.69	1022	20.24	965	-9.46***
Teacher joined school since the previous round (% interviewed Stds 1-3 teachers)	29.26	384	15.09	418	-14.17***
Previous job before becoming a teacher at current school (% Stds 1-3 teachers who joined school since last round)					
Teacher in another school	35.84	114	93.95	63	58.11***
None or other job not in teaching	64.16	114	6.05	63	-58.11***
Location of previous teaching job (% Standards 1-3 teachers who joined school since last round)					
Another school in same district	45.88	42	75.18	58	29.3*
Another school in same region but different district	33.47	42	15.34	58	-18.12

Another school in another region	20.66	42	9.48	58	-11.17
Sources: Impact evaluation midline and endline surveys (teacher interview and teacher roster).					
Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1.					

5.4.2 INCO turnover

Turnover in the INCO post in the programme schools is very high (Table 41). Only 43% of schools have had the same INCO in post since January 2015, when school-based in-service training started. On average, an INCO has been in post for 2.6 years.

In schools where the current INCO has not been in post since January 2015, the majority (64%) had a previous INCO in that post, while 16% of schools had two or more previous INCOS. A fifth of those schools though did not have a previous INCO meaning that since January 2015 those schools did not have any INCO for a certain period of time.

Table 41: Turnover in INCO post

	Endline	
	Estimate	N
School has an INCO (% schools)	96.74	99
Current INCO has been in post since January 2015 or earlier (% current INCOS)	42.65	95
Number of years current INCO has been in post at school (mean years)	2.58	95
Number of teachers who held the INCO post before current INCO was in post (% schools where current INCO has not been in post since Jan 2015)		
None	20.36	53
One	63.97	53
More than one	15.68	53
Sources: Impact evaluation endline survey (INCO interview).		
Notes: (1) This is for all schools and current INCOS.		

5.4.3 Head teacher turnover

Head teacher turnover in the programme schools has been extremely high (Table 42). By midline, 46% of head teachers at baseline were no longer in their posts (either left the school or got demoted to another post in the same school). The head teacher turnover rate between midline and endline was similar at 51%. Over a four year period, just over a quarter (26%) of head teachers at baseline are still head teachers at the same school by endline.

Table 42: Head teacher turnover and reasons (trends in programme areas)

	Midline		Endline		Difference
	Estimate	N	Estimate	N	ML-EL
Head teacher no longer head teacher in same school by the next survey round (% head teachers)	45.77	100	50.97	100	5.2
Head teacher no longer head teacher in same school since baseline in 2014 (% baseline head teachers)			74.18	100	
Reasons for head teacher turnover (% head teachers who are no longer head teachers in the same school by the next survey round)					
Head teacher left the school	88.89	39	74.24	52	-14.64
Head teacher was demoted in same school	11.11	39	25.76	52	14.64
Reasons for leaving the school (% head teachers who left the school by the next survey round)					

Transferred	48.88	32	68.57	37	19.69**
Retired	16.74	32	19.32	37	2.59
Passed away	11.61	32	6.27	37	-5.33
Studies	18.61	32	4.44	37	-14.16*
On secondment	0	32	1.39	37	1.39***
Disciplinary issue	1.15	32	0	37	-1.15***
Other	3.02	32	0	37	-3.02***

Sources: Impact evaluation baseline, midline and endline surveys (head teacher interview).
Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1.

There are two main reasons for the high turnover: head teachers leaving the school or remaining in the school but being demoted to assistant head teachers, academic masters, or teachers (Table 42). Between midline and endline, 74% of head teachers who left their posts had left the school while 26% were demoted within their school. Similarly, between baseline and midline the primary reason for head teacher turnover was head teachers leaving the school (89%), followed by head teachers being demoted within their school (11%). Note that the differences across time are not statistically significant, and this is likely due to the fact that the sample sizes of these groups is small.

Among the head teachers who have left the school between midline and endline, by far the most common reason was being transferred (69%), followed by retiring (20%), passing away (6%), going away for studies (4%), and being seconded (1%) (Table 42). There are some notable changes since midline: significantly more head teachers transferred to another school between midline and endline than between baseline and midline, while significantly more head teachers went to seek further studies between baseline and midline than between midline and endline.

At endline, 45% of head teachers had been head teachers at the school for less than two years (Table 43). Among the endline head teachers who have been head teachers at their current school for less than two years, 67% were teachers in their previous post and 31% were head teachers. The majority (91%) came from another school in the same district, while 8% were promoted from the same school. There have been no significant changes in these shares since midline.

Table 43: Head teacher previous job and location (trends in programme areas)

	Midline		Endline		Difference
	Estimate	N	Estimate	N	ML-EL
Head teacher has been head teacher at current school for less than two years (% head teachers)	36.21	99	44.69	100	8.48
Job before becoming head teacher at this school (% head teachers who had been head teachers for less than two years)					
Head teacher	27.36	34	30.87	46	3.51
Teacher	72.64	34	66.75	46	-5.89
Other job in education	0	34	2.38	46	2.38***
Location of previous job (% head teachers who had been head teachers at current school for less than two years)					
This school	28.21	34	8.23	46	-19.98
Another school in same district	69.79	34	90.89	46	21.11
Another school in same region but different district	0	34	0.88	46	.88***
Another school in another region	2.01	34	0	46	-2.01***

Sources: Impact evaluation midline and endline surveys (head teacher interview).
Notes: (1) Asterisks indicate statistical significance levels: *** p<0.01, ** p<0.05, * p<0.1.

5.4.4 WEO turnover

High turnover in education posts are not only symptomatic of posts at the school level but also of those at the ward level. When asked, 56% of head teachers in programme schools reported that the WEO has changed since the start of 2016 (Table 44).

Table 44: Turnover in WEO post

	Endline	
	Estimate	N
WEO has changed since start of 2016 (% schools)	56.19	99
Sources: Impact evaluation endline survey (head teacher interview).		
Notes: (1) This is for all schools.		

References

- Abadzi, H. (2006) 'Efficient learning for the poor: Insights from the frontier of cognitive neuroscience'. World Bank, Directions in Development Series.
- Bennell, P. and Akyeampong, K. (2007) 'Teacher Motivation in Sub-Saharan Africa and South Asia'. *Researching the Issues*, 71, London, UK DFID.
- Blundell, R. and Costa Dias, M. (2000) 'Evaluation Methods for Non-Experimental Data'. *Fiscal Studies* 21, no. 4: 427–68. www.jstor.org/stable/24437670.
- Boyden, J. and Ennew, J. (eds.)/Save the Children (1997) *Children in Focus – a Manual for Participatory Research with Children*. Stockholm: Save the Children Sweden
- Caliendo, M. and Kopeinig, S. (2005) 'Some practical guidance for the implementation of propensity score matching'. *IZA Discussion Papers*, No. 1588.
- Cambridge Education (2014) 'Final EQUIP-Tanzania Inception Report', 5 February.
- Cartwright, N. and Hardie, J. (2012) *Evidence Based Policy: A Practical Guide To Doing It Better*. Oxford: Oxford University Press.
- Carvalho, S. and White, H. (1997) 'Combining The Quantitative and Qualitative Approaches to Poverty Measurement and Analysis', Technical Paper 366. Washington DC: World Bank.
- Cueto, S. *et al.* (2009) 'Psychometric characteristics of cognitive development and achievements instruments in Round 2 of Young Lives'. Young Lives Technical Note #15, January 2009.
- De Grauwe, A. (2001) *School Supervision in Four African Countries: Vol. I: Challenges and Reforms*. Paris: UNESCO IIEP.
- De Ree, J. (2016) 'How Much Teachers Know and How Much it Matters in Class: Analyzing Three Rounds of Subject-Specific Test Score Data of Indonesian Students and Teachers'. World Bank Policy Research Working Paper
- Diseko, E., Barkhuizen, N. and Schutte, N. (2015) 'The Relationship between Talent Management and Turnover Intentions of Teachers in Botswana'. 20th International Academic Conference
- DFID (2013) 'Intervention Summary, Education Quality Improvement Programme in Tanzania (EQUIP-T)'. London, UK DFID.
- DFID (2011) *DFID Ethics Principles for Research and Evaluation*.
- Dowd, A. *et al.* (2013) 'Literacy Boost Cross Country Analysis Results', Save the Children, Washington D.C.
- Drake, L., Woolnough, A., Burbano, C. and Bundy, D. (2016) *Global school feeding sourcebook: lessons from 14 countries*. London, UK: Imperial College Press.
- EQUIP-T MA (2015) 'EQUIP-Tanzania Annual Report 2015'. United Republic of Tanzania (URT), PMO-RALG, MOEVT, and DFID.
- EQUIP-T MA (2016) 'EQUIP Tanzania, 2016 Annual Report External Version'. URT, DFID.
- EQUIP-T MA (2017) 'EQUIP Tanzania, 2016-17 Annual Summary External Version'. URT, DFID.
- Garbarino, S., and Holland, J. (2009) 'Quantitative and Qualitative Methods in Impact Evaluation and Measuring Results', Issues paper commissioned by DFID. GSDRC Emerging Issues Service.
- Gershoff, E.T. (2017) 'School corporal punishment in global perspective: prevalence, outcomes, and efforts at intervention'. *Psychology, Health & Medicine* 22, 224–239. <https://doi.org/10.1080/13548506.2016.1271955>

- Glewwe, P and Muralidharan, K. (2015) 'Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps and Policy Implications'. RISE-WP-15/001. RISE Programme.
- Glewwe, P., Kremer, M. and Moulin, S. (2009). 'Many children left behind? Textbooks and test scores in Kenya'. *American Economic Journal: Applied Economics* 1 (1) (January 2009): 112-35
- Gove, A., Brunnet, T., Bulat. J., Carrol. B., Henny C., Macon. W., Nderu E., and Sitabkan Y. (2017) 'Assessing the impact of early learning programs in Africa.' In Kenneth R Pugh, Peggy McCardle & Annie Stutzman (Eds) *Global Approaches to Early Learning Research and Practice*. New Directions for Child and Adolescent Development. 158, 25-41.
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). 'Toward a conceptual framework for mixed-method evaluation designs'. *Educational evaluation and policy analysis*, 11(3), 255-274.
- Guest, G., MacQueen, K. and Namey, E. (2012) *Applied Thematic Analysis*. Sage Publications.
- Hallinger, P. and Heck, R. (1996) 'The Principal's Role in School Effectiveness: An Assessment of Methodological Progress, 1980-1995'. In K. Leithwood, *The International Handbook of Research on Educational Leadership and Administration*. New York: Kluwer Press.
- Hardman, F. and Dachi, H. (2012) 'Evaluation of School-Based INSET Pilot Programme'. York, UK: Institute for Effective Education, University of York.
- Hattie (2009) 'Visible learning: A synthesis of over 800 meta-analyses related to achievement' London; New York, Routledge.
- Hoyland, A. Dye, L and Mawton, C (2009) 'A systematic review of the effect of breakfast on the cognitive performance of children and adolescents'. *Nutrition Research Reviews* 22.2: 220-243.
- Imbens, G. and Rubin, D. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press.
- J-PAL (2014) 'Increasing test score performance: What interventions are most effective at increasing student learning.'
- Krippendorff, K. (2004) 'Content Analysis: An Introduction to Its Methodology.' Sage Publications.
- Lechner, M. (2002) 'Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies'. *Review of Economics and Statistics*, 84(2), 205-220.
- Mays, N. and Pope, C. (1995) 'Rigour and Qualitative Research', *British Medical Journal* 311(6997): 109-112.
- Metzler, J. and Woessmann, L. (2012) 'The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation'. *Journal of Development Economics*, 99 (2)
- MOEST (2018) *Education Sector Development Plan 2016-17 to 2020-21 Tanzania Mainland*. URT
- MOEST (2017) *National Strategy on Inclusive Education*. URT, MOEST.
- MOEST and PO-RALG (2018a) *Tanzania Education Program for Results (EPforR) Programme Operation Manual*. URT, GOT.
- MOEST and PO-RALG (2018b) *Tanzania Education Program for Results (EPforR) Annual Report 2017-18 Draft (09 October 2018)*. URT, GOT.
- MOEVT (2016) *Curriculum for Basic Education Standards I and II*. URT, MOEVT.
- MOEVT (2009a) *Basic standards for pre- and primary education in Tanzania*. Dar es Salaam: MOEVT. www.ed-dpg.or.tz/pdf/PE/Basic%20Standards%20for%20Pre%20and%20Primary%20Education%20in%20TZ_2009.pdf
- MOEVT (2005a) *Mathematics Syllabus for Primary Schools Standard I-VII*. URT, MOEVT.
- MOEVT (2005b) *Kiswahili Syllabus for Primary Schools Standard I-VII*. URT, MOEVT.

- Mulkeen, A. (2010) 'Teachers in Anglophone Africa. Issues in Teacher Supply, Training and Management'. Washington, DC: The World Bank.
- Nag, S., Chiat, S., Torgerson, C. and Snowling, M.J. (2014) 'Literacy, Foundation Learning and Assessment in Developing Countries: Final Report'. Education Rigorous Literature Review. DFID.
- Ogando Portela, M.J. and Pells, K. (2015) 'Corporal Punishment in Schools – Longitudinal Evidence from Ethiopia, India, Peru and Vietnam'. Florence: Innocenti Discussion Papers.
- OPM (2015a) EQUIP-Tanzania Impact Evaluation Final Baseline Technical Report, Volume I: Results and Discussion. OPM.
- OPM (2015b) EQUIP-Tanzania Impact Evaluation Final Baseline Technical Report, Volume II: Methods and Technical Annexes. OPM.
- OPM (2016a) 'EQUIP-Tanzania Impact Evaluation: Midline Planning Report'. OPM.
- OPM (2016b) 'EQUIP-Tanzania Impact Evaluation: Preliminary Indicators for the Programme Treatment Districts: Results from the Baseline and Midline Quantitative Surveys'. OPM.
- OPM (2016c) 'EQUIP-Tanzania Impact Evaluation: Propensity Score Analysis Review: A Technical Note'. OPM.
- OPM (2016d) 'EQUIP-Tanzania Impact Evaluation: Midline Quantitative Fieldwork Report'. OPM.
- OPM (2017a) 'EQUIP-Tanzania Impact Evaluation Final Midline Technical Report, Volume I: Results and Discussion'. OPM.
- OPM (2017b) 'EQUIP-Tanzania Impact Evaluation Final Midline Technical Report, Volume II: Methods and Supplementary Evidence'. OPM.
- OPM (2018) 'EQUIP-Tanzania Impact Evaluation: Endline Planning Report: Part I Quantitative Research'. OPM.
- Osim, R.O., Chika, C. and Uchendu, I. O. (2012) 'Class size pressure: An Impediment to teacher's work'. *Global Advanced Research Journal of Educational Research and Review* Vol 1(5)
- Ouane, A. and Glanz, C (2011) 'Executive Summary'. In Ouane, A and Glanz, C (eds.) *Optimizing Learning, Education and Publishing in Africa: the Language Factor. A Review and Analysis of Theory and Practice in Mother-Tongue and Bilingual Education in sub-Saharan Africa*. UNESCO Institute for Lifelong Learning and the Association for the Development of Education in Africa / African Development Bank
- Paternoster, R., Brame, R., Mazerolle, P. and Piquero, A. (1998) 'Using the correct statistical test for the equality of regression coefficients'. *Criminology*, 36: 859–866. doi:10.1111/j.1745-9125.1998.tb01268.
- Pivik, R.T. et al. (2012) 'Eating breakfast enhances the efficiency of neural networks engaged during mental arithmetic in school-aged children' *Physiology and Behaviour* 106.4 2012: 548-555.
- Pritchett, L. and Beatty, A. (2012) 'The negative consequences of overambitious curricula in developing countries' *Faculty Research Working Paper Series RWP 12-035*. Harvard Kennedy School.
- Rosenbaum, P.R. and Rubin, D.B. (1985) 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score'. *The American Statistician* 1985: 39:33–38.
- RTI (2016) 'Assistance to Basic Education All Children Reading (ABE ACR): Preliminary Findings Report, Tanzania National Early Grade Reading Assessment (EGRA)'. North Carolina. USAID.

- RTI (2014) 'National Baseline Assessment for the 3Rs (Reading, Writing, and Arithmetic). Using EGRA, EGMA, and SSME in Tanzania'. Study Report. Draft. Washington, DC: USAID.
- Robinson, Lloyd and Rowe (2008) 'The Impact of Leadership in Student Outcomes: An Analysis of the Differential Effects of Leadership Types', *Education Administration Quarterly*.
- Rubin, D. (2001) 'Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation', *Health Services & Outcomes Research Methodology* 2, 169–188.
- Sabarwal, S., Evans, D. and Marshak, A. (2014) 'The permanent input hypothesis: the case of textbooks and (no) student learning in Sierra Leone'. Policy Research working paper, no. WPS 7021. Washington, DC: World Bank Group
- Schuh Moore, A.-M., DeStefano, J., and Adelman, E. (2012) 'Opportunity to Learn: A high impact strategy for improving educational outcomes in developing countries'. Washington, DC: USAID. Available at: <https://www.fhi360.org/resource/opportunity-learn-high-impact-strategy-improving-educational-outcomes-developing-countries>
- Singh, A. (2015a) <https://blogs.worldbank.org/impac evaluations/how-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations>
- Singh, A. (2015b) Private school effects in urban and rural India: Panel estimates at primary and secondary school ages. *Journal of Development Economics*, Volume 113, March 2015 pp16-32.
- Siraj, I., Taggart, B., Melhuish, E., Sammons, P. and Sylva, K. (2014) *Exploring Effective Pedagogy in Primary Schools: Evidence from Research*. London: Pearson. Available at: https://research.pearson.com/content/plc/prkc/uk/open-ideas/en/articles/explore-eppse/icr_content/par/articledownloadcompo/file.res/Exploring%20Effective%20Pedagogy%20in%20Primary%20Schools.pdf
- The National Commission for the Protection of Human Subjects of Biomedical and Behavioural Research (1979) 'The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Department of Health, Education, and Welfare'. [www.hhs.gov/ohrp/humansubjects/guidance/belmont.html]
- The Open University (no date) 'Ethics Principles for Research Involving Human Participants', [www.open.ac.uk/research/main/sites/www.open.ac.uk.research.main/files/files/OU%20Ethics%20Principles%20for%20Research%20Involving%20Human%20Participant.pdf]
- UN Committee on the Rights of the Child (CRC) (2013) Consideration of reports submitted by States parties under article 44 of the Convention, Third to fifth periodic reports of States parties due in 2012 : United Republic of Tanzania, 4 November 2013, CRC/C/TZA/3-5. Available at: <http://www.refworld.org/docid/54ae89254.html> [accessed 26 October 2018]
- UNESCO (2016) 'If you don't understand how can you learn?' GEM Report Policy Paper 24. Paris: UNESCO.
- UNESCO (2015) 'Education For All Global Monitoring Report 2015: Education for All 2000–2015: Achievements and Challenges'. Paris: UNESCO.
- UNESCO (2014) 'Education For All Global Monitoring Report 2013/14: Teaching and Learning: Achieving Quality for All'. Paris: UNESCO
- UNESCO (2007) 'Education For All Global Monitoring Report 2008: Education for All: Will we make it?' Paris: UNESCO.
- UNESCO (2005) *Guidelines for Inclusion: Ensuring Access to Education for All*. UNESCO
- UNICEF (2009) *Child Friendly Schools Manual*. New York
- Uwezo (2011) 'Are our Children Learning? Annual Learning Assessment Report 2011'. Dar es Salaam: Uwezo Tanzania.
- US Department of Health and Human Services (2009) 'Code of Federal Regulations TITLE 45 Public Welfare. PART 46. Protection of human subjects'. [www.hhs.gov/ohrp/policy/ohrpre regulations.pdf]

- Vogel, I. (2012) 'Review of the use of theory of change in international development'. Review paper for the Department of International Development. DFID.
- Westbrook, J., Durrani, N., Brown, R., Orr, D., Pryor, J., Boddy, J. and Salvi, F. (2013) 'Pedagogy, Curriculum, Teaching Practices and Teacher Education in Developing Countries'. Final Report. Education Rigorous Literature Review. UK DFID. Available at: www.gov.uk/government/uploads/system/uploads/attachment_data/file/305154/Pedagogy-curriculum-teaching-practices-education.pdf
- White, H. (2009) 'Theory-Based Impact Evaluation: Principles and Practice'. 3ie Working Paper 3. International Initiative on Impact Evaluation.
- Wright, B and Stone, M. (1979) *Best Test Design*. Chicago: Mesa Press.

Annex A Impact evaluation districts

The list of regions and districts that are included as treatment and control areas in the impact evaluation is shown in Table 45. This table also lists those regions and districts were part of the EQUIP-T programme (treatment) prior to the programme extension in 2017 but were not included in the impact evaluation. Lindi and Mara regions were excluded from the impact evaluation because EQUIP-T implementation started later than in the original five regions (Dodoma, Kigoma, Shinyanga, Simiyu and Tabora). The districts in the original five regions that were excluded from the impact evaluation were omitted because of the risk of contamination from other programmes (see Section 3.2).

Table 45 Impact evaluation treatment and control districts

Control/treatment	Region	District	
Control regions and districts in impact evaluation study	Arusha	Ngorongoro DC	
	Mwanza	Misungwi DC	
	Pwani	Rufiji DC	
	Rukwa	Nkasi DC	
	Ruvuma	Tunduru DC	
	Singida		Ikungi DC
			Singida DC
Tanga	Kilindi DC		
Treatment regions and districts in impact evaluation study (Note: all 17 districts are part of the quantitative survey, * indicates they are also part of the qualitative research at midline)	Dodoma	Bahi DC	
		Chamwino DC	
		Kongwa DC	
		Mpwapwa DC *	
	Kigoma		Kakonko DC
			Kibondo DC
	Shinyanga		Kishapu DC *
			Shinyanga DC
	Simiyu		Bariadi DC
			Bariadi TC
			Itilima DC
			Maswa DC
			Meatu DC
	Tabora		Igunga DC
Nzega DC			
Sikonge DC			
Uyui DC *			
Treatment districts that are not part of the impact evaluation study before the programme extension in 2017 (Note: districts in Lindi and Mara joined EQUIP-T in 2015)	Dodoma	Chemba DC	
		Kondoa DC	
	Kigoma		Buhigwe DC
			Kasulu DC
			Kigoma DC
			Uvinza DC
Shinyanga	Kahama DC		

		Msalala DC
		Ushetu DC
	Simiyu	Busega DC
	Tabora	Kaliua DC
		Uramba DC
	Lindi	Kilwa DC
		Lindi DC
		Liwale DC
		Ruangwa DC
	Mara	Bunda DC
		Butiama DC
		Musuma DC
		Musuma MC
		Rorya DC

Source: OPM

Annex B Stakeholder engagement and impact evaluation governance

B.1 Stakeholder engagement

Stakeholder engagement is an ongoing process in the impact evaluation, and started from the inception phase with consultations on the overall design of the impact evaluation. Volume II of the baseline impact evaluation report sets out the stakeholder consultations carried out in the inception phase and in disseminating the baseline findings (OPM 2015b). Similarly, Volume II of the midline impact evaluation report (OPM 2016b) sets out the stakeholder engagement activities which took place between October 2015 (prior to the midline research) and December 2016 when the main dissemination of midline findings took place. Plans for stakeholder engagement and dissemination of the quantitative endline findings were set out and agreed in the Quantitative Endline Planning Report (OPM 2018). Table 46 summarises the main stakeholder engagement activities that have taken place since the dissemination of midline evaluation findings, and includes those planned for sharing the quantitative endline findings.

For the quantitative endline evaluation, which is the first component of the overall endline evaluation, stakeholder engagement began in October 2017 with an application to Tanzania's Commission for Science and Technology (COSTECH) to obtain approval for the endline research. This was followed by a visit to Dar es Salaam in January 2018 by the lead researchers from the evaluation team to hold a workshop with staff from the EQUIP-T MA. The purpose of the workshop was to obtain an update on programme implementation, and to further develop the details of the programme theory of change to take account of the programme's extension to January 2020. The evaluation team also met with the education advisors at DFID to discuss the emerging priorities for the endline evaluation design. A second visit by the evaluation team followed in February 2018, in order to consult with government education officials from PO-RALG, MOEST and ADEM (a sub-set of Reference Group members) on the endline evaluation design and methods. The evaluation team also carried out a pre-test of the draft endline survey instruments in a set of schools in the Dodoma region. The final quantitative endline research priorities and design were documented in the Endline Quantitative Planning Report (OPM 2018), and this was submitted to DFID in early March 2018 and circulated to the EQUIP-T MA and the evaluation's Reference Group for feedback. This marked the end of the preparatory phase.

The enumerator training took place in late March 2018 followed by piloting and fieldwork in April and May 2018. The data was checked and cleaned during June and July, and the analysis phase has taken place between August and October. Following the submission of this draft report, there will be a phone call between DFID, the EQUIP-T MA and the evaluation team to discuss initial feedback. After this, the evaluation team will convene a full day Reference Group meeting in Dodoma to present the results and to receive feedback from this wider group of stakeholders. The report will be finalised following this feedback. The final report will be presented at the EQUIP-T steering committee meeting, which includes a wide audience of education sector stakeholders.

The principal audience for this endline quantitative evaluation are EQUIP-T's MA, DFID, and GoT officials. The results are intended to inform further adjustments to the programme before it finishes in January 2020, as well as to promote accountability and lesson learning for DFID and the GoT. The findings will also help to guide the design of the second part of the endline, which will include qualitative research and a cost study. Consultation on the priorities and design of these components of the endline evaluation will start in early 2019.

Table 46: Stakeholder consultations and events—from dissemination of midline findings to plans for dissemination of quantitative endline findings

Date	Purpose
Dissemination of midline evaluation findings	
November 2016	Shared midline evaluation findings at the Joint Education Sector Review.
December 2016	<p>Presentation of midline evaluation findings at the EQUIP-T Annual Review steering committee meeting.</p> <p>Presentation of midline evaluation findings at the impact evaluation Reference Group meeting.</p> <p>Presentation of midline evaluation findings on INSET at GoT/Development partners meeting on harmonising approaches to teacher INSET.</p>
March 2017	Paper on the midline impact estimates presented at African Evaluation Association Conference (8 th AfrEA International Conference, Uganda).
September 2017	Two presentations at the UKFIET Oxford Conference on Education and Development on: (i) Factors affecting the sustainability of a teacher in-service training programme in Tanzania; and (ii) Challenges of assessing cost effectiveness and financial sustainability of teacher INSET—insights from Tanzania.
Quantitative endline evaluation engagement	
October 2017	Application made to by COSTECH for the endline impact evaluation (subsequently granted).
January 2018	<p>Met with DFID Education Advisors and all Component Leads at the EQUIP-T MA in Dar es Salaam to understand implementation progress, plans for the coming months, and expectations of changes so far. This was followed up via emails and phone calls to clarify details.</p> <p>Held 2 day workshop with Component Leads at the EQUIP-T MA in Dar es Salaam to further develop the details of the programme theory of change, and to take into account the extension of the programme and the introduction of new components.</p> <p>Emails and phone calls with the Coordinators of LANES and Tusome Pamoja to inform them of the forthcoming endline evaluation of EQUIP-T, and to obtain information about the implementation of these programmes (activities and geographical location).</p> <p>Started process of updating impact evaluation Reference Group membership via emails and phone calls.</p>
February 2018	<p>Meeting in Dodoma with government officials from PO-RALG, MOEST and ADEM (a sub-set of the Reference Group) to discuss the endline evaluation research focus and methods.</p> <p>Pre-test of draft quantitative endline survey instruments by the impact evaluation team, in Dodoma region.</p> <p>Meeting with government coordinator of LANES with responsibility for engagement on EQUIP-T and Tusome Pamoja, to inform and consult on the endline EQUIP-T evaluation and to obtain further details on LANES implementation and other relevant government initiatives.</p> <p>Continued process of updating impact evaluation Reference Group membership via emails and phone calls.</p>
March 2018	Endline Quantitative Evaluation Planning Report submitted to DFID, EQUIP-T MA and the broader Reference Group for comment and feedback (report approved).
April/May 2018	<p>Supervisor and enumerator training and pilot for quantitative endline survey by the impact evaluation team.</p> <p>Endline quantitative data collection.</p>
November 2018	Submission of draft endline quantitative evaluation report to DFID and EQUIP-T MA
November/December 2018 (planned)	<p>Phone call meeting between evaluation team, DFID and EQUIP-T MA to discuss feedback on the draft report (week of 19 November).</p> <p>Presentation and discussion of quantitative endline findings at Reference Group meeting in Dodoma (week of 03 December).</p> <p>Presentation of quantitative endline findings at EQUIP-T Steering Committee meeting (date TBC).</p>
December 2018/January 2019 (planned)	Consultation on the design of the qualitative endline evaluation and the cost study, to be carried out in 2019 (date TBC)

All of the reports, briefing notes, issues papers and other products produced as part of the impact evaluation are available on OPM's website <https://www.opml.co.uk/projects/assessing-equip-t>

A technical working paper on the innovative approach to impact estimation used in this study is also on OPM's website: <https://www.opml.co.uk/publications/working-paper-matching-differencing-repeat>

Briefing notes and conference papers produced using the impact evaluation findings have been uploaded on to the Social Science Research Network www.ssrn.com (see for example: <https://ssrn.com/abstract=2779240>; <https://ssrn.com/abstract=2579284>; and <https://ssrn.com/abstract=2782747>).

The baseline and midline quantitative survey data (anonymised) is publically available on the World Bank microdata library <http://microdata.worldbank.org/index.php/catalog/2290>.

B.2 Reference Group

At the start of the impact evaluation in 2014, the Ministry of Education led a process to form an EQUIP-T impact evaluation Reference Group to provide technical recommendations and feedback to the OPM evaluation team. The terms of reference for the Reference Group are included in the Midline Planning Report (OPM 2015a, Annex F). At baseline, the Reference Group held its first meeting to review and comment on the overall impact evaluation design (January 2014). A second Reference Group meeting was held in November 2014 where baseline findings were discussed extensively, feedback provided to guide revisions to the report, and members advised the evaluation team on opportunities for dissemination as well as links with other studies and programmes.

The Reference Group membership was refreshed to adjust for members no longer able to represent their organisations, and expanded to include additional education agencies, during the preparatory phase of the midline evaluation in consultation with the Commissioner for Education. A third Reference Group meeting took place in December 2016 to discuss the draft midline evaluation report and plan for dissemination of the findings. During the meeting, members provided useful feedback on the draft report (noted in the meeting minutes, subsequently circulated to all members for corrections or additions), and members were also requested to provide any additional feedback in writing. The evaluation team consolidated all the feedback received on the draft report from DFID, the EQUIP-T MA, and other RG members into a document. From this, the team carefully considered each comment and made changes to the draft report as appropriate. The team also drafted a written response to each comment, explaining how the comment had been dealt with in the final report or justifying why no changes had been made. This commentary was submitted to DFID together with the final report.

The Reference Group membership was again refreshed during the preparatory phase of the endline evaluation in January and February 2018. From its inception, the Commissioner for Education has chaired the Reference Group and it is convened by Professor Herme Moshia from UDSM who is a core member of the evaluation team. The organisations represented on the Reference Group are:

- Government ministries: MOEST and PO-RALG;
- Government education departments and agencies: National Examinations Council of Tanzania (NECTA); Tanzania Institute of Education (TIE); and Agency for the Development of Education Management (ADEM).
- DFID;
- EQUIP-T MA
- USDAM, School of Education; and
- Education research organisation (Twaweza East Africa).

Table 47 EQUIP-T Impact Evaluation Reference Group members (December 2018)

Member	Position	Organisation
Edicome Shrima	Acting Commissioner for Education	MOEST
Tixon Nzunda	Deputy Permanent Secretary Education	PO-RALG
Hilda Mkandawile	Officer (LANES/EQUIP-T Coordinator)	MOEST
Mr Makuru	Assistant Acting Director M & E	MOEST
Joel Masangula	Acting Director Primary education	MOEST
George Gidamva	Assistant Director of Primary Education	PO-RALG
Benjamin Oganga	Assistant Director of Secondary Education	PO-RALG
Odilia moshi	Assistant Director Special Education Needs	PO-RALG
Julius Nestory	Director of Education Administration	PO-RALG
Neema Chamgeni	Officer (Economist/EQUIP-T Coordinator)	PO-RALG
Dr Siston Masanja Mgullah	Chief Executive Officer	ADEM
Fika K. Mwakabungu	Officer	TIE
Ezekiel Kisove	Officer	NECTA
Prof. Kitila Mkumbo	Permanent Secretary	Ministry of Water (former USDM)
Dr Blackson Kanukisya	Professor of Education	UDSM (School of education)
Prof. Kalafunja Osaki	Professor of Education	SAUT
John Lusingu	Education Advisor	DFID
Arianna Zanolini	Education Advisor	DFID
George Senyoni	Leader Monitoring and Evaluation	EQUIP-T
Laura McInerney	Deputy National Coordinator	EQUIP-T
Dr Godfrey Telli	Officer	Twaweza East Africa
Mohamed Msongo	DEO-BAHI	PO-RALG
Lyimo Peter Maria	REO-Dodoma	PO-RALG

B.3 Impact evaluation governance and quality assurance

Oversight and policy direction for the impact evaluation is provided by an OPM Governance Team comprising the OPM Managing Director, the OPM Director of Statistics, Evidence and Accountability, the OPM education portfolio lead, and an OPM Education Associate who is Senior Education Advisor in the impact evaluation core senior team (see Table 48 below).

Management is executed by the Project Manager, an OPM Principal Education Consultant, who in addition to playing a leading technical role is responsible for team management, the coordination of inputs, financial management and liaison with the supporting administration team and research teams in OPM's Oxford office and OPM's Tanzania Office respectively, and OPM's internal reporting and project oversight processes.

The Project Manager is responsible to the OPM Governance Team for successful delivery of the impact evaluation. The Project Manager is supported by a core senior team and a wider team of technical specialists (see Table 48 below). The core senior team comprises, a deputy project manager, a senior education advisor (also part of the OPM governance team), and a senior national education advisor. There are 11 technical specialists in the wider technical team. The project manager ensures that the two teams work together to meet the objectives of the evaluation, and to produce the key deliverables. The core team is responsible for stakeholder engagement including dissemination of findings and engagement with the Reference Group.

Table 48: Endline quantitative impact evaluation team members and roles

Name	Role
Georgina Rawle	Project Manager/Endline Quantitative Design Lead/Quantitative Analyst
Nicola Ruddle	Deputy Project Manager
Paud Murphy	Senior Education Advisor
Professor Herme Mosha	Senior National Education Advisor
Dr. Gunilla Pettersson Gelandner	Senior Education Specialist/ Quantitative Analyst
Dr. Michele Binci	Impact Estimation Lead
Paul Jasper	Senior Impact Estimation Analyst
Safa Khan	Impact Estimation Analyst
Jana Harb	Quantitative Survey Fieldwork Technical Lead/ Quantitative Analyst
Ignatus Jacob	Quantitative Survey Fieldwork Lead
Deogardius Medardi	Quantitative Survey Fieldwork Manager
Andreas Kutka	Quantitative Survey Fieldwork Adviser
Diego Shirima	Quantitative Survey Data Manager
Alessio Romarri	Research assistant
Michelle Rorich	Research assistant

Quality Assurance for the endline quantitative research has been provided using a number of layers of review. In the first stage, each key activity and output has been reviewed internally by other project team members, led by Georgina Rawle, Gunilla Pettersson Gelandner and Jana Harb for quantitative descriptive analysis, Michele Binci and Paul Jasper for impact analysis. The learning outcomes analysis including Rasch modelling and construction of interval scales was reviewed by Dr Joshua McGrane (Psychometrician and Rasch measurement specialist, University of Oxford).

In the second stage, the full drafts of Volume I and Volume II were shared with three reviewers: Paud Murphy, Senior Education Advisor, Professor Herme Mosha (University of Dar es Salaam), Senior National Education Advisor, and Dr Caine Rolleston a leading academic researcher in the field of education and economics (Institute of Education, University College London). This team also reviewed the baseline and midline draft reports.

A final stage of external quality assurance will be provided through the Reference Group meeting, together with review and feedback from DFID.

The full endline evaluation, including this quantitative component, the qualitative research and costing study due to take place in 2019, will be reviewed by EQUALS (DFID's external review body for evaluations).

Annex C Ethical considerations

C.1 Ethical protocols at endline

The endline evaluation survey followed the ethical protocols that were used at midline. There were some minor changes to the quantitative instruments for endline, and one change in the fieldwork protocol to deal with the facilitation of a small-group teacher interview.

OPM gained approval for the endline research from the Tanzania Commission for Science and Technology (COSTECH) which has the mandate of coordinating and promoting research and technology development activities in the country.

The OPM Ethical Review Committee has also reviewed the proposal for endline fieldwork, including the amended instruments and the informed consent statements, and granted approval.

The full fieldwork protocols and consent statements were in the Endline Planning Report, Part I (OPM, 2018, Annex D).

The short sections below set out the ethical principles that the protocols used in the endline survey adhere to. These have been reproduced from the Endline Planning Report, Part I (OPM 2018) and were fully implemented.

C.2 Principles

As this research involves human subjects, it is important to be fully aware of the ethical considerations. A review of best practice was conducted to inform the design and protocols of the midline fieldwork and data use. These have been reproduced for use in the endline survey. This review looked at the protocols OPM used in the baseline, those used in OPM's other education evaluations, those used by other research organisation in Tanzania, and guidance from organisations specialising in children's rights (Save the Children, 2007), research (Open University, US Department of Health and Human Services) and development (DFID, 2011).

There are three basic ethical principles of research with human subjects, as set out in the Belmont Report (1979):

1. Respect for persons
2. Beneficence
3. Justice

C.2.1 Respect for persons

This means the prospective participants should be given the information they need to decide whether or not they want to participate, they should be given the freedom to decide not to participate or to stop at any point. In particular, this means that participants should give informed consent, agreeing to take part voluntarily and with adequate information. Where a participant has diminished autonomy – in this case children – they are entitled to additional protections.

In the quantitative surveys, all participants will be read a statement before the interview/ group discussion begins. The statement will set out what the research is for, how and why they were selected, the confidentiality of their responses, how responses will be used (and in particular that they will not affect their grades or job), that the process is optional and they are welcome to ask questions or leave at any time. After this, the enumerator (or interviewer or facilitator) will ask them if they agree

to continue. At this point, the enumerator ticks a box (in CAPI) to confirm the participant has given oral informed consent to continue.

Where children are being interviewed, we will ask the head teacher to give consent on behalf of the parents, and also ask the children for their own consent, in simple language. The statement will be read slowly, and the enumerator will read it again if necessary. This consent statement and agreement will be done individually for the quantitative fieldwork, away from other teachers or parents, so that they do not feel pressured either way. If the researchers feel that any child is not comfortable during this process, they will tactfully find a way to take the child aside and discuss this with them personally.

We have chosen to seek oral consent based on our experience with research in developing countries, and in particular with respondents who are not literate and/or are not familiar with research. In these cases, respondents become very formal and often even worried if we ask them to sign a piece of paper. We want to make sure our respondents are as relaxed and responsive as possible, so allowing them to consent orally and recording it meticulously (by the research team) will achieve the same function without compromising the quality of our interactions.

C.2.2 Beneficence

This principle requires that no harm is caused by the research. There are a number of ways in which we will adhere to this principle. Participants will be interviewed in a quiet place where others cannot hear their responses. Responses will be confidential – we will not name respondents or tell anyone outside of the research team the specifics of who was interviewed or who gave specific responses. This means that no responses will be attributable, and they will not be written in the report in a way which is traceable. These principles are intended to avoid any social risk from views being overheard by others in the community or those above them in the reporting line, and should allow respondents to speak more honestly. The quantitative data set will be made publically available but anonymised. The research team will be trained in confidentiality and sign agreements to keep the responses confidential.

Particular care will be taken in our engagement with children. The research involves interviewing children in standard 3, who generally are between the ages of 9 and 11 years. Given their age, it is important they are treated with care and respect, and given full opportunity to decide to opt out of the work. The fieldworkers carrying out the interviews will be 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 is fine for them to leave a question or leave the interview entirely, keeping responses confidential and anonymous – verbally but also by carefully handling the data collected. These processes will be set out in the enumerator manuals which will be used during training and be available for reference during the fieldwork. No responses will be coerced, participants will be free to not respond.

C.2.3 Justice

Justice requires that individuals and groups are treated fairly and equitably. In this case, there is no notable benefit (except refreshments in a group discussion) or burden (except time) of taking part in the research, and all participants will be subject to the same benefits and burden.

Annex D Quantitative survey fieldwork

OPM's Tanzania office conducted the impact evaluation endline survey. A detailed report on the fieldwork is available (OPM 2018b). This annex summarises the key points from the fieldwork report.

D.1 Personnel

The fieldwork management team comprised eight members (including six OPM staff) led by a quantitative survey project manager who had overall responsibility for the design, implementation, management and quality of the fieldwork. Since all the survey instruments were administered using computer assisted personal interviewing (CAPI), the team also included several members with very strong computer programming skills in the relevant software (Surveybe). 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, piloting and early fieldwork.

60 enumerators were invited to the training. These were selected based on the following criteria (in order): (i) good performance during the EQUIP-T baseline and midline surveys (24 enumerators from baseline and/or midline attended the endline training); (ii) interviewers with strong track record from other OPM-led surveys; (iii) new recruits—these were selected based on their prior survey experience and knowledge of education. The final fieldwork team is listed below. Supervisors are bolded.

Fieldwork team for the endline evaluation survey

1	Samwel David Mande	21	Robert Sizya	41	Mohammed Fadhili Simba
2	Bazil Mwemezi Kagande	22	Hyacinta Ngatoluwa	42	Benjamin Gerald
3	Michael Martin	23	Erick Kazoka	43	Maria Lumolwa
4	Laban N Batungi	24	Upendo Gervas	44	Baraka Jacob Mtui
5	Sylivester Michael	25	Sarah Ndimangwa	45	Novatus Fredy Luhizo
6	Vaileth Lemmy	26	Josephat Lucian Ritte	46	Happy Mushi
7	Henry Deodatus Arumasi	27	Lameck Jackson	47	Alice Daudi Shelukindo
8	James Bonga	28	Mary J. Ombeni	48	Rose Alexander Mchaki
9	Edeltruda Mtaki	29	Julieth Manyara	49	Juston Leason Bataza
10	Fatuma Said	30	Peter Nyanda	50	Kelvin Mtari
11	Eddna Chandeu	31	Yassin Khalid Kimolo	51	Rehemaely Makotha
12	Subira Doreen Mosha	32	Beda Kakuru Henry		
13	Alexander Katura	33	Praygod Goodluck		
14	Theresia Robson	34	Jeniva Mbalikaki		
15	Michael Joseph Davis	35	Regina Kafanabo		
16	Francis Gervas Swai	36	Christopher Ramadhani		
17	Mboya Mkundelida	37	Mansula Shemera		
18	Zeenath Abdulaziz	38	Peaceman Luinga		
19	Valentine Nemes	39	Kanisia Komba		
20	Masanja Edward Bunyongoli	40	AnnaMaria Mushongi		

D.2 Fieldwork preparation

The early fieldwork preparation consisted of pre-testing and refining the instruments and protocols, obtaining permits from the government for visiting schools during the pre-tests, pilot and fieldwork, and revising the midline fieldwork manual.

D.2.1 Pre-test

A full pre-test of all instruments and protocols took place from 19 to 23 February 2018 in Dodoma. A team of six (four members of the core evaluation team and two experienced survey supervisors who were supervisors during the midline fieldwork) visited eight schools, following one day of classroom based training. The main objectives of the pre-test were to test the functionality of the updated electronic questionnaires in the latest version of the CAPI software (Surveybe); test the changes that were made to the midline instruments, focusing mostly on the head teacher interview; and test the new endline instrument - that is the in-service training coordinator interview.

The pre-test resulted in the following outcomes:

- Refinement of the instruments and data collection protocols;
- Refinement of the translation of instruments from English to Kiswahili; and
- Significant changes made to the development of the instruments in CAPI (Surveybe).

D.2.2 Permits and reporting

As part of preliminary preparations for any survey in Tanzania, there are two types of governmental permits that have to be obtained prior to beginning research work:

- **COSTECH Permit** - Mandatory for any research activity in Tanzania.
- **Ministry Permit** - Different partners in the field require Ministry letters, as few recognise COSTECH. These permits give the order to local administration to cooperate with the research and support the field teams.

Upon receipt of the permits, the anticipated fieldwork needs to be reported at the regional and district level during which letters introducing the study to local leaders are obtained in the process.

For the endline survey, the COSTECH research clearance and an introduction letter were received more than three months prior to the start of actual fieldwork.

For the Ministry permits, OPM reported to The Prime Minister's Office Regional Administration and Local government (PMORALG) and to the Ministry of Education and Vocational Training (MOEVT). Reporting to MOEVT was relatively fast and simple. The initial submitted letters were followed up in person, and an introduction letter to all 12 Regional Administrative Secretaries (RAS) was received after seven days. Getting government approvals from PMORALG and the RASs was more challenging and time-consuming as it required physical reporting to PMORALG's office in Dodoma as well as physical reporting to all regions and districts that are covered by the endline fieldwork, pre-testing and piloting. However, having learned from midline how challenging this process is, the fieldwork management team devised a plan for endline that started the reporting process early on and involved two members of the fieldwork management team and two supervisors physically reporting in person to all 12 regional and 25 district offices during the month of February. This resulted in all permits and approval letters being obtained at least one month prior to piloting.

D.2.3 Fieldwork manual

Using the midline fieldwork manual as a basis, an extensive fieldworker manual was developed that covered basic guidelines on behaviour and attitude, the use of CAPI and data validation procedures, instructions on fieldwork plans and procedures (sample, targets, replacements, communication, and reporting) as well as a dedicated part on the description of all instruments and protocols. Insights from the pre-test were reflected in the manual.

Draft versions of the instrument and protocol sections of the manual were shared in softcopy with interviewers as a reference during the training, and used as guidelines by the trainers. 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 shared in softcopy with all fieldworkers at the end of the pilot phase.

D.3 Training and pilot

Enumerator training and a field pilot took place in Dar es Salaam and Dodoma from 26 March to 14 April 2018. A total of 60 trainees participated in the training. The training was delivered by four members of the fieldwork management team, the overall project manager of the impact evaluation, and another member of the core evaluation team.

The main objective of the training was to ensure that team members would be able to master the instruments, understand and correctly implement the fieldwork protocols, comfortably use CAPI, and be able to perform data validation. Supervisors were furthermore trained on their extra responsibilities of data management, fieldwork and financial management, logistical tasks, and the transmission of data files to the data manager.

The training had two components: a classroom-based training component and a field-based component that included a full scale pilot. The performance of enumerators was assessed on an ongoing basis, using written assessments and observation of performance in the field and these scores were recorded. At the end of the training and pilot phase, the final fieldwork team was selected using this information.

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.

D.4 Fieldwork organisation

D.4.1 Fieldwork plan

The fieldwork plan was designed to cover all 200 schools within all 12 regions and 25 districts for the duration of not more than six weeks. The plan had to cater for the short fieldwork time window dictated by the end of the school mid-term break and the start of exams at the end of the term; rainy season; allowing the fieldwork management team to supervise teams during the first week of implementation; minimising travel days between districts and during the weekdays; suitable allocation of teams to districts to address cultural and language barriers; and flexibility to deal with unforeseen circumstances.

D.4.2 Fieldwork model

The team composition and fieldwork model at endline were the same as those at midline with the exception of adding one more field team to deal with the shorter timeframe at endline and to ensure that the fieldwork is completed within five to six weeks. At endline there were four treatment teams composed of five enumerators and one supervisor, four control teams of four enumerators and one supervisor each, and one team of five enumerators and one supervisor that visited control and treatment areas. Each team visited and completed one school on one day.

D.5 Fieldwork implementation

The fieldwork started on 16 April and ended on 21 May 2018 with no breaks in-between, except for a couple of days of bank holidays and a few travel days for some of the teams. Teams communicated regularly with OPM to report delays and/or any event likely to affect the feasibility of the fieldwork plan.

D.5.1 Replacements

D.5.1.1 Schools

All schools that were interviewed at baseline and midline were revisited and interviewed at endline, and hence no replacement of schools took place. There were only two cases where teams visited a school and were unable to conduct the survey because they had to report to the district office due to security concerns. In those cases, the teams rearranged to come back another day to conduct the survey in those schools.

D.5.1.2 Pupils and teachers

Only 68 pupils (out of 2,999 pupils) were replaced. The reasons for replacement were: 29 were unavailable due to sudden events such as illness, 30 were absent (but had been recorded by the teacher as present and hence were part of the sampling frame), 5 could not speak or see or hear at all and thus they were not asked to sit for an hour long interview, and 4 for some other reason.

No replacement was done for the teacher interviews or lesson observations, as no sampling was required.

D.5.2 Response rates per instrument

Table 2 in Chapter 3 above shows the generally high response rates for each instrument. Here is some further information underlying the response rates for selected instruments:

- If the parent or guardian of the tested pupil or other adult household member could not be reached, as a last resort, the poverty scorecard was administered to the pupil. This happened in 230 out of 2,992 cases (8%). Some of the reasons given by enumerators were that the pupil is boarding and parents live far away, pupil lives too far away to be reached, and parents were not found at home after repeated visits.
- Some 97 of the 889 teacher interviews (11%) were conducted over the phone, as the teacher was absent on the day of the survey.
- In 40 out of the 200 schools, the head teacher was absent on the day of the survey and as a result the assistant head teacher or another teacher was interviewed instead to collect information related to school records. After fieldwork ended, head teachers in 38 of those 40 schools were reached over the phone to complete the missing modules of the head teacher interview.

D.6 Quality control and data checking protocols

At the end of each working day, supervisors collected all interview files from their team members and uploaded them into a shared and organised Dropbox folder that was set up by the data manager. The data manager would receive all files from all nine teams and export them into Stata data files (a statistical programme) and then run daily checks on all files to make sure they are complete and identify potential errors.

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:

D.6.1 Selection and supervision of enumerators

As discussed above, each enumerator was supervised at least once 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 into roles matching individual strengths and weaknesses.

D.6.2 CAPI built-in routing and validations

One important quality control means in CAPI surveys is the use of automatic routing and checking rules built into the CAPI questionnaires that flag simple errors during the interview, that is early enough for them be corrected during the interview. In each CAPI instrument, validations and checks were incorporated in the design in order to significantly reduce errors and inaccuracies during data collection. In addition to having automatic skip patterns built into the design to eliminate errors resulting from wrong skips, the CAPI validations also checked for missing fields, out of range values and simple inconsistencies within instruments.

D.6.3 Secondary consistency checks and cleaning in Stata

The endline survey exploited another 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 consistency and completeness, and make corrections if necessary. The checks comprised the following:

- ID uniqueness and matching across instruments;
- Completeness of observations: target sample size versus actual; and
- Intra and inter-instrument consistency and out of range checks.

The data manager ran the checking do-file on a daily basis on the latest cleaned data. This would return a list of potential issues in the long format which the data manager would then investigate and undertake the necessary cleaning actions. Whenever any issue was flagged, effort to obtain an explanation was undertaken either by reviewing enumerator comments or phoning teams.

On a daily basis, the data manager collated, shared and discussed all flagged errors with the supervisors in the field, who in turn discussed them with their team members. Throughout the fieldwork, occurrences of errors were monitored in order to keep an eye on the performance of data collectors and constantly provide them with feedback to improve.

In addition to the checking and cleaning process, all enumerator comments as well as other specify variables were translated from Kiswahili to English. All translated entries were further reviewed by the data analysis team in order to 1) ensure that they are understandable and properly translated into English and 2) none of the other specify answers for multiple response questions are in fact synonymous to one of the existing response items. The revision resulted in a long list of other specify items that were then recoded into one the available response items.

D.6.4 Monitoring fieldwork progress and performance indicators

In addition to the above checks that were specific to each instrument, the survey team built a dashboard that allowed for daily monitoring of the general progress of the fieldwork and specific indicators revealing the performance of teams and enumerators over time. For example, indicators included number of control/treatment schools completed, number of instruments completed within each school, average interviewing time of each instrument, time of the day when the pupil tests were conducted, number of pupils interviewed for the scorecard instead of their parents, number of teacher

interviews conducted over the phone, number of pupils being replaced, etc. These indicators were constructed in a Stata do-file that ran on the latest cleaned dataset and was then uploaded onto the dashboard (that was created using the visual software, Power BI) that would break down each of the indicators by team, enumerator (where applicable) and week of data collection. This was reviewed on a daily basis by the fieldwork management team and used to feedback to weaker teams and to improve performance.

D.6.5 Field visits by fieldwork management team including back-checking of data

The quality assurance protocol involved visits by the fieldwork management team to the field as well as data back-checks. Two members of the fieldwork management team visited a number of schools and households across 8 of the 12 regions over a two-week period. The purpose was to verify that the school and household interviews were conducted properly, to collect any missing information from these schools or clarify certain issues that were flagged as errors during the daily checking process, and to hold debriefs and retraining sessions with the teams in the field.

D.6.6 Integration of Analysis and Survey Team

Another central element of quality assurance was the strong integration of the fieldwork management team and members of the quantitative analysis team, including the overall project manager. Members of both teams were involved in the fieldwork preparation and implementation, and in the analysis process which followed.

D.7 Fieldwork challenges

The EQUIP-T endline fieldwork incorporated many of the lessons that were learned from the midline experience such as reporting physically to all regions and districts to obtain permit letters and doing so early on, recruiting more enumerators for the training, training all trainees on the pupil test, and having a training facilitator. Yet the endline fieldwork still faced some challenges, most important of which are:

- The fieldwork window between the end of the mid-term break and the start of the exams at the end of the term is narrow. A few alternative options were assessed in December 2017 but it was decided that it was best to stick with the same time window as at baseline and midline to ensure results are comparable. In order to address the narrow window, one extra team of six enumerators was added to the model that balanced between survey being completed on time and having a manageable number of trainees. Furthermore, towards the end of the fieldwork larger schools were prioritised for completion as these were the most likely to be busy towards the end with exam preparations.
- Unscheduled school events on the day of the survey. In order to address this, as at midline, teams contacted the DEOs and head teachers one week in advance (to explore and confirm school timetables and accessibility “in the next month”). The teams did not disclose precisely the date of the visit to avoid schools “preparing” for the visit and to not distort the data collected on teacher attendance and punctuality. There were a number of few cases where teams arrived at the school and found out that the schools were closed or holding some event, and in those cases the fieldwork management team re-planned the school visits and teams were sent to the nearest school in the area that had not yet been visited.
- Rains posed some logistical challenges as roads and bridges were disrupted. OPM was well prepared to minimize the impact. As was done at midline, the Ngorogoro team was dispatched early from the central field start location to be as far ahead of the rains as possible and reduce their exposure to transport issues arising in the district during the rainy season. All supervisors were instructed to check for potential access issues to schools the day prior to visiting. In some

instances the road to a school was not accessible by car and OPM had to transfer additional money to the teams to hire local transport such as motorcycles.

- In a few cases, teams arrived late at the school, due to either transport issues arising from the rain or to rescheduling due to some events at a school. This resulted in teams missing the headcount observation of teachers in the morning. A couple of these cases happened during the first week of fieldwork, and in those cases, a Kiswahili speaking member of the fieldwork management team was sent to those schools on another day to conduct the headcount observation of teachers. Additionally, during the visit to the field in mid-implementation, a member of the fieldwork management team visited one other school to conduct a missing headcount.
- In certain areas, particularly Ngorongoro and Simiyu, there are many non-Kiswahili speakers. This can make it difficult to interview parents. Some teams at endline used teachers who knew the vernacular language to translate for the scorecard interview with parents. It was emphasized with field teams that pupil tests are always administered in Kiswahili.
- Internet and phone network coverage in some areas especially in Ngorongoro district was an issue. This caused teams in these remote areas to delay sending data on time to the data manager. In some very remote areas it also caused some difficulty in reaching the head teachers, and some absent school teachers for the missed teacher interviews that were planned to be done over the phone.
- A very high number of head teachers (20%) were absent on the day of the survey. In those cases, the assistant head teacher was interviewed instead or if not available then the academic master or a teacher who is familiar with school records. All modules that could only have been answered by head teachers were not asked to the alternative respondents, and at the end of the fieldwork almost all the head teachers that were absent were called over the phone to administer the remaining missing modules.
- Two of the 50 fieldworkers selected at the end of the training dropped out before the pilot in Dodoma due to illness and sudden unavailability. Additionally, two weeks after fieldwork started one interviewer dropped out but fortunately one of the two interviewers who had dropped out before the pilot was available for replacement which did not impact on the data collection timeline.

Annex E Measurement of pupil learning outcomes

E.1 Summary of the content of the pupil tests

E.1.1 Rationale for using EGRA- and EGMA- type tests and matching to curriculum criteria

As explained in Chapter 3 which covers the quantitative impact evaluation design, the baseline pupil tests were adapted from existing EGRA and EGMA instruments. These are competency-focused instruments. Part of the decision to use these types of instruments for the EQUIP-T measurement of pupil learning was because at the time of the baseline impact evaluation survey, the Government had recently used EGRA and EGMA instruments in a survey to monitor its then flagship national education programmes BRN-Ed, and as a baseline for another national education programme LANES. The Government was in the process of setting national targets related to the results of these tests.

As set out in the impact evaluation baseline report (OPM 2015b, pp100,103), the skills tested in the impact evaluation tests were matched, as far as possible, to the 'specific objectives' laid out in the existing Standards 1 and 2 Kiswahili and mathematics curricula (MoEVT, 2005a,b) which explained what the pupil should be able to do to reach the curriculum standard (for example, Standard 2 pupils should be able to add numbers to get a sum not exceeding 1000). Two tables in the baseline report (one each for Kiswahili and maths) set out the list of skills that pupils had to demonstrate in the impact evaluation tests to be considered as achieving in one of five curriculum-linked performance bands (OPM 2015b, pp102,104). The competencies required to move up the scale are in a logical order of increasing difficulty, and it was noted that 'these [competencies] appear to be broadly consistent with the order of the competencies expressed in the Standard 1 and 2 curriculum, although in many cases the curriculum statements are fairly general and similar at the two levels' (OPM 2015b, p100).

Subsequent to the baseline impact evaluation research, the Government rolled out a new Standards 1 and 2 curriculum in 2015 which focuses on the 3Rs competencies of reading, writing and arithmetic (MoEVT 2016). Two further rounds of nationally representative EGRA and EGMA surveys have been carried out since then to continue the monitoring of the Government's national education programmes (see RTI 2016, for the results of the EGRA and EGMA surveys conducted in February 2016). Targets for two of the indicators captured in these surveys, form part of the agreed results that trigger disbursements from a group of development partners (DFID, SIDA and World Bank) under the national EPforR programme that replaced BRN-Ed. The indicators and targets are in Box 6 This emphasis on core EGRA and EGMA results is a strong indication that the Government considers these to be valid instruments for measuring early grade learning progress.

Box 6: National 3R assessment targets

Disbursement linked results (DLRs) related to early grade learning in the EPforR programme

DLR 6.2: average words per minute (WPM) read during EGRA assessment

Target: baseline (2013) 17.9 WPM; target (2018) 25 WPM

DLR 6.3: average score on level 2 addition and subtraction questions answered during EGMA assessment

Target: baseline (2013) 22.6% correct; target (2018) 30% correct

Source: MOEST and PO-RALG, 2018a.

Because of the content and nature of the baseline impact evaluation tests, there was no need to adapt them to fit with the new Standards 1 and 2 curriculum. The same pupil tests were used in all three rounds of the impact evaluation surveys. This also has the advantage of making the raw-score results from traditional test analysis comparable over time (see Section E.2 below). A similar exercise of mapping the skills tested in the impact evaluation tests to the 'competence benchmarks' in the new

curricula found that the baseline classification of skills into performance bands (below Standard 1, emerging Standard 1, achieving Standard 1, emerging Standard 2 and achieving Standard 2) is still valid for comparative purposes, while acknowledging that this judgement is necessarily subjective to some extent given the broad descriptions of skills in the curriculum.

The same limitations as were acknowledged at baseline with the old curriculum, also apply to the new curriculum mapping of skills into the performance bands:

- Many of the 3Rs curriculum statements are fairly general (for example, Standard 1 pupils should be able to read aloud with appropriate speed) and so cannot be mapped with precision into performance bands. The hierarchy of skills acquisition is preserved in the performance bands, so that as pupils are able to read faster they fall into higher performance bands.
- The impact evaluation tests do not cover all of the competencies listed in the new 3Rs curriculum. In maths they also cover one additional competency (simple multiplication). The skills that are covered are discussed next. Table and Table 51 in this annex show the mappings of curriculum competencies to the performance bands used in the impact evaluation.

E.1.2 Kiswahili

Skill areas: There are seven subtests in total. Each subtest covers a different skill area:

- Four subtests are timed oral reading tests of: syllables, familiar words, invented words and a short passage;
- The remaining three subtests cover: reading comprehension (five questions) based on the short passage read, listening comprehension (five questions), and writing/spelling dictated sentences (two sentences).

Curriculum levels: the short passage was designed to be a Standard 2 level text and so the reading comprehension questions which relate to this are classified as Standard 2 level questions.²⁵ The remaining subtests combine Standard 1 and Standard 2 curriculum skills by including questions of different levels within each subtest. The 3Rs Standard 2 curriculum requires that pupils read text with accuracy and fluency and, although this is not specified in the curriculum itself, the Government set a national benchmark for reading speed of 50 words per minute for Standard 2 pupils, following the national EGRA and EGMA assessment in 2013.²⁶ The national EPforR programme also monitors Standard 2 reading speed as an intermediate indicator against this benchmark (MOEST/PO-RALG, 2018b, p.153).

E.1.3 Mathematics

Skill areas: There are six subtests containing 60 questions in total. These cover: number comparison/quantity discrimination (eight questions), missing numbers in sequences (eight questions), addition (16 questions), subtraction (16 questions), multiplication (8 questions), and word problems (4 questions).

Curriculum levels: Apart from multiplication, the other five subtests contain a mix of Standard 1 and Standard 2 level questions. Multiplication was part of the previous Standard 2 level curriculum, but it is not part of the new 3Rs curriculum for Standard 2. For comparability with baseline, the multiplication

²⁵ The reading passage was developed by a team of experienced Tanzanian subject and test design specialists (see OPM 2015b for more details).

²⁶ Well-known international research (Abadzi, 2006) found that reading at 45-60 words per minute is a minimum fluency required for comprehension.

subtest was retained in the impact evaluation test. Over the whole test, the balance is skewed towards Standard 1 level material; about 60% of the questions are at the lower curriculum level.

E.2 Notes on traditional test analysis in the IE

Traditional test analysis relies on simple descriptive statistics of the different subtest results, such as mean test scores, mean reading speeds, and the proportion of pupils achieving more than x% of questions correct. These supplementary results are in Chapter 5 Section 5.1. In interpreting these results, it is important to understand how the subtests were marked, and how non-response was treated.

Marking of the Kiswahili subtests: The four reading subtests are ‘marked’ using a simple reading speed indicator: number of words correctly read per minute. Each pupil was given exactly one minute to complete each reading test. If a pupil finished early, this was accounted for in the reading speed. For the remaining subtests, marks are awarded as follows: reading comprehension (five marks: one per question); listening comprehension (five marks: one per question); writing (21 marks for spelling words and punctuation).

Marking of the maths subtests: One mark is given for each question answered correctly. The number of questions in each subtest is given above.

Treatment of non-response: ‘Non-response’ is treated as incorrect on all subtests in the traditional test analysis, except the four reading speed subtests in Kiswahili because this does not affect the ‘reading speed’ indicator. Most non-response happened because of instructions in the test to skip questions, to enhance the efficiency of the subtest, when a pupil got a fixed number of prior questions incorrect. The test designers sought to make the questions in each subtest hierarchically difficult. In Kiswahili, for example, the writing subtests contained two sentences, if the pupil was unable to write any word correctly in the first sentence, then the second sentence was skipped. In mathematics, for example, the addition and subtraction questions were divided into two levels, with level two questions designed to be more difficult than level one questions. If a pupil did not get any level one questions correct (one and two digit problems) then level two questions (two and three digit problems) were skipped. Given this hierarchical ordering of questions within the subtests, it was deemed reasonable to treat the skipped questions as incorrect, as it is very unlikely that the students who were unable to answer the less difficult items correctly would have been able to answer the more difficult items correctly if they were administered to them.

E.3 Application of the Rasch model in the impact evaluation

This section explains the rationale for using Rasch modelling to analyse the Kiswahili and maths test data for the impact evaluation. It discusses the principles underpinning the Rasch model and some of its key assumptions.

The key advantage of using Rasch modelling to analyse pupil test scores for the impact evaluation, is that, under certain assumptions (explained below), this generates estimates of pupil ‘ability’ in Kiswahili and mathematics on an *interval scale* which can be directly linked to criterion-referenced competencies found in the curriculum. On an interval scale, equal differences between numbers (in this case, pupil ability estimates) reflect equal differences in the amount of the underlying attribute being measured. Since the key objective of the impact evaluation is to measure change in learning achievement over time, an interval measurement scale allows for more accurate estimation of change. Using raw scores and traditional test analysis for this purpose can be substantially misleading (Wright and Stone 1979).

A key principle underlying the Rasch model is that of seeking to measure a latent unidimensional trait. This simply means an underlying construct that cannot be measured directly and can be thought of in terms of more or less. The impact evaluation seeks to measure the latent unidimensional traits of literacy skills (in Kiswahili) and numeracy (a type of mathematical) skills.

The Rasch model is the simplest item response theory (IRT) model. It is a probabilistic mathematical model of a person's (in this case, a pupil's) response to a test item whereby, relative to an item of a certain difficulty, as a pupil's level of ability (as estimated across all items) increases, the probability of a correct response increases. The latent trait is conceived as a single dimension along which items can be located in terms of their difficulty and persons can be located in terms of their ability. The model estimates the probability of answering the item correctly as a logistic function of the difference between the person's ability and the item's difficulty. This can be seen in the formula below, which shows the form of the Rasch model for dichotomous responses (either correct or incorrect):

$$P\{x_{vi} = 1 \mid \beta_v, \delta_i\} = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}}$$

Where P depicts that the model is of a probabilistic nature, $x_{vi} = 1$ means a correct response for a particular person and item combination, and β_v and δ_i are respectively the ability of person v and the difficulty of item i

The Rasch model enables the creation of a common interval scale of scores for both the item difficulties and the person abilities, and these scores are scaled in logits. The Rasch model has the property of specific objectivity, which is a statistical form of invariance whereby the ability estimates do not depend upon the specific items used, and the item difficulty estimates do not depend upon the specific sample that were assessed.²⁷ This is its principle advantage over other IRT models. The Rasch model is easily extended under the same core principle to items with ordered-category (polytomous) responses. The analysis in this evaluation applies both the dichotomous and the polytomous Rasch models as applicable to different items, as explained in the next section.

Rasch models have statistics to evaluate the fit of the item responses to the model. This can be used to determine whether to keep all of the items in the analysis, and also to provide insights into how to improve the tests for future applications.

Source: this text was partly extracted from: Cueto et al. 2009.

E.4 Rasch analysis of Kiswahili baseline, midline and endline pupil test data

This section explains the steps taken in producing the estimates of pupil ability in Kiswahili that are presented in Volume I of this report. Where relevant, it also summarises the results from key diagnostic tests that were used to assess the fit of the item response data to the Rasch model. This work builds on the Rasch analysis of baseline and midline Kiswahili test data reported in the baseline and midline evaluation reports (OPM 2015b, pp97-108, OPM 2016b, pp196-202). The Kiswahili performance band descriptor table, which describes the skills that pupils have achieved at each band-level has been reproduced from baseline, with notes to explain a few modifications revealed by the midline and endline data.

²⁷ The Rasch model encompasses a fundamental criterion of measurement, that of invariance (specific objectivity). This requirement is independent of any particular dataset. In the case of pupil test data, the criterion of invariance means that comparison between the measures of pupil ability is independent of the set of test items used, and comparison between measures of item difficulty are independent of which pupils are used.

E.4.1 Overall treatment of Kiswahili items in the Rasch analysis

At baseline, the Rasch analysis of item fit led to two subtests being deleted from the analysis: listening comprehension and reading syllables (OPM 2015b, p105). Similar analyses of item fit using midline and endline data revealed similar misfit to the Rasch model and these subtests were also excluded from the midline and endline analyses. Item fit was primarily explored using item characteristic curves (ICCs) which compare predicted item responses from the Rasch model to observed item responses—if the data fit the Rasch model (and hence satisfy its properties) then observed item responses (for each class interval) will lie on the expected ICC curve. The ICCs of the listening comprehension items showed very poor discrimination, possibly because some of the items could be answered using common sense rather than requiring deduction from the listening passage. At baseline, the syllables subtest was found to systematically discriminate less than the other items, it had disordered categories²⁸, and there was evidence that the subtest was dimensionally divergent from the other subtests. At midline and endline, the syllables subtest was also found to be dimensionally divergent.²⁹

After dropping the two subtests, the three remaining reading subtests (familiar words, non-words, and story passage) are treated as separate polytomous items, which means that there are more than two answer categories. The answer categories are all possible reading speeds up to a cut-off speed where there were very few responses above this. All responses at or above the cut-off speed are included in one answer category. For example, on the familiar words subtest, the answer categories range from one word per minute to a top category of 46 words per minute or above.

For the remaining subtests, each is treated as a testlet in the analysis to account for the dependence between them.³⁰ In the analysis, testlets are treated as polytomous items with thresholds. The number of answer categories for each testlet equals the number of questions in each subtest. Answer categories are of the form 'x correct out of y questions in total'. There are 5 reading comprehension, 13 writing-spelling, and 8 writing-punctuation questions. So, for example, for reading comprehension, answer categories are 1 out of 5, 2 out of 5, 3 out of 5, 4 out of 5 and 5 out of 5.

E.4.2 Steps taken in estimating Kiswahili item difficulty

This subsection briefly explains the treatment of baseline item response data that was used to estimate item difficulty (i.e. the location of items on the common scale) at baseline. In theory, the midline and endline item response data should reveal similar estimates of item locations because of the criterion of invariance embedded in the Rasch model. Hence the second step is to compare estimated item locations at baseline, midline and endline from independent analyses. The final step reports on diagnostic tests used to reveal how well the endline item response data fits the Rasch model (when items have been anchored to the baseline item locations).

STEP 1: Recap assumptions about the treatment of non-response in the baseline dataset used to estimate item difficulties. When pupils did not respond to an item, this was treated as an incorrect response for some questions and as missing data for other questions. It is not necessary to have every pupil answer every question to estimate item difficulty accurately (because of the specific objectivity property of the Rasch model), and so where it was more difficult to determine whether

²⁸ The term 'disordered categories' means that the ordinal numbering of categories does not correspond with their substantive meaning. In this case, it meant that some slower syllable reading speeds were found higher on the scale than some faster reading speeds.

²⁹ Dimensionality was assessed by looking at the principle components (PC) analysis of residuals.

³⁰ After combining the reading comprehension, spelling and punctuation items into three testlets, the residual correlations between items were acceptable.

pupils who did not respond to questions were in reality unlikely to know the answers, the data was coded as missing:

- Reading speed subtests: non-response is not relevant to the answer categories which simply require the number of words read correctly.
- Reading comprehension: non-response is treated as incorrect. There are two cases of non-response: the first is where the enumerator asks the pupil a question based on the passage which the pupil has just read and the pupil does not give an answer; the second is where the pupil is not asked a particular question by the enumerator because he/she did not read at sufficient speed to reach the part of the passage relevant to the question.
- Writing: in the first sentence, non-response is treated as incorrect, while in the second sentence, non-response is treated as incorrect unless all responses are non-responses; in the latter case these are treated as missing, and the pupil's response for the entire testlet treated as missing. The second sentence was automatically skipped if the pupil failed to write any word correctly in the first sentence. (Note that the treatment of missing data here is different to the treatment described in Section E.2 above for the traditional test analysis.)

STEP 2: Compare item locations from independent analyses of baseline, midline and endline data. Table 49 below shows that item locations from independent Rasch analyses of the baseline, midline and endline Kiswahili item responses, are fairly similar, as expected under Rasch model assumptions, for all of the subtests except punctuation. The item location estimate for punctuation is very similar in the midline and endline analysis, but this differs from the baseline estimate by about 0.4 logits. As noted in the midline report, this implies that the punctuation subtest became considerably easier for pupils of the same overall estimated ability between baseline and midline. It is difficult to know why this might have happened. One possibility is that, following the introduction of the 3Rs curriculum which has writing as a subject after baseline, pupils became more used to writing sentences and so the test format is more familiar making it easier for pupils to demonstrate their skills.³¹

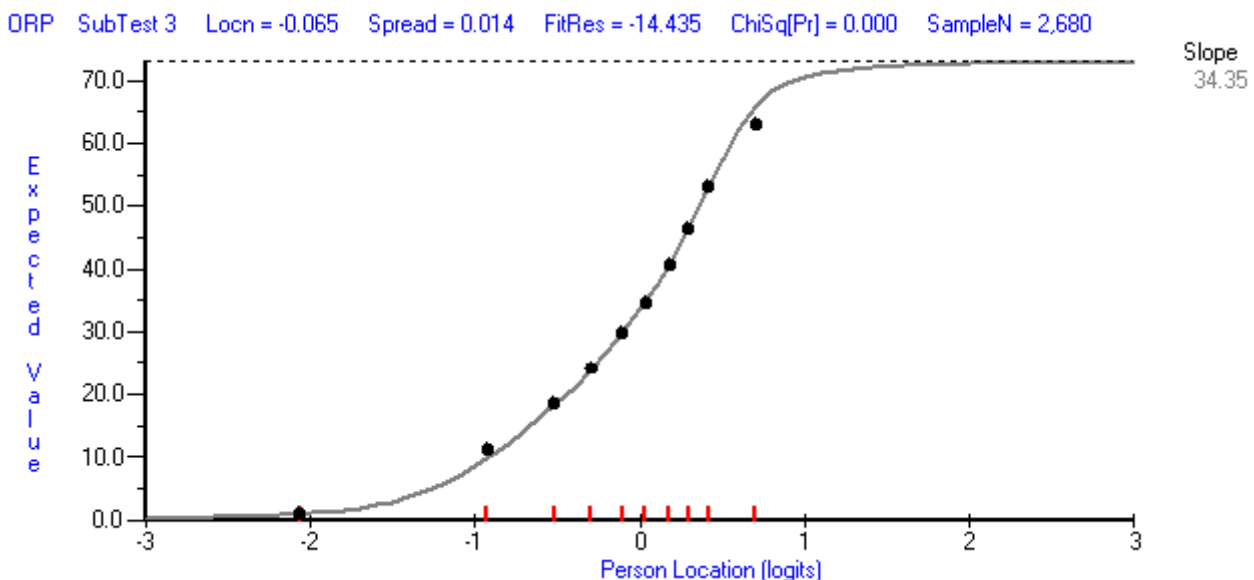
³¹ As noted in the midline report, this could help to explain why the estimated locations for the first two thresholds in the punctuation testlet are so much lower in the midline analysis compared to the baseline, but this does not explain why the top threshold is considerably higher at midline than baseline.

Table 49: Comparison of estimated Kiswahili item locations from independent baseline, midline and endline Rasch analyses

	Item locations (logits)			Differences (logits)	
	Baseline	Midline	Endline	EL-BL	EL-ML
Reading familiar words	-0.11	-0.05	-0.02	0.10	0.03
Reading non-words	0.03	0.18	0.18	0.15	0.00
Reading passage	-0.07	-0.01	0.00	0.07	0.01
Reading comprehension	0.63	0.72	0.65	0.03	-0.06
Spelling	-0.57	-0.50	-0.51	0.06	-0.01
Punctuation	0.09	-0.35	-0.31	-0.40	0.03

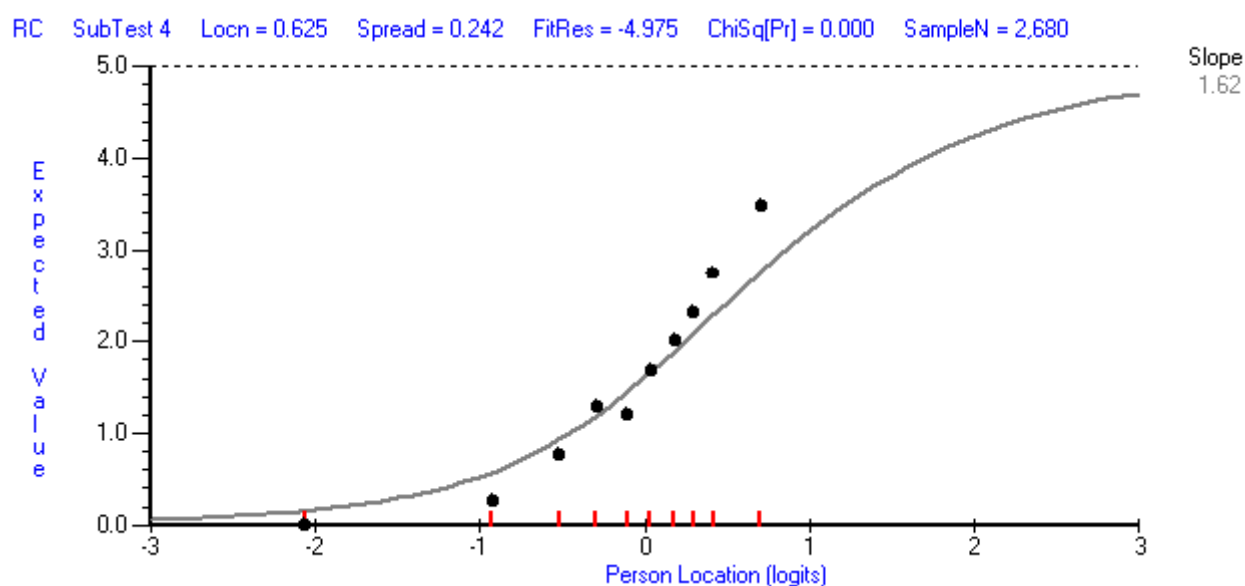
Source: Baseline, midline and endline Kiswahili pupil test data

STEP 3: Use baseline item locations to anchor items (except for punctuation items) in the Rasch analysis of endline Kiswahili item responses and then assess item fit. Using the same method as was applied to the midline item response data, the item locations in the Rasch analysis of endline test response data were constrained ('anchored') to the baseline locations shown in the table above, apart from the punctuation item. Item fit was then examined primarily using ICCs. The ICCs show good fit for the three reading speed tests in particular, with observed values for all class intervals either lying on, or very close to, the ICC curve (which shows the values predicted by the Rasch model). The figure below is the ICC for the oral reading passage subtest revealing that this item fits the Rasch model well (the corresponding ICCs for the baseline and midline data are in OPM 2015b, p.106 and OPM 2016b, p. 192).

Figure 37 ICC for oral reading passage subtest, endline

Source: Endline survey, pupil Kiswahili test.

As at baseline and midline, the worst fitting item is the reading comprehension testlet. Its ICC below shows that this item is discriminating higher than the average discrimination across all items. This was also the case with this item at baseline and midline. On balance, over-discrimination is less of a problem than under-discrimination and the misfit is not extreme, and so this item was retained.

Figure 38 ICC for reading comprehension subtest, endline

Source: Endline survey, pupil Kiswahili test.

All of the subtests have ordered thresholds, except for non-words where at higher categories, a faster reading speed did not necessarily correspond with higher level of ability as assessed across all items. However, this issue is confined to the very top categories (above 30 words per minute) where there are far fewer observations and so the results are less reliable. This was also observed with the baseline and midline data, and is not considered serious enough to warrant deleting this item.

Generally, the tests of item fit to the Rasch model applied to the endline data gave very similar results to the same tests applied to the midline data, which gives confidence that the enumerators administered the tests in the same way in each round.

E.4.3 Steps taken in estimating person abilities in Kiswahili

This subsection first explains why the test data used to estimate person abilities requires different assumptions about non-response to those used above to estimate item difficulty. It then explains the steps taken to estimate person abilities (pupil Kiswahili Rasch scores) reported in Volume I, and reports on the key diagnostic tests used to examine person fit to the Rasch model.

Step 1: Make appropriate assumptions about the treatment of non-response in the pupil test data. If non-response is treated as missing in some of the subtests (as was assumed in Step 1 above for the estimation of item difficulty), and then used to estimate person Rasch scores, it can advantage persons who were administered or attempted less items, and generally leads to different estimates for pupils who achieved the same overall score but were administered and/or attempted different numbers of items. In other words it can cause incoherence, particularly in the estimates of person ability at the lower end of the ability range. This was the case with the test data from this survey i.e. when some non-response was treated as missing, students who did not get any items correct were getting different estimates based on the number of items they were administered or attempted.

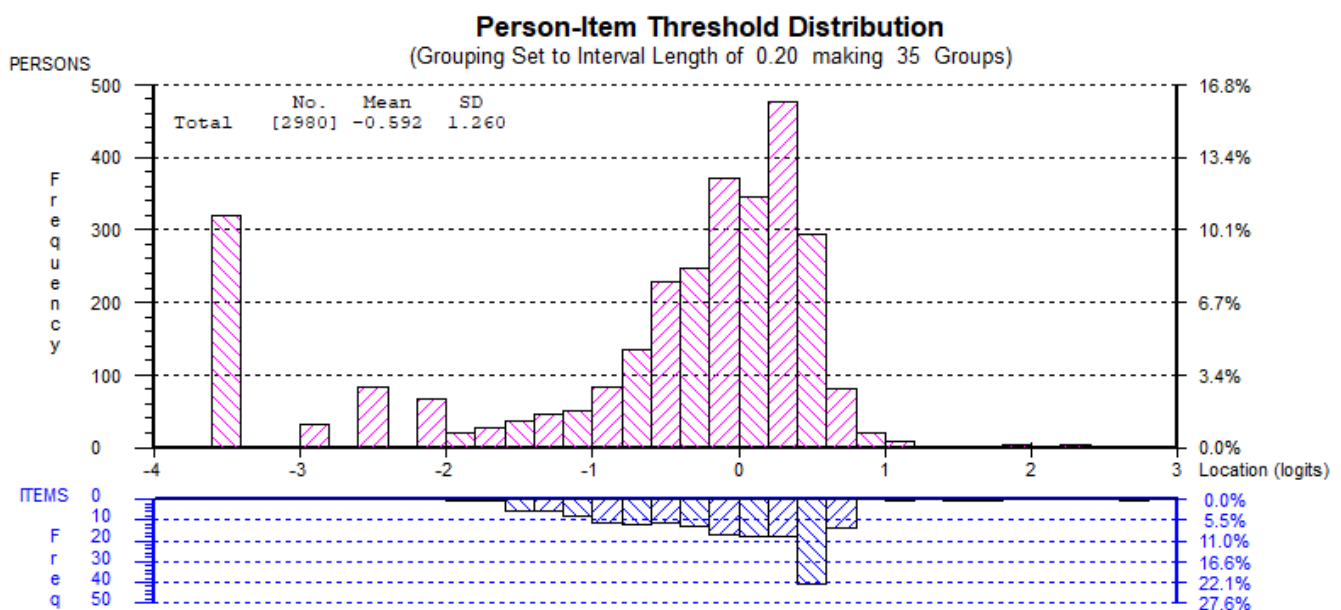
In order to estimate person abilities as accurately as possible, similar to the discussion in Section E.2., it was deemed reasonable to assume that pupils with missing responses in the second sentence of the writing test were highly unlikely to have been able to write the second sentence if they were unable

to write the first sentence at all.³² So in the analysis which follows, all non-response is treated as incorrect in the estimation of person abilities.

STEP 2: Estimate endline person abilities using the data treated as in Step 1, and use baseline item locations to anchor items (except for punctuation items). This is the Rasch analysis that produces the endline pupil ability estimates presented in Volume I (Chapter 3). This is the same method that was applied to the baseline and midline data. Here is a summary of the results from the diagnostic tests to assess fit of the endline data to the Rasch model:

- **Test score reliability:** The person separation index (PSI, which is Rasch's equivalent of Cronbach's alpha used in traditional test analysis) is high at 0.91 demonstrating good internal consistency reliability for the test, and excellent power to detect misfit.
- **Test targeting:** The average difficulty of the items (constrained to be 0) was quite difficult relative to the average pupil ability estimate (unweighted mean = -0.592, standard deviation 1.260). However, the distribution of persons and items in Figure 39 clearly shows the bi-modal distribution of pupil ability estimates with a large-floor effect. If the pupils at the lower extreme are excluded, then the test is slightly too easy for the average pupil.
- **Person fit:** the mean person fit residual is -0.431, which is fairly close to the expected value of 0, which suggests that the misfit to the Rasch model is not extreme.

Figure 39 Kiswahili person-item distribution at endline



Source: Endline survey, pupil Kiswahili test.

E.4.4 Kiswahili performance band descriptors

The description of the skills required to achieve at each of the five Kiswahili curriculum-linked performance bands has not changed since baseline, and so the table below from the baseline report (OPM 2015b, p102) is still applicable.

³² The writing subtest consists of two sentences. The second sentence was designed to be of a similar standard to the first.

Table 50: Kiswahili performance band descriptors

Score range	Items	Competency descriptor
Band 0 Below emerging skills at std 1 level		
< -1.61 logits	None	Not applicable
Band 1E Emerging skills at std 1 curriculum level: pupils have achieved at least some of the skills below		
Between -1.61 and -0.76 logits	FW: 1 to 9	Read familiar words at a speed of between 1 word and 9 words per minute
	NW: 1 to 5	Read non-words at a speed of between 1 word and 5 words per minute
	ORP: 1 to 13	Read a simple story at a speed of between 1 and 13 words per minute
	WSSp: 1 to 5	Spell between 1 and 5 words correctly out of 13. The spelling test included five simple short words of up to 4 letters (na, la, je, lina, letu).
	WSPu: 1 to 3	Partly punctuate sentences correctly, by getting between 1 and 3 punctuation requirements out of 8 correct. The punctuation requirements included writing text from left to right and using spacing between words. ¹
Band 1A Achieving skills at std 1 curriculum level: pupils have achieved all band 1E skills and at least some of the skills below		
Between -0.76 and -0.08 logits	FW: 10 to 20	Read familiar words at a speed of between 10 words and 20 words per minute
	NW: 6 to 13	Read non-words at a speed of between 6 words and 13 words per minute
	ORP: 14 to 30	Read a simple story at a speed of between 14 and 30 words per minute
	WSSp: 6 to 10	Spell between 6 and 10 words correctly out of 13, including very familiar words (shamba, shule), and simple longer words (kuvutia, darasa)
	WSPu: 4 to 5	Partly punctuate sentences correctly, by getting between 4 and 5 punctuation requirements out of 8 correct. The punctuation requirements included the use of capital letters at the start of a sentence.
Band 2E Emerging skills at std 2 curriculum level: pupils have achieved all band 1E and band 1A skills and at least some of the skills below		
Between -0.08 and 0.37 logits	FW: 21 to 30	Read familiar words at a speed of between 21 words and 30 words per minute
	NW: 14 to 21	Read non-words at a speed of between 14 and 21 words per minute
	ORP: 31 to 49	Read a simple story at a speed of between 31 and 49 words per minute
	RC: 1 to 2	Answer 1 to 2 out of 5 simple reading comprehension questions correctly based on a reading a short passage, including 2 fact-based qns.
	WSSp: 11	Spell 11 words correctly out of 13. The spelling test included simple longer words (e.g. linapendenza).
Band 2A Achieving std 2 curriculum level or above: pupils have achieved all band 1E, band 1A, and band 2E skills and at least some of the skills below		
More than 0.37 logits	FW: 31 or above	Read familiar words at a speed of 31 words or more per minute
	NW: 22 or more	Read non-words at a speed of 22 or more words per minute
	ORP: 50 or more	Read a simple story at a speed of at least 50 words per minute
	RC: 3 to 5	Answer 3 to 5 out of 5 reading comprehension qns correctly based on a reading a short passage. The test included deductive and inferential qns.
	WSSp: 12 to 13	Spell 12 to 13 words correctly out of 13. The test included simple words containing r/l (karoti) and more complex words (njegere).
	WSPu: 6 to 8	Punctuate sentences correctly, by getting between 6 and 8 punctuation requirements out of 8 correct. The punctuation requirements included the use of a full stop at the end of a sentence, and the use of a question mark at the end of a sentence.

Source: OPM 2015b, p102. Note: (1) The estimated item locations for punctuation skills differ between baseline and midline, and are very similar between midline and endline. Questions were systematically easier for midline and endline pupils of the same ability compared with baseline pupils, but these differences do not change the bands that the different levels of punctuation skills fall into, except for the first skill level (getting 1 punctuation question correct) where the midline and endline item locations fall into band 0 rather than band 1E.

E.5 Rasch analysis of maths baseline, midline and endline pupil test data

This section explains the steps taken in producing the estimates of pupil ability in maths that are presented in Volume I of this report.

E.5.1 Overall treatment of maths items in the Rasch analysis

Each question in the maths test is treated as a dichotomous item, which means that there are two answer categories: correct or incorrect. At baseline, one item was dropped (number discrimination, question 6) because the ICC for this item showed a pattern consistent with guessing, and poor, and at times negative discrimination, i.e., lower ability students performed better than higher ability students. This pattern was also picked up in the midline data, although the misfit was less extreme, and so this item was dropped from the midline data as well. The endline data did not reveal such poor misfit as to warrant deleting this item from the analysis.

E.5.2 Steps taken in estimating maths item difficulties

This subsection briefly explains the treatment of baseline item response data that was used to estimate item difficulty (i.e. the location of items on the common scale) at baseline. In theory, the midline and endline item response data should reveal similar estimates of item locations because of the criterion of invariance embedded in the Rasch model. Hence the second step is to compare estimated item locations at baseline, midline and endline for all items³³, to identify whether any items are showing differential item functioning (DIF) between survey waves. This analysis combines a statistical approach to identifying DIF with an inspection of the ICCs split by survey wave. The same type of analysis was carried out at midline, and this found that 14 items showed DIF between baseline and midline. Hence to produce the midline estimates, all items were anchored to their baseline locations, except for these 14 items. The same method has been applied to the endline data, and the final step reports on diagnostic tests used to reveal how well the endline item response data fits the Rasch model (when all items, except the set that exhibit DIF, have been anchored to the baseline item locations). These steps are explained in more detail below.

STEP 1: Recap assumptions about the treatment of non-response in the baseline dataset used to estimate item difficulties. In the baseline maths dataset, all non-response is treated as missing. Most non-response occurs automatically in the test because of automatic skips (explained earlier in Section E.2). Some non-response also occurs when pupils are asked a question and they do not reply in the time allocated. The rationale for leaving the non-response data as missing when estimating item locations is that it is not necessary to have every pupil answer every question to estimate item difficulty accurately because of the specific objectivity property of the Rasch model, and so no assumptions were necessary regarding the reasons for the non-responses.

STEP 2: Investigate DIF by surveywave (baseline to midline to endline). A Rasch analysis of the combined baseline, midline and endline datasets identified 17 items out of 59 common items which showed patterns of uniform DIF by surveywave. An iterative approach was then taken to split these items, starting with the item with the largest statistical indicator of DIF, in order to potentially identify items with real DIF, as opposed to artificial DIF, which is an artefact of parameter estimation when some items have real DIF. At each stage, the ICCs for the items which showed statistical DIF were inspected to confirm that they support the DIF statistics. This process

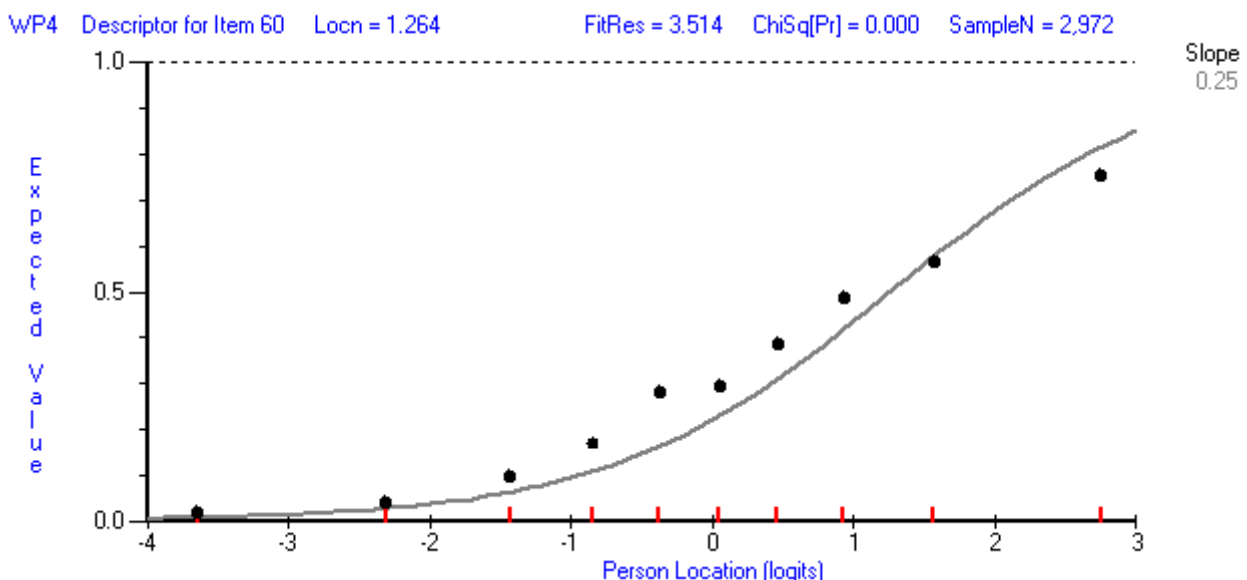
³³ Excluding the number discrimination question 6, which had been dropped from the baseline and midline analysis.

confirmed that the 17 items exhibit survey wave DIF. In addition, two further items showed clear patterns of uniform DIF in their ICCs bringing the total that show surveywave DIF to 19 items.

STEP 3: Conduct Rasch analysis of endline maths item responses, using baseline item locations to anchor all items, except for the 19 items identified in Step 2 with surveywave DIF and the item that is only in the endline data (see E.5.1 above). Thus there are 40 items with common locations between baseline and endline. These are distributed across the different subtests as follows: number discrimination (2 items); sequences (6 items); addition (14 items); subtraction (16 items); and word problems (2 items). This means that the linking items cover all the main skills being tested, apart from multiplication. Multiplication items have become systematically more difficult for endline pupils compared to baseline pupils of the same overall ability level (this is not unexpected, see explanation in E.1.3) and so it is not possible to include these items as linking items. This means that construct coverage underpinning the person Rasch scores is slightly different between survey rounds.

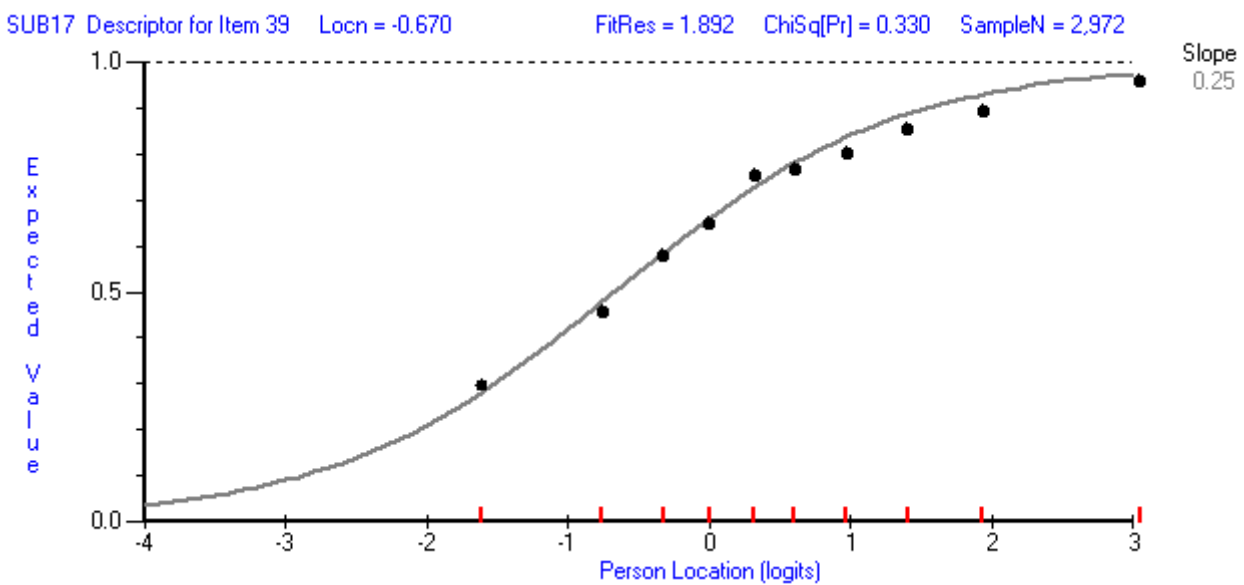
STEP 4: Examine endline item fit to the Rasch model, primarily using ICCs. The ICCs show reasonable fit in most cases, with observed values for all class intervals either lying on, or not far from, the ICC curve (which shows the values predicted by the Rasch model). Figure 40 is the ICC for the fourth item in the word problems subtest, presented here because the corresponding ICCs for the same item in the baseline and midline data are in previous reports (OPM 2015b, p108 and OPM 2016b). For this word problem, the endline ICC shows some predicted values on the ICC while many are slightly above. Item fit is worse than at baseline and midline, and overall this item shows a pattern of underdiscrimination. A better fitting item is subtraction question 7—its ICC in Figure 41 shows most observed values on or close to the ICC curve. The mean item fit residual is -0.296 which is somewhat different to the expected value of 0, but this is not large enough to suggest that overall item misfit is a serious problem.

Figure 40 ICC for word problem 4, endline



Source: Endline survey, pupil maths test.

Figure 41 ICC for subtraction question 7 from level 1, endline



Source: Endline survey, pupil maths test.

E.5.3 Steps taken in estimating person abilities in maths

This subsection first explains why the test data used to estimate person abilities requires different assumptions about non-response to those used above to estimate item difficulty. It then explains the steps taken to estimate person abilities (pupil Rasch maths scores) reported in Volume I, and reports on the key diagnostic tests used to examine person fit to the Rasch model.

STEP 1: Make appropriate assumptions about the treatment of non-response in the pupil test data. Similar to the rationale explained in Section E.4.3 for Kiswahili, for the purpose of estimating person abilities, it was deemed reasonable to assume that pupils with missing responses to maths items were highly unlikely to be able to answer the items correctly (because of the hierarchically difficult design of the test items as discussed in Section E.2). If the missing responses are not treated as incorrect for person ability estimates, the resulting estimates can be biased in the manner explained in Section E.4.3. So in the analysis which follows, all non-response is treated as incorrect in the estimation of person abilities. This is the same approach that was followed with the baseline and midline person ability estimates.

STEP 2: Estimate endline person abilities using the endline data treated as in Step 1, and use baseline item locations to anchor items (except for 20 items highlighted in Section E.5.2). This is the Rasch analysis that produces the endline pupil ability estimates presented in Volume I (Chapter 3). Here is a summary of the results from the diagnostic tests to assess fit with the Rasch model:

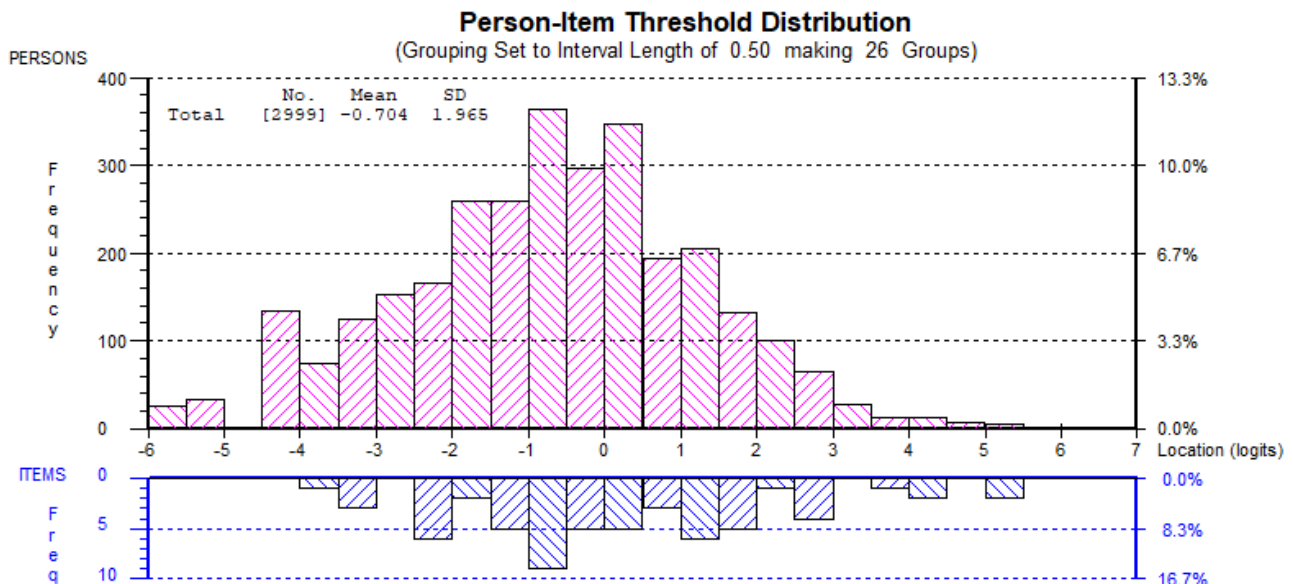
- **Test score reliability:** The person separation index (PSI, which is Rasch's equivalent of Cronbach's alpha used in traditional test analysis) is high at 0.95 demonstrating good internal consistency reliability for the test.
- **Test targeting:** The average difficulty of the items (estimated at 0.1 because of the anchoring procedure³⁴) was quite difficult relative to the average pupil ability estimate (unweighted mean = -0.704, standard deviation 1.965). Figure 42 shows that the distribution of pupil ability

³⁴ If no anchoring is applied then the mean item estimate is fixed at 0.

estimates is somewhat similar to a normal bell shape but is slightly skewed to the lower ability levels.

- **Person fit:** the mean person fit residual is -0.297, which deviates somewhat from the expected value of 0, but is not large enough to be considered as indicative of serious person misfit.

Figure 42 Maths person-item distribution at endline



Source: ML IE survey, pupil maths test.

E.5.4 Maths performance band descriptors

The description of the skills required to achieve at each of the five maths curriculum-linked performance bands is in Table 51 below. This is replicated from the baseline report (OPM 2015b, p104). As explained in E.5.2 above, some of the estimated item locations changed over the survey rounds which could potentially move them into different performance bands if the change is large enough.

Between baseline and midline, the estimated item locations for 14 out of 59 items changed, but the movement is not large enough to shift any of these items (and thus the skills they are measuring) into different performance bands. Between baseline and endline, the estimated item locations for 19 out of 59 items changed. For 12 of these items, the change in location is not large enough to move the items into different performance bands, but for seven items the endline estimate falls in an adjacent performance band (three of these are multiplication items, which, as already noted, got substantially harder for pupils at endline). These movements are noted in the footnote to Table 51. The additional item (number discrimination question 6) in the endline analysis that was dropped at baseline and midline is located in band 2E (emerging Standard 2 level)—this is also noted in the footnote to Table 51.

Table 51 Maths performance band descriptors

Score range	Items ¹	Competency descriptor
BAND 0 Below emerging skills at std 1 level		
<-3.29 logits	None	Not applicable
BAND 1E Emerging skills at std 1 curriculum level: pupils have achieved at least some of the skills below		
Between -3.29 and -1.40 logits	ND: 1,2,3,4	Compare two whole numbers containing one or two digits, and identify which is larger
	ADD1: 1, 2,3,4	Add whole numbers containing one digit to get a total not exceeding 10
	SUB1: 1	Subtract whole numbers with values less than five
	SEQ: 1,2	Fill in missing numbers in a sequence of whole numbers containing one or two digits (less than 20) going up in steps of one
BAND 1A Achieving skills at std 1 curriculum level: pupils have achieved band 1E skills and at least some of the skills below		
Between -1.40 and -0.11 logits	ND: 5,7,8	Compare two whole numbers containing three digits, and identify which is larger (first digit is identical in both numbers, so essentially it is a comparison of two digit numbers)
	ADD1: 5,6,7,8	Add whole numbers containing one and two digits to get a total between 10 and 20
	ADD2: 1,2,7	Add whole numbers containing one, two digits and three digits to get a total between 20 and 999 (no carrying needed)
	SUB1: 2, 3, 4, 6, 7	Subtract whole numbers containing one or two digits (less than 20) (no borrowing required)
	SUB2: 1,5,7	Subtract whole numbers containing two or three digits (no borrowing needed)
	WP: 1	Solve real-life problems given in words using addition of one digit numbers to get a total not exceeding 10
	MULT: 1,2	Multiply one digit numbers with value less than five (from the 2, 3 and 4 times tables)
BAND 2E Emerging skills at std 2 curriculum level: pupils have achieved band 1E and band 1A skills and at least some of the skills below		
Between -0.11 and 2.04 logits	ADD2: 3, 4, 5, 6, 8	Add whole numbers containing two digits and three digits to get a total between 20 and 999 (carrying needed)
	SUB1: 8, 5	Subtract whole numbers containing one or two digits (less than 20) (borrowing required)
	SUB2: 2, 4, 6, 8	Subtract whole numbers containing one, two or three digits (borrowing required)
	SEQ: 3, 5	Fill in missing numbers in a sequence of whole numbers containing two digits going up in steps of 10 Fill in missing numbers in a sequence of whole numbers containing three digits going up in steps of one
	WP: 2, 3, 4	Solve real-life problems given in words using: (i) subtraction of one digit numbers to get a total not exceeding 10; (ii) multiplication of one digit numbers to get a total not exceeding 20
	MULT: 3, 4, 5	Multiply whole numbers to get a product not exceeding 72
BAND 2A Achieving std 2 curriculum level or above: pupils have achieved band 1E, band 1A, band 2E and at least some of the skills below		
More than 2.04 logits	SUB2: 3	Subtract whole numbers containing one, two or three digits (borrowing required)
	SEQ: 4, 6, 7, 8	Fill in missing numbers in a sequence of whole numbers containing one, two or three digits: (i) going <u>down</u> in steps of two or steps of 10; (ii) going up in steps of two and five.
	MULT: 6, 7, 8	Multiply whole numbers to get a product not exceeding 72 (including 8,9 and 12 times tables)
Source: OPM 2015b, p104. Notes (1): the items highlighted in small grey boxes (e.g. SUB1:7) are exceptions to the description given on the adjacent line. (2) Although the estimated item locations for 14 of the 59 items differ between BL and ML, these differences do not change the bands that these items fall into. (3) 7 items moved into adjacent categories between BL and EL as follows: ND1 (1E to 0); SEQ1 (1E to 0); SEQ3 (2E to 1A); MULT1 (1A to 2E); MULT4 (2E to 2A); MULT 5 (2E to 2A); and WP1 (1A to 1E).		

Annex F Definition of key quantitative indicators

F.1 Chapter 3 Pupil learning and background characteristics

Indicator name	Indicator definition	Respondent / unit of analysis	Notes
Pupil learning in Kiswahili			
Standard 3 pupils rasch ability score in Kiswahili (mean score)	Mean pupil rasch ability estimate in Kiswahili, in logits	Standard 3 pupils	Estimates of pupil ability and item difficulty are estimated using Rasch analysis (item-response theory modelling). Both are mapped on to a common scale. The items relate to statements in the standard one and standard two curriculum, and can be used to draw performance band boundaries to mark, for example, the increasingly difficult skills required to move from one curriculum level to another. The performance band boundaries are defined using estimates of item difficulties linked to curriculum competencies and mapped on to the same scale as the pupil ability estimates.
Standard 3 pupils in Kiswahili performance band x (% Standard 3 pupils):	Number of Standard 3 pupils with ability scores that fall in or on the boundary of Kiswahili performance band x/all assessed Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	
Band 0: below Standard 1 level			
Band 1E: emerging Standard 1 level			
Band 1A: achieving Standard 1 level			
Band 2E: emerging Standard 2 level			
Band 2A achieving Standard 2 level			
Correct words from passage read per minute (mean words per minute)	Mean number of words read correctly from a passage per minute	Standard 3 pupils	
Pupil learning in mathematics			
Standard 3 pupils rasch ability score in mathematics (mean score)	Mean pupil rasch ability estimate in mathematics, in logits	Standard 3 pupils	Estimates of pupil ability and item difficulty are estimated using Rasch analysis (item-response theory modelling). Both are mapped on to a common scale. The items relate to statements in the standard one and standard two curriculum, and can be used to draw performance band boundaries to mark, for example, the increasingly difficult skills required to move from one curriculum level to another. The performance band boundaries are defined using estimates of item difficulties linked to curriculum competencies and mapped on to the same scale as the pupil ability estimates.
Standard 3 pupils in mathematics performance band x (% Standard 3 pupils):	Number of Standard 3 pupils with ability scores that fall in or on the boundary of mathematics performance band x/all assessed Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	
Band 0: below Standard 1 level			
Band 1E: emerging Standard 1 level			
Band 1A: achieving Standard 1 level			
Band 2E: emerging Standard 2 level			
Band 2A achieving Standard 2 level			
Pupil background characteristics			
Pupil is female (% Standard 3 pupils)	Number of Standard 3 female pupils/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	
Age (mean)	The average age of Standard 3 pupils	Standard 3 pupils	
Pupil is overage (% Standard 3 pupils)	Number of Standard 3 pupils aged 11 years or older/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	

Pupil has repeated a class (% Standard 3 pupils)	Number of Standard 3 pupils that were in Standard 3 or 4 last year/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	This is a proxy repetition rate, which uses a different denominator to the standard definition.
Pupil is from a household below poverty line (% Standard 3 pupils)	Number of Standard 3 pupils that come from a poor household/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	A pupil is considered 'poor' if he/she comes from a household that has a greater than 50% probability of being below the Tanzania national poverty line, and 'rich' otherwise.
Pupil ate before school (% Standard 3 pupils)	Number of pupils reporting that they ate something before school on the day of the survey/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	
Pupil does paid work outside household (% Standard 3 pupils)	Number of Standard 3 pupils who do paid work outside the household/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil does unpaid work outside household (% Standard 3 pupils)	Number of Standard 3 pupils who do unpaid work outside the household/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Main language spoken at home not Kiswahili (% Standard 3 pupils)	Number of Standard 3 pupils reporting that the main language spoken at home is not Kiswahili/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	
Books, newspapers at home (% Standard 3 pupils)	Number of Standard 3 pupils who have books, newspapers or other reading materials available in their home/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	
Pupil receives help at home with homework (% Standard 3 pupils)	Number of Standard 3 pupils who have someone at home to help with homework/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil never reads aloud at home (% Standard 3 pupils)	Number of Standard 3 pupils who never read aloud to someone at home/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil is never read to aloud by someone at home (% Standard 3 pupils)	Number of Standard 3 pupils who never have someone at home reading aloud to them/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil attended preschool (% Standard 3 pupils)	Number of Standard 3 pupils who attended preschool/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	
Type of preschool attended by pupil (% Standard 3 pupils):			
Government pre-primary			
Nurse / kindergarten			
Madrasah			
EQUIP-T SRP			
Other short programme			
Other			
Pupil attends extra tuition classes (% of Standard 3 pupils)	Number of Standard 3 pupils who attend paid extra tuition classes/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil absence from school, using headcount (% Standard 1-3 pupils)	The number of Standard 1-3 pupils who were absent from school on the day of the survey using a pupil headcount/all Standard 1-3 pupils enrolled in the school, expressed as a percentage.	Standards 1-3 pupils	Enumerators record all Standards 1-3 classes and count the number of pupils present on the day of the survey. The difference between this headcount and the number of enrolled Standards 1-3 pupils in the school records are considered the number of absent pupils on the day of survey.
Pupil absence from school, using school records (% Standard 1-3 pupils)	The number of Standard 1-3 pupils who were absent from school on the day of the survey using school records/all Standard 1-3 pupils enrolled in the school, expressed as a percentage.	Standards 1-3 pupils	Enumerators record the number of Standards 1-3 pupils enrolled in the school and the number of pupils present in each stream of Standards 1-3 on the day of the survey.

F.2 Chapter 4 Teacher performance

Indicator name	Indicator definition	Respondent / unit of analysis	Notes
Have teachers received EQUIP-T in-service training? (EQUIP-T input)			
Attended EQUIP-T in-service training last two years (% Standards 1-2 teachers)	Number of teachers of Standards 1-2 that report attending EQUIP-T in-service training the previous two years/all interviewed teachers of Standards 1-2, expressed as a percentage.	Standards 1-2 teachers	
Attended away or school-based EQUIP-T training (% Standards 1-2 teachers who attended EQUIP-T)	The number of teachers of Standards 1-2 that reported attending EQUIP-T training only away/only at school/ away and at school/all interviewed teachers of Standards 1-2 who attended EQUIP-T training, expressed as a percentage.	Standards 1-2 teachers	
Only away from school			
Only school-based			
Away and school-based			
Number of Kiswahili literacy training days away from school by year (school mean)	The average number of total Kiswahili literacy training days away from school by year	Schools	
Number of numeracy training days away from school by year (school mean)	The average number of total numeracy training days away from school by year	Schools	
Number of GRP training days away from school by year (school mean)	The average number of total GRP training days away from school by year	Schools	
Number of 3Rs training days away from school by year (school mean)	The average number of total 3Rs training days away from school by year	Schools	
Number of numeracy modules covered per session away from school (mean)	The average number of numeracy modules covered per session away from school	Schools	
Number of Kiswahili literacy modules covered per session away from school (mean)	The average number of Kiswahili literacy modules covered per session away from school	Schools	
Number of school-based training sessions held in 2015 (% schools)	The number of schools that held x days of school-based training sessions in 2015/all schools, expressed as a percentage.	Schools	
0 to 4 days			
5 to 9 days			
10 to 14 days			
15 or more days			
Number of school-based training sessions held in 2016 (% schools)	The number of schools that held x days of school-based training sessions in 2016/all schools, expressed as a percentage.	Schools	
0 to 4 days			
5 to 9 days			
10 to 14 days			
15 or more days			
Number of school-based training sessions held in 2017 (% schools)	The number of schools that held x days of school-based training sessions in 2017/all schools, expressed as a percentage.	Schools	
0 to 4 days			
5 to 9 days			

10 to 14 days			
15 or more days			
Number of school-based EQUIP-T training sessions by year (school mean)	The average number of total school-based training sessions held by school by year	Schools	
School-based training sessions that trained on content in each year: (mean % school-based training sessions in each year)			
Kiswahili literacy	The average of the (number of training sessions in each year which trained on content X/all training sessions held by a school in a given year, expressed as a percentage) across all schools.	Schools	
Numeracy			
GRP			
Particular curriculum competency			
3Rs			
Number of hours a typical daily school-based session took (mean hours)	The average number of hours a typical daily school-based training session took	School-based training sessions	
Completed all EQUIP-T early grade Kiswahili literacy modules (% Standards 1-2 teachers)	The number of teachers of Standards 1-2 that report competing all of the 13 EQUIP-T early grade Swahili literacy modules/ all interviewed teachers of Standards 1-2, expressed as a percentage.	Standards 1-2 teachers	
Number of EQUIP-T early grade Kiswahili literacy modules completed (mean)	The average number of EQUIP-T early grade Swahili literacy modules completed by teacher.	Standards 1-2 teachers	
Completed all EQUIP-T early grade numeracy modules (% Standards 1-2 teachers)	The number of teachers of Standards 1-2 that report competing all of the 9 EQUIP-T early grade numeracy modules/ all interviewed teachers of Stds 1-2, expressed as a percentage.	Standards 1-2 teachers	
Number of EQUIP-T early grade numeracy modules completed (mean)	The average number of EQUIP-T early grade numeracy modules completed by teacher.	Standards 1-2 teachers	
Completed the EQUIP-T gender responsive pedagogy module (% Standards 1-2 teachers)	The number of teachers of Standards 1-2 that report competing the EQUIP-T gender responsive pedagogy module/ all interviewed teachers of Standards 1-2, expressed as a percentage.	Standards 1-2 teachers	
Completed numeracy modules in school-based training (% schools)	The number of schools that have completed all 9 numeracy modules in school-based training/all schools, expressed as a percentage.	Schools	
Completed Kiswahili literacy modules in school-based training (% schools)	The number of schools that have completed all 13 Kiswahili literacy modules in school-based training/all schools, expressed as a percentage.	Schools	
Completed Gender-responsive pedagogy module in school-based training (% schools)	The number of schools that have completed the GRP module in school-based training/all schools, expressed as a percentage.	Schools	
Has teacher in-service training coordinator (% schools)	The number of schools that have a coordinator for teacher in-service training/all schools, expressed as a percentage.	Schools	
Teachers attending training away from school are Standards 1-3 teachers (mean % of teachers attending training away from school)	The average of the (number of teachers attending training sessions away from school that are Standards 1-3 teachers/all teachers attending training sessions away from school, expressed as a percentage) across all schools.	Schools	
Teachers attending training away from school are Standards 1-3 Kiswahili/maths teachers (mean % of teachers attending training away from school)	The average of the (number of teachers attending training sessions away from school that are Standards 1-3 Kiswahili/maths teachers/all teachers attending training sessions away from school, expressed as a percentage) across all schools.	Schools	
Teachers who attended the last school-based training session teach Standards 1-3 (mean % teachers attending)	The average of the (number of teachers at the school who attended the last school-based training session who teach Standards 1-3/all teachers who attended the last school-based training, expressed as a percentage) across all schools.	Schools	

Teacher attendance at EQUIP-T in-service training (EQUIP-T input to output assumption)			
Duration of school-based EQUIP-T training attended by Standard 1-2 teachers (mean days)	The average number of days of school-based training that Standards 1-2 teachers attended in the last two years	Standards 1-2 teachers	
Proportion of EQUIP-T training attended (% Standards 1-2 teachers):	The number of teachers of Standards 1-2 that report they attended all/most/some of the EQUIP-T school-based training sessions/all interviewed teachers of Standards 1-2 who attended EQUIP-T school-based training, expressed as a percentage.	Standards 1-2 teachers	
All sessions			
Most sessions			
Some sessions			
Relevance and accessibility of EQUIP-T in-service training (EQUIP-T input to output assumption)			
View of EQUIP-T training (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 that reported they found the EQUIP-T training useful/somewhat useful/not useful/all interviewed teachers of Standards 1-3 who attended EQUIP-T training, expressed as a percentage.	Standards 1-3 teachers	
Useful			
Somewhat useful			
Not useful			
Gains from EQUIP-T training (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 reporting gain x from the EQUIP-T training/all interviewed teachers of Standards 1-3 who attended EQUIP-T training and thought it was (somewhat) useful, expressed as a percentage.	Standards 1-3 teachers	
Curriculum knowledge			
Subject knowledge			
Teaching skills			
Gender-responsive teaching skills			
Inclusive teaching skills			
Classroom management/disciplinary skills			
Lesson planning skills			
Confidence in my teaching			
Support network			
Other			
What difficulties did you experience with EQUIP-T training (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 reporting difficulty x with the EQUIP-T training/all interviewed teachers of Standards 1-3 who attended EQUIP-T training and thought it was (somewhat) useful, expressed as a percentage.	Standards 1-3 teachers	
None			
Not relevant to my job			
Materials difficult to understand			
Too much content			
Too theoretical			
Took too much time/work load			
Limited training time			
Time lag between training events			
Sessions inconvenient time/day			

Transport difficult / venue too far			
No/insufficient payment			
No/insufficient direct training			
Envy from colleagues			
Not enough training material			
Content not completed			
Problems with trainers			
Other			
Challenges with EQUIP-T training away from school, reported at school-level (% schools):			
Materials difficult to understand	The number of schools that reported challenge x with EQUIP-T training away from school/all schools, expressed as a percentage.	Schools	
Too much content			
Too theoretical			
Took too much time/work load			
Limited training time			
Time lag between training events			
Transport difficult / venue too far			
Venue inadequate			
No/insufficient payment			
Envy from colleagues			
Not enough training material			
Content not completed			
Too few trainers			
Trainers not competent			
Training groups too large			
Short notice			
Other			
Challenges with EQUIP-T school-based training, reported at school-level (% schools):			
Materials difficult to understand	The number of schools that reported challenge x with EQUIP-T school-based training/all schools, expressed as a percentage.	Schools	
Too much content			
Too theoretical			
Took too much time/work load			
Limited training time			
Time lag between training events			

No/insufficient payment			
Not enough training material			
Content not completed			
Trainers not competent			
Participants not motivated			
Sessions inconvenient time/day			
Loss of information			
No /insufficient direct training outside school			
Other			
Improvements to the in-service training for teachers (% schools):			
Allowance for school-based training			
More allowance			
Train more teachers			
Supply more training materials	The number of schools that suggested improvement x to the EQUIP-T training/all schools, expressed as a percentage.	Schools	
Train when the school is closed			
Less content/more time			
More training for inspectors/WEOs/DEOs			
Reduce other teacher tasks			
Other			
School-based training sessions had: (mean % school-based training sessions in a school):	The average of the (number of school-based training sessions which had X number of facilitators/all training sessions held by a school, expressed as a percentage) across all schools.	Schools	
1 facilitator			
2 facilitators			
3 or more facilitators			
School-based training session had INCO as one of the facilitators (mean % school-based training sessions)	The average of the (number of school-based training sessions which had INCO as one of the facilitators/all training sessions held by a school, expressed as a percentage) across all schools.	Schools	
Age of facilitators of school-based training sessions (mean years)	The average age of the facilitators of school-based training sessions.	Schools	All facilitator indicators are weighted by the proportion of sessions each facilitator has facilitated.
Time facilitators of school-based training sessions have been teaching (mean years)	The average length of service of the facilitators of school-based training sessions.	Schools	
Facilitators of school-based training sessions are (mean % facilitators):	The average of the (number of facilitators of school-based training sessions who have title x/all facilitators of training sessions, expressed as a percentage) across all schools.	Schools	
Standards 1-3 teachers			
Head teacher			
INCO			

Facilitators of Kiswahili school-based training sessions attended at least one Kiswahili training session away from school (mean % facilitators of Kiswahili sessions)	The average of the (number of facilitators of Kiswahili school-based training sessions who attended at least one Kiswahili training session away from school/all facilitators of Kiswahili training sessions, expressed as a percentage) across all schools.	Schools	
Facilitators of numeracy school-based training sessions attended at least one numeracy training session away from school (mean % facilitators of numeracy sessions)	The average of the (number of facilitators of numeracy school-based training sessions who attended at least one numeracy training session away from school/all facilitators of numeracy training sessions, expressed as a percentage) across all schools.	Schools	
Facilitators of GRP school-based training sessions attended at least one GRP training session away from school (mean % facilitators of GRP sessions)	The average of the (number of facilitators of GRP school-based training sessions who attended at least one GRP training session away from school/all facilitators of GRP training sessions, expressed as a percentage) across all schools.	Schools	
School based training was held on (mean % school-based training sessions in a school):	The average of the (number of school-based training sessions that took place on x time of the day/all training sessions held by a school, expressed as a percentage) across all schools.	Schools	
On school days during teaching hours			
On school days after teaching hours			
Outside schools days			
Equal balance of school days and outside school days			
Records are available for a school-based training session (mean % school-based training sessions in a school):	The average of the (number of school-based training sessions for which records were available/all training sessions held by a school, expressed as a percentage) across all schools.	Schools	
Minutes are available for a school-based training session (mean % school-based training sessions in a school):	The average of the (number of school-based training sessions for which minutes were available/all training sessions held by a school, expressed as a percentage) across all schools.	Schools	
Experience as INCO at current school (mean years)	The average number of years INCO has been in the post at the current school.	INCOs	
INCO has been in post since Jan 2015 or earlier (% INCOs)	The number of INCOs who have been in post since Jan 2015 or earlier/all interviewed INCOs, expressed as a percentage.	INCOs	
Number of teachers who held the INCO post before current INCO was in post (% schools where current INCO has not been in post since Jan 2015)	The number of schools that had none, one or more than one teacher responsible for coordinating in-service training before current INCO was in post/all schools where current INCO has not been in post since January 2015, expressed as a percentage.	Schools	
None			
One			
More than one			
INCO is female (% INCOs)	Number of INCOs that are female/all interviewed INCOs, expressed as a percentage.	INCOs	
Age of INCO (mean years)	Average INCO age in years.	INCOs	
Time INCO working as a teacher (mean years)	The average number of years INCO has worked as a teacher.	INCOs	
Time INCO teaching at current school (mean years)	The average number of years INCO has been working at the current school.	INCOs	
INCO's highest professional education qualification (% INCOs):	The number of INCOs whose highest professional qualification is x/all interviewed INCOs, expressed as a percentage.	INCOs	
Bachelors of Education or higher			

Diploma or advanced diploma			
Certificate in education			
Other professional qualification			
No professional qualification			
INCO's highest academic qualification apart from professional education qualification (% INCOs):			
Primary school	The number of INCOs whose highest academic qualification (apart from their professional education qualification) is x/all interviewed INCOs, expressed as a percentage.	INCOs	
Form 4			
Form 6			
Certificate			
Diploma or advanced diploma			
Bachelors or higher			
Other			
INCO teaches maths this school term (% INCOs)	The number of INCOs reporting they are teaching maths to any standard this school term/all interviewed INCOs, expressed as a percentage.	INCOs	
INCO teaches Kiswahili this school term (% INCOs)	The number of INCOs reporting they are teaching Kiswahili to any standard this school term/all interviewed INCOs, expressed as a percentage.	INCOs	
INCO teaches maths or Kiswahili this school term (% INCOs)	The number of INCOs reporting they are teaching maths or Kiswahili to any standard this school term/all interviewed INCOs, expressed as a percentage.	INCOs	
INCO teaches Standards 1-3 this school term (% INCOs)	The number of INCOs reporting they are teaching Standards 1-3 this school term/all interviewed INCOs, expressed as a percentage.	INCOs	
INCO's number of teaching periods per week (mean)	The average number of teaching periods INCO has per week in this term	INCOs	
INCO holds other positions at school (% INCOs):			
Head teacher	The number of INCOs who also hold job x/all interviewed INCOs, expressed as a percentage.	INCOs	
Assistant head teacher			
Academic master			
INCO received EQUIP-T training away from school on numeracy (% INCOs)	The number of INCOs who have attended at least one training session from EQUIP-T away from school on numeracy/all interviewed INCOs, expressed as a percentage.	INCOs	
INCO attended all EQUIP-T training away from school on numeracy (% INCOs)	The number of INCOs who have attended all training from EQUIP-T away from school on numeracy/all interviewed INCOs, expressed as a percentage.	INCOs	
INCO received EQUIP-T training away from school on Kiswahili literacy (% INCOs)	The number of INCOs who have attended at least one training session from EQUIP-T away from school on Kiswahili literacy/all interviewed INCOs, expressed as a percentage.	INCOs	
INCO attended all EQUIP-T training away from school on Kiswahili literacy (% INCOs)	The number of INCOs who have attended all training from EQUIP-T away from school on Kiswahili literacy/all interviewed INCOs, expressed as a percentage.	INCOs	
INCO attended all EQUIP-T training away from school on GRP (% INCOs)	The number of INCOs who have attended all training from EQUIP-T away from school on GRP/all interviewed INCOs, expressed as a percentage.	INCOs	
Are teacher COL structures operating? (EQUIP-T output)			

Teachers attended ward cluster reflection meetings in 2016 and 2017 (% schools)	The number of schools where at least one teacher attended any meeting with teachers from other schools in the ward to reflect on in-service training modules or teaching practices in 2016 and 2017/all schools, expressed as a percentage.	Schools	
Number of days teachers attended ward cluster reflection meetings in 2016 and 2017 (mean days)	The average number of days on which teachers from the school attended ward cluster meetings in 2016 and 2017	Schools	
Number of hours ward cluster meetings usually took in 2016 and 2017 (mean hours)	The average number of hours ward cluster meetings usually lasted in 2016 and 2017	Schools	
Main topic of discussion at last ward cluster reflection meeting (% schools):	The number of schools where the main topic of discussion of the last ward cluster meeting was x/all schools that participated in a meeting in 2016-17, expressed as a percentage.	Schools	
Kiswahili			
Numeracy			
Gender responsive pedagogy			
Particular curriculum competency			
3R			
Other			
Time of the day when the ward cluster reflection meetings were usually held (% schools):	The number of schools where ward cluster meetings attended were held during x time of the day/all schools that participated in a meeting in 2016-17, expressed as a percentage.	Schools	
On school days during teaching hours			
On school days after teaching hours			
Outside school days			
Equal balance of both			
Attended at least one SPMM in the last 60 days (% Standards 1-3 teachers)	Number of Standards 1-3 teachers that attended at least one SPMM in the last 60 days/all interviewed Standards 1-3 teachers, expressed as a percentage.	Standards 1-3 teachers	SPMMs are intended to be held weekly, chaired by a teacher, and attended by HTs and other teachers, to discuss teaching and learning performance.
Attended four or more SPMMs in the last 60 days (% Standards 1-3 teachers)	Number of Standards 1-3 teachers that attended four or more SPMMs in the last 60 days/all interviewed Standards 1-3 teachers, expressed as a percentage.	Standards 1-3 teachers	
School held four or more SPMMs in the last 60 days (% schools)	Number of schools that held four or more SPMMs in the last 60 days/all schools, expressed as a percentage.	Schools	
Discussion at most recent SPMM was on teaching, learning or teacher/pupil attendance (% schools)	Number of schools where discussion at most recent SPMM (main or partial) was on teaching, learning or teacher/pupil attendance/all schools, expressed as a percentage.	Schools	Where minutes of the most recent SPMM were available, interviewers reviewed the minutes with the head teacher to answer this question. If minutes were not available, then a verbal response was accepted.
Has teacher capacity and confidence improved? (EQUIP-T output)			
Confidence in teaching new Standards 1-2 curriculum (% Standards 1-2 teachers):	The number of teachers of Standards 1-2 that report they feel very/fairly/not confident teaching the new Standards 1-2 curriculum/all interviewed teachers of Standards 1-2, expressed as a percentage.	Standards 1-2 teachers	
Very confident			
Fairly confident			
Not confident			

Class size (EQUIP-T output to intermediate outcome assumptions)			
Pupil enrolment by Standard (school mean)	The average number of pupils enrolled in Standard x in the current school year	Schools	
Pre-school			
Standard 1			
Standard 2			
Standard 3			
Standards 1-7			
Class size by Standard (mean)	The average number of pupils per class in Standard x in the current school year	Schools	
Pre-school			
Standard 1			
Standard 2			
Standard 3			
Standards 1-7			
Number of Standards 1-7 pupils per teacher (mean)	The average number of pupils (all Standards) per teacher in the current school year	Schools	
Pupils with a useable desk space (mean % pupils present during lessons)	The average of the (number of pupils with useable desk space/the total number of pupils present during the observed Standard 2 lesson, expressed as a percentage) across all lessons.	Standard 2 lessons observed	
Level of teacher turnover (EQUIP-T output to intermediate outcome assumptions)			
Teacher no longer at school by the next survey round (% all Standards 1-3 teachers)	The number of teachers of Standards 1-3 that were present at midline (baseline) but not at endline (midline)/all teachers of Standards 1-3 present at midline (baseline), expressed as a percentage.	Standards 1-3 teachers	
Teacher no longer at school by the next survey round (% all Standards 1-7 teachers)	The number of teachers of Standards 1-7 that were present at midline (baseline) but not at endline (midline)/all teachers of Standards 1-7 present at midline (baseline), expressed as a percentage.	Standards 1-7 teachers	
Baseline teacher no longer at school by endline (% all baseline Standards 1-3 teachers)	The number of teachers of Standards 1-3 that were present at baseline but not at endline/all teachers of Standards 1-3 present at baseline, expressed as a percentage.	Standards 1-3 teachers	
Baseline teacher no longer at school by endline (% all baseline Standards 1-7 teachers)	The number of teachers of Standards 1-7 that were present at baseline but not at endline/all teachers of Standards 1-7 present at baseline, expressed as a percentage.	Standards 1-7 teachers	
Teacher still teaching Standards 1-3 in same school two years later (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 that are present at the same school at midline (baseline) and endline (midline) and still teaching Standards 1-3/all teachers of Standards 1-3 present at midline (baseline), expressed as a percentage.	Standards 1-3 teachers	
Baseline teacher still teaching Standards 1-3 in same school at endline (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 that are present at the same school at baseline and endline and still teaching Standards 1-3/all teachers of Standards 1-3 present at baseline, expressed as a percentage.	Standards 1-3 teachers	
Reason for leaving for teachers who are no longer at the school by the next survey round (% Standards 1-7 teachers who left):	The number of former teachers reported by the head teacher to have left school by the next survey round for reason x/all former teachers, expressed as a percentage.	Former Standards 1-7 teachers	This is head teachers reporting on former teachers.
Transferred to another school			

Disciplinary issue			
Quit job			
Retired			
Passed away			
Long term sick			
Maternity leave			
Went for studies			
Other			
Approaching retirement age 60 (% Standards 1-3 teachers)	The number of Standards 1-3 teachers who are 59 or 60 years old/all interviewed Standards 1-3 teachers, expressed as a percentage.	Standards 1-3 teachers	
Teacher joined school since the previous round (% Standards 1-7 teachers)	The number of Standards 1-7 teachers who joined the school since the previous survey round/all Standards 1-7 teachers, expressed as a percentage.	Standards 1-7 teachers	
Teacher joined school since the previous round (% Standards 1-3 teachers)	The number of Standards 1-3 teachers who joined the school since the previous survey round/all interviewed Standards 1-3 teachers, expressed as a percentage.	Standards 1-3 teachers	
Previous job before becoming a teacher at current school (% Standards 1-3 teachers who joined school since last round)	The number of teachers of Standards 1-3 reporting previous job was x/all interviewed teachers of Standards 1-3 who joined the school since the previous survey round, expressed as a percentage.	Standards 1-3 teachers	
Teacher in another school			
None or other job not in teaching			
Location of previous posting for teachers who joined the school since the last survey round (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 reporting previous teaching job in location x/all interviewed teachers of Standards 1-3 who joined the school since the previous survey round and were teaching before, expressed as a percentage.	Standards 1-3 teachers	
Another school in same district			
Another school in same region			
A school in another region			
Teacher job satisfaction and motivation (EQUIP-T output to intermediate outcome assumptions)			
Teacher job satisfaction (mean rating)	Mean of self-reported ratings of Standards 1-3 teachers' job satisfaction on the day of the survey.	Standards 1-3 teachers	The rating scale is from one to ten, where 1 indicates 'completely unsatisfied' and ten indicates 'completely satisfied.'
Community appreciation of teachers' role (mean rating)	Mean of Standards 1-3 teachers' ratings of how valued they feel by the community on the day of the survey.	Standards 1-3 teachers	
Head teacher appreciation of teachers' role (mean rating)	Mean of Standards 1-3 teachers' ratings of how much they feel their head teacher values their role as a teacher on the day of the survey	Standards 1-3 teachers	
Reported teacher job satisfaction compared to two years ago (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 who reported feeling more satisfied/less satisfied/similarly satisfied with their job today than two years ago/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teachers	
More satisfied			
Less satisfied			
Similarly satisfied			
Time to school (mean minutes)	The average number of minutes it takes Standards 1-3 teachers to travel from home to school each morning.	Standards 1-3 teachers	

Teacher has outstanding non-salary claims (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who reported having outstanding non-salary claims/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teachers	A non-salary claim is an allowance that is due to a teacher for a variety of reasons including: leave, studies, INSET, medical, transfer, new employment, retirement, subsistence, travel abroad, disturbance, funeral, transport, and head of department.
Has the use of inclusive and gender-responsive teaching practices in the classroom increased? (EQUIP-T intermediate outcome)			
Teacher interactions with pupils are (% lessons observed):	The number of lessons where teachers' interaction with pupils is gender balanced/more with boys/more with girls/all Standards 2 lessons observed, expressed as a percentage.	Standard 2 lessons observed	Collection of information: Enumerators observed the entire duration of each lesson and recorded which pupils' teachers interacted with, noting if the pupil was a boy or girl, and how many boys and girls respectively were present. A classroom gender map was completed for each subject. Indicator construction: First, teacher interactions with girls as a proportion of total teacher interactions with all pupils is computed. Second, the proportion of girls present in the classroom is computed. Teacher interaction is considered gender balanced if the difference between the proportion of interactions with girls and the proportion of girls present during the lesson is smaller than 10 percentage points.
Gender balanced			
More with boys			
More with girls			
Teacher uses examples that challenge gender stereotyping (% lessons observed)	The number of observed Standard 2 lessons where teachers use examples that challenge gender stereotyping/all Standard 2 lesson observations, expressed as a percentage.	Standard 2 lessons observed	
Teacher engaged with at least one pupil from all six areas of the classroom (% lessons observed)	The number of Standard 2 lessons where teacher engaged with at least one pupil from all six areas in the classroom/all Standard 2 lessons observed, expressed as a percentage.	Standard 2 lessons observed	Collection of information: A classroom mapping instrument that divides the classroom into six approximately equally-sized areas was used by enumerators to record the number of interactions between teachers and pupils across the six classroom areas.
Distribution of teacher-pupil interactions (mean % all interactions):	The average of the (number of teacher interactions with pupils in the front two/middle two/back two areas of the classroom/all observed interactions, expressed as a percentage) across all lessons.	Standard 2 lessons observed	
Front two areas			
Middle two areas			
Back two areas			
Have teaching and assessment practices improved? (EQUIP-T intermediate outcome)			
Practices demonstrated by teachers during the introductory stage of a lesson (% lessons observed):	The number of observed Standard 2 lessons where teachers display teaching practice x fully or partly during the lesson introductory stages/all Standard 2 lesson observations, expressed as a percentage.	Standard 2 lessons observed	For each teaching practice enumerators recorded responses as follows: 'no' if they did not observe the practice, 'partly' if they observed some of parts of the practice and 'yes' if they observed all required aspects of the practice.
States objectives of lesson			
States new skills to be acquired			
Checks prior knowledge			
Practices demonstrated by teachers during the concluding stage of a lesson (% lessons observed):	The number of observed Standard 2 lessons where teachers display teaching practice x fully or partly during the lesson concluding stages/all Standard 2 lesson observations, expressed as a percentage.	Standard 2 lessons observed	

Checks pupils have acquired new skills or knowledge			
Holds a plenary to summarise and extend learning			
Practices demonstrated by teachers during the middle stages of a lesson (% lessons observed):			
Pupils demonstrate in front of class			
Teachers asks open ended questions			
Teacher probes pupil answers			
Teacher encourages pupil questions			
Teacher gives feedback on pupil work			
Teacher uses paired or group work			
Teacher makes effective use of blackboard			
Uses different instructional materials			
Relates well with and praises pupils			
Teacher listened to individual pupils reading a list of sounds, words or paragraph during the lesson (% Kiswahili lessons):			
Yes, to most pupils			
Yes, to some pupils			
No			
Teacher demonstrates at least seven positive teaching practices (% lessons)	The number of observed Standard 2 lessons where teachers demonstrate at least seven out of 14 selected teaching practices/all Standard 2 lessons observed, expressed as a percentage.	Standard 2 lessons observed	For each teaching practice enumerators recorded responses as follows: 'no' if practice not observed, 'yes, infrequently' if practice partly observed and 'yes, frequently' if the practice was frequently observed.
Lesson plan available and seen (% lessons)	The number of observed Standard 2 lessons where teacher had a lesson plan available/all Standard 2 lessons observed, expressed as a percentage.	Standard 2 lessons observed	
Teacher reports assessing pupil academic progress during the last five days (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report assessing pupil academic progress during the last five days/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teacher	
Teacher shows evidence of any pupil assessment conducted in the past 5 days (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who can show marked examples of any pupil assessment conducted in the last five days/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teacher	
Teacher shows evidence of the following types of pupil assessment conducted in past 5 days (% Standards 1-3 teachers):			
Class exercise			
Written class tests			
Homework			
Other written assessment			
Oral evaluation			
Teacher shows evidence of (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 able to show marked example of x assigned during the last five days/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teacher	
		Standards 1-3 teacher	

Two or more types of pupil assessments conducted in the past 5 days	The number of teachers of Standards 1-3 able to show marked examples of x number of assessments assigned during the last five days/all interviewed teachers of Standards 1-3, expressed as a percentage.		
One type of pupil assessment conducted in the past 5 days			
No types of pupil assessments conducted in the past 5 days			
Reports individually on pupils' academic progress to their parents (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report that they report individually on their pupils' academic progress to their parents/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teacher	
Number of times teacher reported on pupils' academic progress in the last year (mean)	The average number of times teachers reported to parents on pupil's academic progress in the last year, <i>reported by teachers</i>	Standards 1-3 teacher	
Number of times in 2017 HH received written info on pupils academic progress (mean)	The average number of times parents received written information from school about pupil's academic progress in 2017	Standard 3 pupils	
Number of times in 2017 HH met with teacher to discuss pupils academic progress (mean)	The average number of times parents met with a teacher to receive information about pupil's academic progress in 2017	Standard 3 pupils	
Have schools received EQUIP-T TLMs? (EQUIP-T input)			
Received supplementary since baseline (% schools)	The number of schools that received supplementary readers since baseline/all schools, expressed as a percentage.	Schools	
Received big books since baseline (% schools)	The number of schools that received big books since baseline/all schools, expressed as a percentage.	Schools	
Received read aloud books since baseline (% schools)	The number of schools that received teacher read aloud books since baseline/all schools, expressed as a percentage.	Schools	
Received teaching material/toolkits for Swahili literacy since baseline (% schools)	The number of schools that received teaching material/toolkits for Swahili literacy since baseline /all schools, expressed as a percentage.	Schools	
Received teaching material/toolkits for maths/numeracy since baseline (% schools)	The number of schools that received teaching material/toolkits for maths/numeracy since baseline/all schools, expressed as a percentage.	Schools	
Has the availability of TLMs in classrooms increased? (EQUIP-T output)			
Kiswahili supplementary readers available in classroom (% Standard 2 Kiswahili lessons)	The number of observed Standard 2 Kiswahili lessons where none, 1 to 20, 21 to 50, and more than 50 Kiswahili supplementary readers are available in the classroom/all observed Kiswahili Standard 2 lessons in schools that received the readers, expressed as a percentage.	Standard 2 Kiswahili lessons observed	
None			
1 to 20			
21 to 50			
More than 50			
Teaching and learning materials displayed on walls (% lessons)	The number of observed Standard 2 lessons where TLMs were displayed on walls/all observed Standard 2 lesson, expressed as a percentage.	Standard 2 lessons observed	
Pupils had a pencil during lesson (mean % pupils in a lesson)	The average of the (number of pupils in observed Standard 2 lessons that had a pencil during the lesson/all pupils present during lessons, expressed as a percentage) across all lessons.	Standard 2 lessons observed	
Pupils had a maths exercise book (mean % pupils in math lessons)	The average of the (number of pupils in observed Standard 2 maths lessons that had a maths exercise book during the lesson/all pupils present during maths lessons, expressed as a percentage) across all maths lessons.	Standard 2 maths lessons observed	
Pupils had a Kiswahili exercise book (mean % pupils in Kiswahili lessons)	The average of the (number of pupils in observed Standard 2 Kiswahili lessons that had a Kiswahili exercise book during the lesson/all pupils present during Kiswahili lessons, expressed as a percentage) across all Kiswahili lessons.	Standard 2 Kiswahili lessons observed	
Number of pupils per maths textbook in use (mean)	The average number of pupils per maths textbook being used in maths lessons	Standard 2 maths lessons observed	

Number of pupils per Kiswahili textbook in use (mean)	The average number of pupils per Kiswahili textbook being used in Kiswahili lessons	Standard 2 Kiswahili lessons observed	
Teacher has access to Standards 1 and 2 curriculum (% Standards 1-2 teachers):	The number of Standards 1-2 teachers who have good/limited/no access to the Standards 1-2 curriculum/all interviewed Standards 1-2 teachers, expressed as a percentage.	Standards 1-2 teacher	
Yes, good access			
Yes, limited access			
No access	The number of Standard 1 teachers who have good/limited/no access to the syllabi for Standard 1/all interviewed Standard 1 teachers, expressed as a percentage.	Standard 1 teacher	
Teacher has access to syllabi for Standard 1 (% Standard 1 teachers):			
Yes, good access			
Yes, limited access	The number of Standard 2 teachers who have good/limited/no access to the syllabi for Standard 2/all interviewed Standard 2 teachers, expressed as a percentage.	Standard 2 teacher	
No access			
Teacher has access to syllabi for Standard 2 (% Standard 2 teachers):			
Yes, good access	The number of Standards 1-2 Kiswahili teachers who have good/limited/no access to the teachers' guide for reading/all interviewed Standards 1-2 Kiswahili teachers, expressed as a percentage.	Standards 1-2 Kiswahili teacher	
Yes, limited access			
No access			
Teacher has access to teachers' guide for reading (% Standards 1-2 Kiswahili teachers):	The number of Standards 1-2 Kiswahili teachers who have good/limited/no access to the teachers' guide for writing/all interviewed Standards 1-2 Kiswahili teachers, expressed as a percentage.	Standards 1-2 Kiswahili teacher	
Yes, good access			
Yes, limited access			
No access	The number of Standards 1-2 maths teachers who have good/limited/no access to the teachers' guide for arithmetic/all interviewed Standards 1-2 maths teachers, expressed as a percentage.	Standards 1-2 maths teacher	
Teacher has access to teachers' guide for writing (% Standards 1-2 Kiswahili teachers):			
Yes, good access			
Yes, limited access	The number of Standards 1 maths teachers who have good/limited/no access to maths textbooks for the majority of Standard 1 pupils/all interviewed Standard 1 maths teachers, expressed as a percentage.	Standard 1 maths teacher	
No access			
Teacher has access to teachers' guide for arithmetic (% Standards 1-2 maths teachers):			
Yes, good access	The number of Standards 1 maths teachers who have good/limited/no access to maths textbooks for the majority of Standard 1 pupils/all interviewed Standard 1 maths teachers, expressed as a percentage.	Standard 1 maths teacher	
Yes, limited access			
No access			
Teacher has access to maths textbooks for the majority of Standard 1 pupils (% Standard 1 maths teachers) :	The number of Standards 1 maths teachers who have good/limited/no access to maths textbooks for the majority of Standard 1 pupils/all interviewed Standard 1 maths teachers, expressed as a percentage.	Standard 1 maths teacher	
Yes, good access			
Yes, limited access			
No access			

Teacher has access to maths textbooks for the majority of Standard 2 pupils (% Standard 2 maths teachers) :	The number of Standards 2 maths teachers who have good/limited/no access to maths textbooks for the majority of Standard 2 pupils/all interviewed Standard 2 maths teachers, expressed as a percentage.	Standard 2 maths teacher	
Yes, good access			
Yes, limited access			
No access			
Teacher has access to Kiswahili textbooks for the majority of Standard 1 pupils (% Standard 1 Kiswahili teachers) :	The number of Standards 1 Kiswahili teachers who have good/limited/no access to Kiswahili textbooks for the majority of Standard 1 pupils/all interviewed Standard 1 Kiswahili teachers, expressed as a percentage.	Standard 1 Kiswahili teacher	
Yes, good access			
Yes, limited access			
No access			
Teacher has access to Kiswahili textbooks for the majority of Standard 2 pupils (% Standard 2 Kiswahili teachers) :	The number of Standards 2 Kiswahili teachers who have good/limited/no access to Kiswahili textbooks for the majority of Standard 2 pupils/all interviewed Standard 2 Kiswahili teachers, expressed as a percentage.	Standard 2 Kiswahili teacher	
Yes, good access			
Yes, limited access			
No access			
Has the use of TLMs in classrooms increased? (EQUIP-T intermediate outcome)			
Teacher uses big books or Teacher Read Aloud books (% Standard 2 Kiswahili lessons)	The number of Standard 2 Kiswahili lessons where teacher used big books or teacher read aloud books during the lesson/all observed Kiswahili Standard 2 lessons in schools that received the books, expressed as a percentage.	Standard 2 Kiswahili lessons observed	
Pupils used maths learning materials during lessons (% Standard 2 maths lessons)	The number of Standard 2 maths lessons where most/some/no pupils used maths learning materials besides textbooks/all observed maths Standard 2 lessons in schools that received the materials, expressed as a percentage.	Standard 2 maths lessons observed	
Yes, most pupils			
Yes, some pupils			
No			
Pupils read supplementary readers during lessons (% Standard 2 Kiswahili lessons)	The number of Standard 2 Kiswahili lessons where most, some or no pupils read supplementary readers to themselves or out-loud/all observed Kiswahili Standard 2 lessons in classrooms with available readers, expressed as a percentage.	Standard 2 Kiswahili lessons observed	
Yes, most pupils			
Yes, some pupils			
No			
Pupils used textbooks during the lesson (% lessons)	The number of Standard 2 lessons where pupils used textbooks during the lesson/all observed Standard 2 lessons, expressed as a percentage.	Standard 2 lessons observed	
Teacher absence from school and classrooms, and punctuality (EQUIP-T output to intermediate outcome assumptions)			
On the day of the survey, of all teachers in the roster: Absent from school (%)	The number of teachers who were not present for the teacher head count on the day of the survey/all teachers working at the school, expressed as a percentage.	All teachers in schools' teacher roster	Collection of information: The school and classroom absenteeism measures rely on two different headcounts of teachers carried out by enumerators. At the start of the day of the school visit, enumerators first recorded teachers who were present at school and second, during
Of teachers present on the day of the survey and timetabled to teach: Absent from class (%)	The number of teachers who were not present at their timetabled lesson before lunch despite being in school and timetabled to teach/all teachers present on the day of the survey and timetabled to teach the lesson before lunch, expressed as a percentage.	Teachers in schools' teacher roster who were scheduled to teach	

		before lunch and present at school on the day of the survey	the lesson before lunch, recorded if teachers were in classrooms teaching. In the head teacher instrument we record whether each teacher in the roster is timetabled to teach in the period before lunch.
Of teachers present on the day of the survey: Arrived late (%)	The number of teachers who arrived after the school is officially supposed to start/all teachers present on the day of the survey, expressed as a percentage.	All teachers present in school on day of survey	Classroom absenteeism was measured during the lesson before lunch because it is a 'typical' lesson time to make the observation that was the same across all surveyed schools, but that avoided the start of the day so that classroom absenteeism was not confounded with lateness.
Reasons for classroom absenteeism for teachers who reported being absent from class the last 30 days (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 who reported being absent from class in the last 30 days and reported reason x/all interviewed teachers of Standards 1-3 who reported being absent from class during the last 30 days, expressed as a percentage.	Standards 1-3 teachers	
Large workload			
Meeting with head teacher			
Meeting with teachers			
Lack of motivation			
Illness			
Feeling tired			
Other			
Has instructional time increased? (EQUIP-T intermediate outcome)			
Actual weekly timetabled minutes for mathematics in Standards 1 and 2 (before adjustment).	Minutes per week timetabled for mathematics in Standards 1 and 2 (school mean).	Schools	Data on timetables for each class in Standards 1 and 2 were used to identify how many periods by subject are timetabled each week. For each class in a standard, the total number of weekly periods assigned for mathematics and Kiswahili were multiplied by the number of minutes assigned to each period to calculate total weekly minutes in each subject at the class level. These totals were then averaged across the number of classes to get the number of minutes timetabled for each subject by standard. Finally, the weekly minutes were averaged across standards one and two to get the number of weekly minutes timetabled for each subject by school. To estimate to what extent available instructional time is reduced by classroom absenteeism, indicators on weekly minutes timetabled were adjusted for whether teachers were present in a classroom.
Actual weekly timetabled minutes for mathematics in Standards 1 and 2 after adjusting for the classroom absenteeism rate of Standards 1 and 2 teachers	The minutes per week timetabled for mathematics in Standards 1 and 2 after adjusting for the classroom absenteeism rate of Standards 1 and 2 teachers (school mean).	Schools	
Actual weekly timetabled minutes for Kiswahili in Standards 1 and 2 (before adjustment).	Minutes per week timetabled for Kiswahili in Standards 1 and 2 (school mean).	Schools	
Actual weekly timetabled minutes for Kiswahili in Standards 1 and 2 after adjusting for the classroom absenteeism rate of Standards 1 and 2 teachers	The minutes per week timetabled for Kiswahili in Standards 1 and 2 after adjusting for the classroom absenteeism rate of Standards 1 and 2 teachers (school mean).	Schools	

			This is a rough estimate of actual instructional time.
Actual weekly minutes for mathematics before adjustment meets requirements (% schools)	The number of schools where the timetabled weekly minutes for mathematics for Standards 1 and 2 meets the official instructional time	Schools	At baseline, the official instructional time for mathematics was 210 minutes per week; at midline and endline it is 240 minutes per week
Actual weekly minutes for Kiswahili before adjustment meets requirements (% schools)	The number of schools where the timetabled weekly minutes for mathematics for Standards 1 and 2 meets the official instructional time	Schools	At baseline, the official instructional time for Kiswahili was 180 minutes per week; at midline and endline it is 480 minutes per week
Early grade teacher background characteristics			
Female (% Standards 1-3 teachers)	Number of Standards 1-3 teachers that are female/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teachers	
Age (mean years)	Average age of Standards 1-3 teachers in years.	Standards 1-3 teachers	
Time working as a teacher (mean years)	The average number of years Standards 1-3 teachers have worked as a teacher.	Standards 1-3 teachers	
Time teaching at current school (mean years)	The average number of years Standards 1-3 teachers have been working at the current school.	Standards 1-3 teachers	
Highest professional education qualification (% Standards 1-3 teachers):			
Bachelors of Education or higher	The number of Standards 1-3 teachers whose highest professional qualification is x/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teachers	
Diploma or advanced diploma			
Certificate in education			
Other professional qualification			
No professional qualification			
Highest academic qualification apart from professional education qualification (% Standards 1-3 teachers):			
Primary school	The number of Standards 1-3 teachers whose highest academic qualification (apart from their professional education qualification) is x/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standards 1-3 teachers	Due to a change in the administration of this question, the midline data is not comparable to the baseline and endline data. The data from the baseline and endline survey rounds are comparable.
Form 4			
Form 6			
Certificate			
Diploma or advanced diploma			
Bachelors or higher			
Other			

F.3 Chapter 5 School leadership and management

Indicator name	Indicator definition	Respondent / unit of analysis	Notes
Head teacher background characteristics			
Female (% head teachers)	Number of female head teachers/all head teachers, expressed as a percentage.	Head teachers	
Age (mean years)	Average head teacher age in years.	Head teachers	
Time working as a head teacher (mean years)	The average number of years head teachers have worked as a head teacher.	Head teachers	
Time working as a head teacher at current school (mean years)	The average number of years head teachers have worked as a head teacher at their current school.	Head teachers	
Highest professional qualification (% head teachers):	The number of head teachers whose highest professional qualification is x/all head teachers, expressed as a percentage.	Head teachers	
Bachelors of Education or higher			
Diploma or advanced diploma			
Certificate in education			
Other professional qualification			
No professional qualification	The number of head teachers whose highest academic qualification, apart from the professional education qualification, is x/all head teachers, expressed as a percentage.	Head teachers	Due to a change in the administration of this question, the midline data is not comparable to the baseline and endline data. The data from the baseline and endline survey rounds are comparable.
Highest academic qualification, apart from the professional education qualification (% head teachers):			
Primary school			
Form 4			
Form 6			
Certificate			
Diploma or advanced diploma			
Bachelor's degree or higher			
Other			
Level of head teacher turnover (EQUIP-T input to output and output to intermediate outcome assumption)			
Head teacher was head teacher at same school at baseline and endline (% head teachers)	The number of head teachers who were head teachers at the same school at baseline and endline/all head teachers, expressed as a percentage	Head teachers	
Head teacher was head teacher at same school since the last survey round (% head teachers)	The number of head teachers who were head teachers at the same school since the last survey round/all head teachers, expressed as a percentage	Head teachers	
Reasons for head teacher turnover (% head teachers who are no longer head teachers in the same school by the next survey round):	The number of head teachers who are no longer head teachers in the same school by the next survey round for reason x/all head teachers who are no longer head	Head teachers	

Left the school	teachers in the same school by the next survey round, expressed as a percentage.		
Demoted within the same school			
Reasons for head teachers who left the school (% head teachers who left the school by the next survey round):	The number of head teachers who left the school for reason x/all head teachers who left the school by the next survey round, expressed as a percentage.	Head teachers	
Transferred			
Retired			
Passed away			
Studies			
On secondment			
Disciplinary issue			
Other			
Head teacher has been head teacher at current school for less than 2 years (% head teachers)	The number of head teachers who had been in the head teacher post at their current school for less than 2 years/all head teachers, expressed as a percentage.	Head teachers	
Job before becoming head teacher at this school (% head teachers who had been HTs less than two years)	The number of head teachers who were doing job x before becoming head teacher at current school/all head teachers who had been head teachers at current school for less than two years, expressed as a percentage.	Head teachers	
Head teacher			
Teacher			
Other job in education			
Location of previous job (% head teachers who had been head teachers at current school for less than two years):	The number of head teachers who used to work at a school in location x/all head teachers who had been head teachers at current school for less than two years, expressed as a percentage.	Head teachers	
This school			
Another school in this district			
Another school in this region			
Another school in another region			
Has EQUIP-T provided SLM in-service training for head teachers? (EQUIP-T input)			
Attended SLM in-service training last two years (% head teachers)	Number of head teachers that reported attending any SLM in-service training the previous two years/all interviewed head teachers, expressed as a percentage.	Head teachers	The relevant period for BL is 2012-2013; for ML 2014-2015; and for EL 2016-2017.
Attended in-service SLM training provided by (% head teachers):	Number of head teachers that reported attending in-service SLM training from provider x the previous two years/all interviewed head teachers, expressed as a percentage.	Head teachers	
EQUIP-T			
LANES			
BRN			
STEP			
Other			

Head teacher received EQUIP-T training on following content since baseline (% head teachers):	The number of head teachers reporting receiving training on content x since baseline/ all interviewed head teachers, expressed as a percentage	Head teachers	
School leadership / HT role / school standards			
SDPs			
PTP grant 1 (application / management)			
PTP grant 2 (application / management)			
Reporting and record keeping			
SIS			
SCs			
PTP roles			
School performance management meetings			
Business plans and IGAs			
Pupil welfare / JUU clubs			
Duration of SLM training in last two years (mean days)	Average number of total days of SLM training head teacher attended in last two years	Head teachers	
Head teacher's view of EQUIP-T SLM training (% head teachers):	The number of head teachers reporting that they found the EQUIP-T SLM training useful/somewhat useful/not useful/all interviewed head teachers who attended EQUIP-T SLM training, expressed as a percentage.	Head teachers	
Useful			
Somewhat useful			
Not useful			
Gains from EQUIP-T SLM training (% head teachers):	The number of head teachers reporting gain x from the EQUIP-T SLM training/all interviewed head teachers who attended EQUIP-T SLM training and thought it was (somewhat) useful, expressed as a percentage.	Head teachers	<p>Gains means skills, knowledge or behaviour change mentioned by respondents in any of the areas listed.</p> <p>Pupil welfare relates to any policies, actions or clubs that promote the welfare of pupils in the school and that are not related to academic or extra-curricular activities (these are typically sport, music). This covers health, safety (including child protection issues), well-being (including positive learning environment e.g. positive discipline rather than corporal punishment, appropriate roles/responsibilities/behaviour for teachers and pupils, planting trees so pupils have shade, menstruation support for girls, anti-female genital mutilation, rights to education campaigns, counselling, etc.).</p> <p>At endline, there were some additions and changes to the categories listed at baseline and midline.</p>
Head teacher responsibilities			
Teacher management			
Financial management			
School development planning			
Reporting/record keeping			
Academic programme management			
Confidence in role as head teacher			
Support network			
Relationship with teachers			
Relationships with parents/community			
School committee			
Pupil welfare			

Other						
Difficulties head teachers experienced with EQUIP-T SLM training (% head teachers):						
None						
Not relevant to my job						
Materials difficult						
Too much content						
Too theoretical						
Took too much time/work load						
Limited training time	The number of head teachers reporting difficulty x with the EQUIP-T SLM training/all interviewed head teachers who attended EQUIP-T SLM training and thought it was (somewhat) useful, expressed as a percentage.	Head teachers	At endline, there were some additions and changes to the categories listed at baseline and midline.			
Time lag between training events						
Sessions inconvenient time/day						
Transport difficult / venue too far						
No/insufficient payment						
No/insufficient direct training						
Envy from colleagues						
Not enough training material						
Content not completed						
Problems with trainers						
Other						
Head teacher incurred out of pocket expenses for attending EQUIP-T SLM training away from school (% head teachers)				Number of head teachers reporting that they had to make an out of pocket payment/all interviewed head teachers who attended EQUIP-T SLM training away from school, expressed as a percentage	Head teachers	
Has EQUIP-T provided early grade teaching in-service training to head teachers? (EQUIP-T input)						
Attended Early Grade teaching in-service training last two years (% head teachers)				Number of head teachers that reported attending Early Grade teaching in-service training the previous two years/all interviewed head teachers, expressed as a percentage.	Head teachers	The relevant period for BL is 2012-2013; for ML 2014-2015; and for EL 2016-2017.
Attended in-service early grade training provided by (% head teachers):						
EQUIP-T	Number of head teachers that reported attending in-service Early Grade teaching training from provider x the previous two years/all interviewed head teachers, expressed as a percentage.	Head teachers				
LANES						
BRN						
STEP						
Other						

Main content of EQUIP-T early grade training last two years (% head teachers who attended the training):	The number of head teachers reporting content x/all interviewed head teachers who attended EQUIP-T early grade in-service training, expressed as a percentage	Head teachers	
Standards 1 and 2 curriculum			
Standards 3 and 4 curriculum			
EG Swahili literacy			
EG numeracy			
EG other subjects			
Upper grad subject			
General teaching methods			
Gender-responsive pedagogy			
Pre-school teaching			
Health/nutrition			
Other			
SLM			
Has head teacher capacity changed? (EQUIP-T output)			
Has SDP for current school year (% schools)	The number of head teachers reporting they have a school development plan (SDP) for the current school year/all interviewed head teachers, expressed as a percentage.	Head teachers	Head teachers were questioned about whether they had a SDP for year x. To check the reliability of this response, head teachers were asked to present this SDP to the interviewer.
SDP comprehensiveness (% schools):	The number of schools with a SDP that contains no/one/two/three of the core elements/all schools, expressed as a percentage.	Head teachers	The core elements are: (1) a budget, (2) teaching and learning objectives and (3) baseline data and targets.
Has SDP but it is not available			
SDP has none of the core elements			
SDP has one of the core elements			
SDP has two of the core elements			
SDP has three of the core elements			
SDP contents (% schools):	The number of schools with SDP that contains element x/all schools, expressed as a percentage.	Head teachers	
Improvements to school facilities			
Teaching and learning objectives			
Strategy to improve Standards 4 and 7 exam scores			
Strategy to reduce dropout or pupil absenteeism			
Strategy to improve girls' learning			
Strategy to improve transition to secondary school			
Budget			
Baseline data and targets			

Have head teachers' SLM practices changed? (EQUIP-T intermediate outcome)			
Implementation of current year's SDP started (% schools)	Number of schools where implementation has started on at least one activity from the current SDP/all schools, expressed as a percentage.		
Reported most common teacher performance management practices (% head teachers):	The number of head teachers reporting teacher performance management practice x as the most common/all interviewed head teachers, expressed as a percentage.	Head teachers	In the head teacher interview, only head teachers were asked this question, not assistant head teachers or academic masters answering on behalf of the head teacher if absent. But some non-interviewed head teachers were phoned for this information to reduce the number of missing observations.
Pupil academic results			
Lesson preparations			
Teaching performance in class			
Teacher punctuality and attendance			
Use of continuous pupil assessment			
Other			
Report lesson plans were checked by head teacher (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report that the head teacher checks their lesson plans/all interviewed teachers of Standards 1-3 expressed as a percentage.	Standard 1-3 teachers	
Report written lesson plan feedback from head teacher (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report receiving written lesson plan feedback from the head teacher/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standard 1-3 teachers	
Report lesson observation by head teacher (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report that the head teacher observes their teaching/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standard 1-3 teachers	
Report written lesson observation feedback from head teacher (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report receiving written lesson observation feedback from the head teacher/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standard 1-3 teachers	
Report lesson observation by others in last 30 days (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report that the academic master or other teachers observed their teaching in the last 30 days/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standard 1-3 teachers	
Report lesson observation by head teacher or others in last 30 days (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report that the head teacher, academic master or other teachers observed their teaching in the last 30 days/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standard 1-3 teachers	
Report receiving at least one performance appraisal in the previous school year (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 who report that the head teacher, assistant head teacher or academic master held at least one individual meeting with them to discuss their performance and professional development needs during the previous school year/all interviewed teachers of Standards 1-3, expressed as a percentage.	Standard 1-3 teachers	
Four or more staff meetings in the last 60 days (% Standards 1-3 teachers)	The number of Standards 1-3 teachers reporting that at least four staff meetings were held in the last 60 days/all interviewed Standards 1-3 teachers, expressed as a percentage.	Standards 1-3 teachers	Staff meetings are typically chaired by the HT, attended by teachers and (sometimes) non-teaching staff, to discuss administrative and other school matters.
Number of staff meetings held in last 60 days (mean days)	Average number of staff meetings held in last 60 days as reported by Standards 1-3 teachers	Standards 1-3 teachers	

Rewards for teachers who perform well exist (% head teachers)	The number of head teachers reporting that there are rewards in their school for teachers who perform well/all interviewed head teachers, expressed as a percentage.	Head teachers	In the head teacher interview, the CAPI instrument was designed only to ask teacher management related questions of <i>actual</i> head teachers (not academic masters or other persons answering on behalf of the head teacher). However, subsequent to the initial survey, head teachers were phoned for this information to reduce the number of missing responses.
Types of teacher performance rewards (% head teachers)	The number of head teachers reporting reward type x/all interviewed head teachers, expressed as a percentage.	Head teachers	
Financial			
Material (in-kind resources)			
Verbal recognition			
Certificate, cup or medal			
In-school promotion			
Trips or events			
Other			
Action is taken for teachers performing poorly (% head teachers)	The number of head teachers reporting that action is taken at their school for teachers who perform poorly/all interviewed head teachers, expressed as a percentage.	Head teachers	
Types of actions for poor teacher performance (% head teachers)	The number of head teachers reporting action type x/all interviewed head teachers, expressed as a percentage.	Head teachers	
Extra support to improve teaching			
Increased lesson observation			
Increased checking of lessons plans etc.			
Teachers required to give extra classes			
Warning from HT			
HT reports to WEO			
Warning from WEO			
Warning from SC			
Other			
Rewards for teachers who perform well exist (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 reporting that there are rewards in their school for teachers who perform well/all interviewed Standards 1-3 teachers, expressed as a percentage.	Teachers of standards 1-3	
Types of teacher performance rewards (% Standards 1-3 teachers)	The number of Standards 1-3 teachers reporting reward type x/all interviewed Standards 1-3 teachers, expressed as a percentage.	Teachers of standards 1-3	
Financial			
Material (in-kind resources)			
Verbal recognition			
Certificate, cup or medal			
In-school promotion			
Trips or events			

Other			
Action is taken for teachers performing poorly (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 reporting that action is taken for teachers who perform poorly/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	
Types of actions for poor teacher performance (% Standards 1-3 teachers)			
Extra support to improve teaching			
Increased lesson observation			
Increased checking of lessons plans etc.			
Teachers required to give extra classes	The number of teachers of Standards 1-3 reporting action type x/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	
Warning from HT			
HT reports to WEO			
Warning from WEO			
Warning from SC			
Other			
Head teacher considers teachers attendance at school 'good' or 'very good' (% head teachers)	The number of head teachers who consider that teachers' attendance at school is good or very good/all interviewed head teachers, expressed as a percentage.	Head teachers	
Teacher attendance today compared to two years ago (% head teachers who have been at school for at least 2 years):	The number of head teachers who consider that teachers' attendance at school compared to two years ago is better/worse/similar/all interviewed head teachers who had been at school for at least two years, expressed as a percentage.	Head teachers	
Better			
Worse			
Same			
Most common reasons teachers are absent from school (% head teachers):			
Illness			
Family reasons			
Attending training			
Official education work / meeting	The number of head teachers reporting that teachers are most commonly absent from school for reason x/all interviewed head teachers, expressed as a percentage.	Head teachers	In the head teacher interview, the CAPI instrument was designed only to ask teacher management related questions of <i>actual</i> head teachers (not academic masters or other persons answering on behalf of the head teacher). However, subsequent to the initial survey, head teachers were phoned for this information to reduce the number of missing responses.
Transport problem			
Collecting salary			
Other official government work			
Other private work			
Lack of motivation			
Alcoholism/drugs			

Other			
Head teacher reports that sometimes teachers are absent from classrooms (% head teachers)	The number of head teachers reporting that teachers are sometimes absent from classroom/all interviewed head teachers, expressed as a percentage.	Head teachers	
Most common reasons teachers absent from classrooms (% head teachers):			
Large work load			
Meeting with head teacher			
Meeting with other teachers			
Lack of motivation	The number of head teachers reporting that teachers are most commonly absent from classroom for reason x/all interviewed head teachers, expressed as a percentage.	Head teachers	
Illness			
Feeling tired/exhausted			
Other			
Head teacher took action to improve education in last school year (% head teachers)	The number of head teachers reporting taking action in last school year to improve education in the school/all interviewed head teachers who have been at school for at least a year, expressed as a percentage.	Head teachers	
Actions taken by head teacher in previous school year to improve education (% head teachers):			
Teacher attendance and punctuality			
Teacher in-service training / other teaching support			
Teaching and learning materials			
School infrastructure			
Relationship with parents / community			
Relationship with WEO / district			
Extra tuition classes	The number of head teachers reporting taking action x in last school year to improve education in the school/all interviewed head teachers who have been at school for at least a year, expressed as a percentage.	Head teachers	The relevant period for ML is 2015 and for EL 2017.
Pupil absenteeism			
School feeding			
IGAs			
Fundraising			
More tests/exams			
Pupil welfare			
Extra-curricular activities			
Other			
Head teacher took action to improve education in last school year (% Standards 1-3 teachers)	The number of Standards 1-3 teachers reporting that the head teacher took action in last school year to improve education in the school/all interviewed Standards 1-3	Teachers of standards 1-3	

	teachers who have been at school for at least a year, expressed as a percentage.		
Head teacher absence and job satisfaction (Equip-t output to intermediate outcome assumptions)			
Head teachers absent on day of survey using headcount observation (%)	The number of head teachers who were not present at the headcount on the day of the survey/all head teachers, expressed as a percentage.	Head count	A head count of all head teachers was conducted by enumerators on the day of the survey.
Head teacher reports being absent from school during the last 30 days (% head teachers)	The number of head teachers reporting being absent from school in the last 30 days/ all interviewed head teachers, expressed as a percentage.	Head teachers	
Reasons for school absenteeism in the last 30 days (% head teachers):	The number of head teachers who report being absent from school in the last 30 days for reason x/all interviewed head teachers, expressed as a percentage.	Head teachers	
Illness			
Family responsibility			
Attending training			
Official education work/meeting			
Transport problem			
Collecting salary			
Other official work			
Other private work			
Lack of motivation			
Alcoholism / drugs			
Other			
Head teacher has outstanding non-salary claims (% head teachers)	The number of head teachers who report having outstanding non-salary claims/all interviewed head teachers, expressed as a percentage.	Head teachers	A non-salary claim is an allowance that is due to a teacher for a variety of reasons including: leave, studies, INSET, medical, transfer, new employment, retirement, subsistence, travel abroad, disturbance, funeral, transport, and head of department.
Head teacher job satisfaction (mean rating)	Mean of self-reported ratings of head teachers' job satisfaction on the day of the survey.	Head teachers	The rating scale is from one to ten, where 1 indicates 'completely unsatisfied' and ten indicates 'completely satisfied.'
Compared to two years ago, head teacher job satisfaction is (% head teachers):	The number of head teachers reporting that their job satisfaction compared to two years ago is higher/lower/similar/all interviewed head teachers, expressed as a percentage.	Head teachers	
Higher			
Lower			
Similar			
Capitation grant payments received in full and in-kind resources received (Equip-t output to intermediate outcome assumption)			
Estimate of capitation grant payments per pupil received in previous school year (mean TZS)	The average amount of capitation grants received in previous school year per enrolled pupil	Head teachers	The relevant period for ML is 2015 and for EL 2017.
Received capitation grant in full in previous school year (% schools)	The number of schools that received the capitation grant in the previous school year in full/all schools, expressed as a percentage.	Head teachers	The relevant period for ML is 2015 and for EL 2017. The expected capitation grant is TZS 6,000 per pupil.

School received any in-kind resources in the last two school years (% schools)	The number of schools that received any in-kind resources from any provider in the last two school years/all schools, expressed as a percentage.	Head teachers	The relevant period for ML is 2014-2015 and for EL 2016-2017.
School received the following in-kind resources in the last two school years (% schools)			
Textbooks			
Pupil uniform including shoes			
Classroom furniture			
Classrooms	The number of schools that received resource x from any provider in the last two school years/all schools, expressed as a percentage.	Head teachers	The relevant period for ML is 2014-2015 and for EL 2016-2017.
Toilets/latrines			
Water			
Electricity			
Teacher housing			
School feeding			
Key school characteristics and infrastructure			
Pupils per classroom in use (school mean)	The average number of pupils (all Standards) per usable classroom in the current school year	Schools	
Schools with more than 60 pupils per classroom in use (% schools)	The number of schools with more than an average of 60 pupils per classroom in use/all schools, expressed as a percentage.	Schools	
School has second shift (% schools)	The number of schools that have a second shift/all schools, expressed as a percentage.	Schools	
Pupils per functional toilet (school mean)	The average number of pupils (all Standards) per functional toilet	Schools	
School has a source of drinking water on school premises (% schools)	The number of schools that have a source of drinking water on school premises/all schools, expressed as a percentage.	Schools	
School has functioning electricity on school premises (% schools)	The number of schools that have functioning electricity on school premises/all schools, expressed as a percentage.	Schools	
Has EQUIP-T trained head teachers on SIS and provided SIS tablets? (EQUIP-T input)			
Received tablet for school information system (SIS) in last two years (% schools)	Number of schools that received SIS tablet in the last two years/all schools, expressed as a percentage.	Schools	
School has a functioning SIS tablet (% schools)	Number of schools that report having a functioning SIS tablet/all schools, expressed as a percentage.	Schools	
SIS tablet seen by survey enumerator (% schools)	Number of schools where enumerator saw the SIS tablet/all schools, expressed as a percentage.	Schools	
Charged SIS tablet seen by survey enumerator (% schools)	Number of schools where enumerator saw the SIS tablet and observed it was charged/all schools, expressed as a percentage.	Schools	
Is the EQUIP-T SIS functional? (EQUIP-T output)			

SIS has up to date records of pupils enrolled and teachers employed (% schools with functional tablets)	Number of schools that report having completed records in their SIS tablet of all pupils enrolled and teachers employed at the school/all schools with a functional SIS tablet, expressed as a percentage.	Schools	
Actual mean hours it took to enter/update pupil and teacher data	Average number of hours it took to enter all the data on pupils enrolled and teachers employed at the school (for the schools that have up to date records)	Schools	
Reasons why records of pupil enrolment / current teachers not complete (% schools):	Number of schools that report reason x for incomplete records of pupil enrolment and teacher employment/all schools that have incomplete records, expressed as a percentage.	Schools	
Tablet not working properly			
Too much time			
No benefit			
Do not understand			
Did not receive training			
Other			
Pupil attendance for all classes recorded on at least one day in current school year (% of schools)	Number of schools where pupil attendance for all classes recorded on at least one day in the current school year/all schools with a functioning tablet, expressed as a percentage.	Schools	
Reasons pupil attendance data not recorded last week (% schools)	Number of schools that report reason x for not recording pupil attendance data/all schools with a functioning tablet, expressed as a percentage.	Schools	
Tablet not working properly			
Too much time			
Did not receive training			
Teacher attendance recorded in the tablet on at least one day in the current school year (% of schools)	Number of schools where teacher attendance recorded on at least one day in the current school year/all schools with a functioning tablet, expressed as a percentage.	Schools	
EQUIP-T SIS tablet software is fit for purpose and tablet-based SIS is appropriate (EQUIP-T input to output assumption)			
Difficulties in using SIS tablet (% schools)	Number of schools that report difficulty x with using SIS tablet/all schools with a functioning tablet, expressed as a percentage.	Schools	
None			
Poor internet connectivity			
Insufficient direct training on SIS/tablet			
Insufficient on-going support			
High work load/too much time to enter SIS data			
No feedback after submitting SIS data			
Electricity not reliable to regularly charge SIS tablet			
SIS data were lost			
Don't know how to use			

Does the SIS provide useful reports to support SLM? (EQUIP-T intermediate outcome)			
Head teacher used SIS for SLM or community engagement (% of head teachers)	Number of head teachers that have used the SIS for SLM tasks or for sharing information with the community in the current school year/all head teachers with a functioning tablet, expressed as a percentage.	Head teachers	
Head teacher used SIS for discussions with WEO (% of head teachers)	Number of head teachers that have used the SIS for discussions with WEO/all head teachers with a functioning tablet, expressed as a percentage.	Head teachers	
SIS tablet has replaced other written records/reports in the school (% schools)	Number of schools where SIS tablet has replaced other written records/reports in the school/all schools with a functioning tablet, expressed as a percentage.	Schools	
Head teachers' view of SIS (% head teachers)	The number of head teachers reporting that they found the SIS very/somewhat/not useful/all head teachers with a functioning tablet, expressed as a percentage.	Head teachers	
Very useful			
Somewhat useful			
Not useful			
Is WEO support for head teachers effective? (EQUIP-T output, component 3B)			
Schools visited by School Quality Assurers in the previous school year (% schools)	The number of head teachers who report being visited by School Quality Assurers in the previous school year/all interviewed head teachers, expressed as a percentage.	Head teachers	District Inspectors are now called School Quality Assurance Officers. The relevant period for BL is 2014, ML 2015; and EL 2017.
Number of SQA visits in previous school year (mean)	The average of the total number of visits by SQA to a school during the previous school year as reported by the head teacher (school mean for the schools that report a SQA visit in the previous year).	Head teachers	
Schools visited by WEO in the previous school year (% schools)	The number of head teachers who report being visited by WEO in the previous school year/all interviewed head teachers, expressed as a percentage.	Head teachers	Ward Education Coordinators are now called Ward Education Officers. The relevant period for BL is 2014, ML 2015; and EL 2017.
School received WEO visits in previous school year 12 or more times (% schools)	The number of head teachers who report being visited by WEO in the previous school year 12 or more times/all interviewed head teachers, expressed as a percentage.	Head teachers	
Duration of last WEO visit (% schools):	The number of head teachers who reported WEO stayed for length x during last visit/all interviewed head teachers, expressed as a percentage.	Head teachers	
30 minutes or less			
31-60 minutes			
61-120 minutes			
121-180 minutes			
More than 180 minutes			
Have WEO management practices changed? (EQUIP-T intermediate outcome)			
WEO activities during the last visit (% schools):	The number of head teachers who reported WEO conducted activity x during last visit/all interviewed head teachers, expressed as a percentage.	Head teachers	
Checked school records			
Checked teacher records			
Checked pupils' work			

Observed lessons			
Observed school facilities			
Observed school management practices			
Observed SC meeting			
Observed PTP meeting			
Attended school-based INSET			
Bringing/supervising exams			
Coaching/participating in sports			
Other			
Areas WEO advised on / supported during the last visit (% schools):			
WEO did not provide any advice / support during last visit			
Teaching and learning			
Teacher attendance/punctuality			
Pupil attendance/punctuality			
Pupil welfare			
Extra-curricular activities			
Community/parental engagement			
SC			
SDP			
School finances			
Inclusive education for girls			
Communication/reporting to higher levels			
Other			
Helpfulness of WEO's last visit (% schools)			
Very helpful			
Fairly helpful			
Not helpful			
WEO support to school is very good or good (% head teachers)			
WEO turnover (EQUIP-T input to output assumption)			
WEO has changed since beginning of 2016 (% schools)			

The number of head teachers who reported WEO advised or supported on topic x during last visit/all interviewed head teachers, expressed as a percentage.

Head teachers

The number of head teachers who reported WEO's last visit was very/fairly/not helpful/all interviewed head teachers who had been at school for at least a year, expressed as a percentage.

Head teachers

The number of head teachers who stated that the support of the WEO to the school is good or very good/all interviewed head teachers, expressed as a percentage.

Head teachers

The number of head teachers who report that the WEO has changed since beginning of 2016 /all interviewed head teachers, expressed as a percentage.

Head teachers

Are head teachers reporting to WEOs / districts? (EQUIP-T output)			
Head teacher provides written school reports to WEO/district (% schools):	The number of head teachers who provide written school reports to the WEO/district on basis x/all interviewed head teachers, expressed as a percentage.	Head teachers	
Monthly			
Quarterly			
Annually			
Head teacher does not provide written reports			
Content of the written reports (% schools):	The number of head teachers who report content x/all interviewed head teachers, expressed as a percentage.	Head teachers	
No report available			
Teacher attendance			
Teacher in-service training			
Other teacher information			
Pupil enrolment			
Pupil attendance			
Pupil academic performance			
Infrastructure/furniture			
Teaching and learning materials			
School committee information			
Parents/community information			
School budget or finance			
Extra-curricular activities			
Other			
Are head teachers attending ward education and COL meetings? (EQUIP-T output)			
Attended ward education meeting in last 60 days (% head teachers)	The number of head teachers who report they attended a ward education meeting in the last 60 days/all interviewed head teachers, expressed as a percentage.	Head teachers	Ward education meetings are chaired by the WEO and attended by HTs from the ward.
Number of ward education meetings attended in 2017 (mean)	The average of the total number of ward education meetings attended by the HT in 2017 (school mean).	Head teachers	Community of learning meetings are peer support meetings chaired by a head teacher and attended by other head teachers. The WEO can attend but not chair.
Head teacher attended COL meeting in last 60 days (% of head teachers)	Number of head teachers that attended a peer support meeting in the last 60 days/all interviewed head teachers, expressed as a percentage	Head teachers	

F.4 Chapter 6 Community participation and demand for accountability

Indicator name	Indicator definition	Respondent/unit of analysis	Notes
Has EQUIP-T provided training for school committees (SCs)? (EQUIP-T input)			
School committee exists (% schools)	The number of schools reporting that a school committee exists/all schools, expressed as a percentage.	Head teachers	
School committee received training (% schools)	The number of schools reporting that the school committee received training on its roles and responsibilities in the last two years/all schools, expressed as a percentage.	Head teachers	The relevant period for midline is 2014-2015; and for endline 2016-2017.
Provider of school committee training (% schools):	The number of schools reporting that the SC received training on its roles and responsibilities from provider x in the last two years/all schools, expressed as a percentage.	Head teachers	The relevant period for midline is 2014-2015; and for endline 2016-2017.
EQUIP-T			
LANES			
WEO and/or HT			
Other government official			
Other			
Has SC capacity increased? (EQUIP-T output)			
Head teachers rating SC support to school as 'very good' or 'good' (% head teachers)	The number of head teachers reporting the support of the SC to the school is 'very good' or 'good'/all interviewed head teachers, expressed as a percentage.	Head teachers	
School committee met in the last quarter (% schools)	The number of schools reporting that the SC met in the last quarter/all schools, expressed as a percentage.	Head teachers	
Minutes from last school committee meeting exist (% schools)	The number of schools where the head teacher could show minutes from the last meeting of the SC/all schools, expressed as a percentage.	Head teachers	
Main topics discussed at last SC meeting (% schools):	The number of schools where the last meeting of the SC covered topic x/all schools, expressed as a percentage.	Head teachers	
Academic progress			
Pupil absenteeism, discipline and/or dropout			
Teacher discipline			
Teacher supervision/support			
School development plan			
School finance including parental contributions			
Infrastructure development			
PTP / community engagement			
Pupil welfare			
Other			
Has EQUIP-T provided training for parent teacher partnerships (PTPs)? (EQUIP-T input)			
School has a PTP (% schools)	The number of schools reporting that a PTP exists/all schools, expressed as a percentage.	Head teachers	
PTP received training in last two years (% schools)	The number of schools reporting that the PTP received training on its roles and responsibilities in the last two years/all schools, expressed as a percentage.	Head teachers	The relevant period for midline is 2014-2015; and for endline 2016-2017.

Provider of PTP training (% schools):			
EQUIP-T	The number of schools reporting that the PTP received on its roles and responsibilities from provider x in the last two years/all schools, expressed as a percentage.	Head teachers	The relevant period for midline is 2014-2015; and for endline 2016-2017.
LANES			
WEO and/or HT			
Other			
Has EQUIP-T provided training on SCs, PTPs, grants and business plans for head teachers? (EQUIP-T input)			
Head teacher received EQUIP-T training on SC roles and responsibilities (% head teachers)	The number of head teachers reporting that they received training from EQUIP-T on SC roles and responsibilities in the last 4 years/all interviewed head teachers, expressed as a percentage.	Head teachers	
Head teacher received EQUIP-T training on PTP roles and responsibilities (% head teachers)	The number of head teachers reporting that they received training from EQUIP-T on PTP roles and responsibilities in the last 4 years/all interviewed head teachers, expressed as a percentage.	Head teachers	
Head teacher received EQUIP-T training on PTP grant 1 application and management (% head teachers)	The number of head teachers reporting that they received training from EQUIP-T on PTP grant 1 application and management in the last 4 years/all interviewed head teachers, expressed as a percentage.	Head teachers	
Head teacher received EQUIP-T training on business plan development and income generation (% head teachers)	The number of head teachers reporting that they received training from EQUIP-T on business plan development and income generation in the last 4 years/all interviewed head teachers, expressed as a percentage.	Head teachers	
Has PTP capacity increased and are PTPs active? (EQUIP-T output)			
Number of times PTP met in 2017 (mean)	The average number of times the PTP met in 2017	Schools	
PTP met at least four times in 2017 (% schools)	The number of schools where PTP met at least four times in 2017/all schools with a PTP, expressed as a percentage.	Schools	
PTP did not meet in 2017 (% schools)	The number of schools where PTP did not hold any meetings in 2017/all schools with a PTP, expressed as a percentage.	Schools	
Number of male and female teacher members of the PTP (mean)	The average number of male and female teacher members of the PTP	Schools	
Number of male and female parent members of the PTP (mean)	The average number of male and female parent members of the PTP	Schools	
Parents are members of the PTP (% parents of Standard 3 pupils)	The number of parents of Standard 3 pupils who report being members of the PTP/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Parents are aware that a PTP exists at the school (% parents of Standard 3 pupils)	The number of parents of Standard 3 pupils who are aware of the existence of a PTP at the school /all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Parents attended meetings to receive information about the PTP (% parents of Standard 3 pupils)	The number of parents of Standard 3 pupils who report attending any meeting that provides information about the PTP/all interviewed parents of Standard 3 pupils who are not members of the PTP, expressed as a percentage.	Parents of Standard 3 pupils	
Head teachers reporting that role of PTP and SC in school is (% schools):			
About the same	The number of head teachers reporting that the role of the PTP and SC in the school is about the same/somewhat different/very different/all interviewed head teachers, expressed as a percentage.	Head teachers	
Somewhat different			
Very different			
Has PTP taken action to improve education? (EQUIP-T intermediate outcome)			

PTP took action to improve education in the school in the last school year (% head teachers)	The number of head teachers reporting that the PTP took action to improve education in the school in the last school year/all interviewed head teachers, expressed as a percentage.	Head teachers	The relevant period for midline is 2015; and for endline 2017.
PTP took action to improve education in the school in the last school year (% Standards 1-3 teachers)	The number of Standards 1-3 teachers reporting that the PTP took action to improve education in the school in the last school year/all interviewed Standards 1-3 teachers, expressed as a percentage.	Standards 1-3 teachers	
PTP took action to improve education in the school in the last school year (% parents of Standard 3 pupils)	The number of parents of Standard 3 pupils reporting that the PTP took action to improve education in the school in the last school year/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Action taken by PTP to improve education in the school in the last school year (% head teachers):	The number of head teachers reporting that the PTP took action x in the last school year/all interviewed head teachers, expressed as a percentage.	Head teachers	
Teacher attendance and punctuality			
Pupil attendance and punctuality			
Community members assisting in classrooms			
Extra teaching and learning materials			
Extra tuition classes			
Extra tests/exams			
School infrastructure			
School feeding			
IGA			
Fundraising			
Pupil welfare			
Extra-curricular activities			
Other			
Has school and community interaction improved? (EQUIP-T intermediate outcome)			
Head teachers rating community support to the school as 'very good' or 'good' (% head teachers)	The number of head teachers reporting the support of the community to the school is 'very good' or 'good'/all interviewed head teachers, expressed as a percentage.	Head teachers	This means a meeting where all parents are invited, not a meeting of the parent teacher partnership (PTP).
Head teacher holds at least one meeting per year with teachers and all parents (% schools)	The number of schools reporting that they held at least one meeting with teachers and all parents last year/all schools, expressed as a percentage.	Head teachers	
Main topics discussed at last teacher and all parents meeting (% schools):	The number of schools where the last meeting of the teachers and <i>all</i> parents meeting covered topic x/all schools, expressed as a percentage.	Head teachers	
Academic progress			
Pupil discipline			
Pupil absenteeism, discipline and/or dropout			
Teacher discipline			
Teacher supervision/support			
School development plan			
School finance incl. parental contributions			
Infrastructure development			

PTP / community engagement			
Pupil welfare			
School committee			
Other			
Parents only received written information from the school about their child's academic progress in 2017 (% parents of Standard 3 pupils)	The number of parents reporting receiving written information from school about their child's academic progress in 2017 but not meeting with a teacher/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Parents only met with a teacher to receive information about their child's academic progress in 2017 (% parents of Standard 3 pupils)	The number of parents reporting meeting with a teacher to receive information about their child's academic progress in 2017 but not receiving written information from the school/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Parents received written information from the school and met with a teacher on their child's academic progress in 2017 (% parents of Standard 3 pupils)	The number of parents reporting receiving written information from the school and meeting with a teacher on their child's academic progress in 2017/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Parents did not receive written information from the school nor meet with a teacher on their child's academic progress in 2017 (% parents of Standard 3 pupils)	The number of parents reporting not receiving written information from the school nor meeting with a teacher on their child's academic progress in 2017/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Has EQUIP-T provided PTP grants (EQUIP-T input) and have PTP grants been spent? (EQUIP-T output)			
School received PTP grant 1 (% schools)	The number of schools reporting receiving PTP grant 1/all schools, expressed as a percentage.	Head teachers	
School received correct amount of PTP grant 1 (% schools)	The number of schools reporting receiving the correct amount of PTP grant 1/all schools who received PTP grant 1, expressed as a percentage.	Head teachers	The correct amount is TZS 550,000 per grant
PTP grant 1 has been spent (% schools)	The number of schools reporting that they have spent the PTP grant 1/all schools, expressed as a percentage.	Head teachers	
PTP grant 1 was spent on (% schools):			
Infrastructure and furniture			
Admin expenses			
Pupil welfare			
Teaching and learning			
Extra-curricular			
Other			
Has EQUIP-T provided IGA grants (EQUIP-T input) and have IGAs started? (EQUIP-T output)			
School and community developed a business plan for IGA and submitted a proposal to EQUIP-T in 2016 or 2017 (% schools)	The number of schools reporting that the school together with the community developed a business plan for income generating activities and submitted a proposal to EQUIP-T in 2016 or 2017/all schools, expressed as a percentage.	Head teachers	
Type of IGA proposed (% schools):			
Agriculture/horticulture			
Livestock /livestock products			
Trading/sales			

Manufacturing/processing			
IGA business plan proposal was successful (% schools)	The number of schools reporting that their IGA business plan proposal to EQUIP-T in 2016 or 2017 was successful/all schools, expressed as a percentage.	Head teachers	
School received IGA grant (% schools)	The number of schools reporting receiving IGA grant/all schools, expressed as a percentage.	Head teachers	
School received correct amount of IGA grant (% schools)	The number of schools reporting receiving the correct amount of IGA grant/all schools that received IGA grant, expressed as a percentage.	Head teachers	The correct amount is TZS 1,500,000 per grant
Some IGAs have started (% schools)	The number of schools reporting that at least some of the income generating activities the school received the grant for have started/all schools that received IGA grant, expressed as a percentage.	Head teachers	
Has EQUIP-T provided school notice boards? (EQUIP-T input)			
School received notice board from EQUIP-T in the last two years (% schools)	The number of schools that received a notice board in the last two years supplied by EQUIP-T/all schools, expressed as a percentage.	Schools	
Are school notice boards publicly accessible and used? (EQUIP-T output)			
Schools with notice board publicly displayed on school premises (% schools)	The number of schools that have a notice board displayed publicly on school premises/all schools, expressed as a percentage.	Schools	
Types of info displayed on school notice board (% schools):			
SDP/budget/financial/grants			
Academic results/teaching and learning related	The number of schools that display information x/all schools with a publicly displayed notice board, expressed as a percentage.	Schools	
Pupil/teacher attendance			
Community and school events			
JUU clubs/pupil welfare			
Are parents aware of and reading school notice boards? (EQUIP-T intermediate outcome)			
Parents are aware that a school notice board exists at school (% parents of Standard 3 pupils)	The number of parents of Standard 3 pupils that reported that a school notice board exists at school/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Parents read the notice board at least once in Jan-Mar 2018 (% parents of Standard 3 pupils)	The number of parents of Standard 3 pupils who read the notice board at least once in Jan-Mar 2018/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Number of times parents read the notice board in Jan-Mar 2018 (mean)	The average number of times parents read the notice board in Jan-Mar 2018	Parents of Standard 3 pupils	
Was a community education needs assessment (CENA) undertaken and actions taken based on it? (EQUIP-T input and intermediate outcome)			
Community carried out CENA over last four years (% schools)	The number of schools reporting that the community carried out its own education needs assessment and wrote it down in last 4 years (2014-2017)/all schools, expressed as a percentage.	Head teachers	
Action was taken by school or community based on CENA in last four years (% schools)	The number of schools reporting that the school or the community took action based on the CENA in the last four years/all schools with a CENA, expressed as a percentage.	Head teachers	
School/community actions taken based on CENA (% schools):			
School infrastructure	The number of schools reporting that the school or the community took action x in the last two years/all schools with a CENA, expressed as a percentage.	Head teachers	
Pupil attendance and punctuality			
School feeding			

Teacher attendance and punctuality			
Fundraising			
Extra-curricular activities			

F.5 Chapter 7 Conducive learning environments for marginalised children, particularly for girls and children with disabilities

Indicator name	Indicator definition	Respondent/unit of analysis	Notes
Profile of pupil vulnerability in programme schools			
Pupil has visual difficulties (% Standard 3 pupils)	Number of Standard 3 pupils reporting seeing difficulties/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	Pupils self-reported their disability status by answering four of the Washington Group's short set of questions on disability: 'Do you have difficulties seeing, even if wearing glasses?'; 'Do you have difficulties hearing, even if using a hearing aid?'; 'Do you have difficulties walking or climbing steps?'; 'Do you have difficulties remembering or concentrating?'. These questions were taken from DFID's guide to disaggregating programme data by disability (undated) that was shared with the evaluation team in early 2016, just prior to the midline survey.
Pupil has hearing difficulties (% Standard 3 pupils)	Number of Standard 3 pupils reporting hearing difficulties/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil has movement difficulties (% Standard 3 pupils)	Number of Standard 3 pupils reporting walking or climbing difficulties/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil has memory/concentration difficulties (% Standard 3 pupils)	Number of Standard 3 pupils reporting memory or concentration difficulties/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil has any disability (% Standard 3 pupils)	Number of Standard 3 pupils reporting seeing, hearing, walking or climbing, or memory or concentration difficulties/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil has any physical disability (% Standard 3 pupils)	Number of Standard 3 pupils reporting seeing, hearing, walking or climbing difficulties/all Standard 3 pupils, expressed as a percentage	Standard 3 pupils	
Pupil takes 45 minutes or longer to get to school (% Standard 3 pupils)	Number of Standard 3 pupils that take 45 minutes or longer to get to school/all Standard 3 pupils, expressed as a percentage.	Standard 3 pupils	
Absenteeism rate in Jan-Mar 2018 (% school days in Q1-2018)	Number of days absent in January, February and March of 2018/total number of schools days in that period, expressed as a percentage, for each Standard 3 pupil.	Standard 3 pupils	
Have schools received the various inputs under sub-component 4B as intended? (EQUIP-T input)			
School received training on setting up and running a JUU club in 2016 or 2017 (% schools)	The number of schools reporting that the school received training on setting up and running a JUU club in 2016 or 2017/all schools, expressed as a percentage.	Schools	
School received PTP girls' education grant (% schools)	The number of schools reporting that they received PTP girls' education grant (grant 2)/all schools, expressed as a percentage.	Schools	
School received correct amount of PTP girls' education grant (% schools)	The number of schools reporting that they received the correct amount of PTP girls' education grant/all schools that received grant, expressed as a percentage.	Schools	The correct amount is TZS 550,000 per grant
School received copies of Shujaaz magazine in 2016 or 2017 (% schools)	The number of schools that received copies of Shujaaz magazine in 2016-2017/all schools, expressed as a percentage.	Schools	
Are JUU clubs established, active and supporting marginalised children? (EQUIP-T output and intermediate outcome)			
School has a JUU club (% schools)	The number of schools reporting that the school has a JUU club/all schools, expressed as a percentage.	Schools	

Number of male and female pupils in the JUU club (mean)	The average number of male and females pupils in the JUU club	Schools	
Number of times JUU club met in 2017 (mean)	The average number of times the JUU club met in 2017	Schools	
JUU club carried out activities in 2017 (% schools)	The number of schools reporting that the JUU club has carried out activities in 2017/all schools with a JUU club, expressed as a percentage.	Schools	
Activities the JUU club has taken in 2017 (% schools):	The number of schools reporting that the JUU club carried out activity x in 2017/all schools reporting that the JUU club took any activity in 2017, expressed as a percentage.	Schools	
Environment			
Pupil attendance/punctuality			
Health/hygiene			
Pupils' safety			
Rights to education			
Extra-curricular			
Outside school community			
IGAs			
Gender equality			
Other			
Some of the activities the JUU club carried out in 2017 were for particular groups of vulnerable pupils (% schools):			
None			
Girls			
Pupils with disabilities			
Pupils with learning difficulties			
Pupils from poor households			
Pupils who don't speak Kiswahili at home			
Pupils with parents not interested in education			
Pupils who live far from school			
Pupils who are regularly absent	The number of schools reporting that the JUU club has access to copies of the Shujaaz magazine/all schools with a JUU club, expressed as a percentage.	Schools	
JUU club has access to copies of the Shujaaz magazine (% schools)			
Has the PTP girls' education grant been spent and on what? (EQUIP-T output and intermediate outcome)			
PTP girls' education grant has been spent (% schools)	The number of schools reporting that the PTP girls' education grant has been spent/all schools, expressed as a percentage.	Schools	
PTP girls' education grant was spent on (% schools):	The number of schools reporting that the PTP girls' education grant was spent on expenditure x/all schools reporting that the PTP grant had been spent, expressed as a percentage.	Schools	
Infrastructure and furniture			
Admin expenses			

Pupil welfare			
Teaching and learning			
Extra-curricular			
Other			
Some of the expenditures from the PTP girls' education grant were targeted at particular groups of vulnerable pupils (% schools):			
None	The number of schools reporting that some of the expenditures from the PTP girls' education grant were targeted at group x of vulnerable pupils/all schools reporting that the PTP grant had been spent, expressed as a percentage.	Schools	
Girls			
Pupils with disabilities			
Pupils with learning difficulties			
Pupils from poor households			
Pupils who don't speak Kiswahili at home			
Pupils with parents not interested in education			
Pupils who live far from school			
Pupils who are regularly absent			
Have inputs related to child protection and positive behaviour management been received and are schools promoting child protection and anti-violence? (EQUIP-T input, output, and intermediate outcome)			
Schools received posters on positive and safe learning environment (% schools)	The number of schools that report receiving posters on positive and safe learning environment in the last two years/all schools, expressed as a percentage.	Schools	
Material displayed on walls about expected teacher and student behaviour or classroom rules (% lessons)	The number of Standard 2 lessons where materials are displayed on walls about expected teacher and student behaviour or classroom rules/all observed Standard 2 lessons in schools that received the poster, expressed as a percentage.	Standard 2 lessons observed	
School beats pupils as a punishment (% parents of Standard 3 pupils)	The number of parents of Standard 3 pupils who reported that beatings of pupils take place at the school as a form of punishment/all interviewed parents of Standard 3 pupils, expressed as a percentage.	Parents of Standard 3 pupils	
Pupil was beaten at school as a punishment in 2017 (% parents of Standard 3 pupils)	The number of parents of Standard 3 pupils who reported that their child was beaten at the school as a punishment in 2017/all interviewed parents of Standard 3 pupils that report schools practices corporal punishment, expressed as a percentage.	Parents of Standard 3 pupils	
School has a student suggestion box where students can raise issues anonymously (% schools)	The number of schools reporting that the school has a student suggestion box where students can raise issues anonymously/all schools, expressed as a percentage.	Schools	
Location of student suggestion box (% schools):	The number of schools reporting that the student suggestion box is located in location x/all schools with a suggestion box, expressed as a percentage.	Schools	
Not seen			
No permanent location			
Open space			
Discrete space			
Are teachers using inclusive teaching practices in the classroom to support marginalised children? (EQUIP-T intermediate outcome)			

Notices groups of pupils with learning difficulties (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 that report noticing groups of pupils in their classes that have learning difficulties/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	
Group of pupils identified to have learning difficulties (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 that report group x having learning difficulties in their classes/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	
With disabilities			
Girls			
Boys			
Don't speak Kiswahili at home			
Poor pupils			
Haven't attended preschool			
With health problems			
Parents not interested in education			
Live far from school			
Are regularly absent			
Other			
No particular group			
Able to help groups of pupils with learning difficulties (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 that report they are able to help pupils with learning difficulties/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	
Action to help pupils with learning difficulties (% Standards 1-3 teachers):	The number of teachers of Standards 1-3 that report action x to help pupils with learning difficulties/all interviewed teachers, expressed as a percentage.	Teachers of standards 1-3	
Adapt materials and teaching to app level			
Use regular assessment to monitor progress			
Ensure pupil engagement in lessons			
Give extra tuition classes			
Suggest extra tuition classes by others			
Switch btw Kiswahili and vernacular language			
Talk to pupil's parents			
Group pupils together			
Give more exercises and work			
Repeat topics until pupils understand			
Other			
Teacher reports speaking Kiswahili when teaching (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 reporting they speak Kiswahili when teaching/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	

Teacher reports speaking Kiswahili with pupils outside the classroom (% Standards 1-3 teachers)	The number of Standards 1-3 teachers reporting they speak Kiswahili with pupils outside the classroom/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	
Teacher reports switching between Kiswahili and a vernacular language when teaching (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 reporting they switch between Kiswahili and a vernacular language when teaching/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	
Teacher switches between Kiswahili and a vernacular language when teaching (% Standard 2 lessons)	The number of observed Standard 2 lessons where teacher switches between Kiswahili and a vernacular language while teaching/all observed Standard 2 lessons, expressed as a percentage	Standard 2 lessons observed	
Teacher provided extra support to non-native Kiswahili speaking pupils (% Standard 2 lessons)	The number of observed Standard 2 lessons where teacher provided extra support to non-native Kiswahili speaking pupils/all observed Standard 2 lessons, expressed as a percentage	Standard 2 lessons observed	
Teacher can speak same local language as pupil (% Standard 3 pupils)	Number of Standard 3 pupils reporting their teacher can speak the same local language as they/all assessed Standard 3 pupils whose main language spoken at home is vernacular, expressed as a percentage.	Standard 3 pupils	
Teacher speaks Kiswahili at home (% Standards 1-3 teachers)	The number of teachers of Standards 1-3 reporting they speak Kiswahili at home/all interviewed teachers of Standards 1-3, expressed as a percentage.	Teachers of standards 1-3	
Are there inclusive strategies at the school-level that support marginalised children? (EQUIP-T intermediate outcome)			
School development plan includes strategies to improve pupil welfare or to improve girls' learning (% schools)	The number of schools that have a school development plan with strategies to improve pupil welfare or to improve girls' learning/all schools, expressed as a percentage.	Schools	
School development plan includes strategies targeted at particular groups of vulnerable pupils (% schools):	The number of schools with a SDP with strategies targeted at group x of vulnerable pupils/all schools, expressed as a percentage.	Schools	
None			
Girls			
Pupils with disabilities			
Pupils with learning difficulties			
Pupils from poor households			
Pupils who don't speak Kiswahili at home			
Pupils with parents not interested in education			
Pupils who live far from school			
Pupils who are regularly absent			
Head teacher took action in 2017 to improve pupil welfare (% head teachers)	The number of head teachers that reported taking action in 2017 to improve pupil welfare/all head teachers who had been at school for at least a year, expressed as a percentage.	Head teachers	
Some of the head teacher actions in 2017 to improve education in the school were targeted at particular groups of vulnerable pupils (% head teachers):	The number of head teachers reporting that some of their actions in 2017 to improve education in the school were targeted at group x of vulnerable pupils/all head teachers who had been at school for at least a year, expressed as a percentage.	Head teachers	
None			
Girls			

Pupils with disabilities			
Pupils with learning difficulties			
Pupils from poor households			
Pupils who don't speak Kiswahili at home			
Pupils with parents not interested in education			
Pupils who live far from school			
Pupils who are regularly absent			
Some of the head teacher actions in 2017 to improve education in the school were targeted at particular groups of vulnerable pupils, as reported by teachers (% Standards 1-3 teachers):			
None			
Girls			
Pupils with disabilities	The number of Standards 1-3 teachers reporting that some of the head teachers actions in 2017 to improve education in the school were targeted at group x of vulnerable pupils/all Standards 1-3 teachers who had been at school for at least a year, expressed as a percentage.	Standards 1-3 teachers	
Pupils with learning difficulties			
Pupils from poor households			
Pupils who don't speak Kiswahili at home			
Pupils with parents not interested in education			
Pupils who live far from school			
Pupils who are regularly absent			
WEO's advice or support to the school in the last visit was about pupil welfare or inclusive education for girls (% schools)			The number of schools reporting that on his/her last visit, the WEO provided advice or support to the school on pupil welfare or inclusive education for girls/all schools, expressed as a percentage.
WEO's advice or support to the school in the last visit was targeted at particular groups of vulnerable pupils (% schools):			
None			
Girls			
Pupils with disabilities	The number of schools reporting that on his/her last visit, the WEO provided advice or support to the school targeted at group x of vulnerable pupils/all schools, expressed as a percentage.	Schools	
Pupils with learning difficulties			
Pupils from poor households			
Pupils who don't speak Kiswahili at home			
Pupils with parents not interested in education			
Pupils who live far from school			
Pupils who are regularly absent			
PTP took action in 2017 to improve pupil welfare (% schools)	The number of schools reporting that the PTP took action in 2017 to improve pupil welfare/all schools, expressed as a percentage.	Schools	

Some of the PTP's actions in 2017 to improve education in school were about a particular group of vulnerable pupils (% schools)	The number of schools reporting that some of the PTP's actions in 2017 to improve education in school were for group x of vulnerable pupils/all schools, expressed as a percentage.	Schools	
None			
Girls			
Pupils with disabilities			
Pupils with learning difficulties			
Pupils from poor households			
Pupils who don't speak Kiswahili at home			
Pupils with parents not interested in education			
Pupils who live far from school			
Pupils who are regularly absent			
Main topic of discussion at the last all parents-teachers meeting was on pupil welfare (% schools)	The number of schools reporting that the main topic of discussion at the last all parents-teachers meeting was on pupil welfare/all schools, expressed as a percentage.	Schools	
School has a school feeding programme that has provided food to pupils in last five school days (% schools)	The number of schools reporting that there is a school feeding programme that has provided food to pupils in the last five school days/all schools, expressed as a percentage.	Schools	
School has a teacher responsible for coordinating gender issues in the school (% schools)	The number of schools reporting that the school has a teacher responsible for coordinating gender issues in the school/all schools, expressed as a percentage.	Schools	
Teacher responsible for coordinating gender issues in the school attended specific gender training in 2016 or 2017 (% schools)	The number of head teachers reporting that the teacher responsible for coordinating gender issues in the school has attended specific gender training such as GRP or gender-inclusive environments in 2016 or 2017/all schools with a gender coordinator, expressed as a percentage.	Schools	
School has special classes (% schools)	The number of schools reporting that the school has special classes/all schools, expressed as a percentage.	Schools	These are classes for pupils with special needs such as children with disabilities

Annex G Statistical tables of results from programme areas

See separate document in excel format.

Annex H Implementation of other large education programmes

H.1 LANES

Table 52: LANES activities in 2014 and 2015

Overview	2014/15 to 2016/17 funded by Global Partnership for Education US\$95m budget
Objectives	Improved basic skills in literacy and numeracy for children aged 5-13 years
Expected outputs	Improved teaching and learning; improved education sector management; increased community participation
Geographical coverage	14 regions for training ¹ : <i>Kagera, Mwanza, Geita, Arusha, Kilimanjaro, Tanga, Manyara, Dar es Salaam, Morogoro, Singida, Pwani, Rukwa, Katavi and Ruvuma</i> ; national for materials distribution
Main activities in 2014 & 2015	Training of 18,656 standards one and two teachers on the new 3Rs curriculum (9 days, centralised training model in Dodoma, delivered by TTC tutors)
	Training of 10,870 head teachers, and 2,480 WECs in school leadership and management (3 days, regional training model, delivered by ADEM; 3 additional regions Iringa, Mbeya, Njombe)
	Materials development and distribution to schools via these trainees of: standards 1&2 curriculum; std 1 syllabus; std 2 syllabus; stds1&2 teachers guide for reading/writing; stds 1&2 teachers guide for maths; school leadership and management guidelines (on general school management and 3Rs programme implementation) ²
	Materials development and procurement of (national distribution planned for 2016): 6 std 1 textbooks: reading, story book writing, maths, health, art & sports).
	Production and distribution of Primary School Leaving Examination (PSLE) item analysis booklets for each of 8 subjects to regions and districts for forwarding to all schools

Sources: (i) MoEVT (2015) (ii) Interview with LANES National Co-ordinator (January 2015). Note: (1) The IE control districts are in the regions highlighted in italics. (2) BRN-Ed developed the general SLM guideline.

Table 53: LANES activities in 2016 and 2017

Overview	2014/15 to 2018/19 funded by Global Partnership for Education US\$95m budget ¹
Objectives	Improved basic skills in literacy and numeracy for children aged 5-13 years
Expected outputs	Improved teaching and learning of literacy and numeracy; improved education sector management; increased community engagement in literacy and numeracy support
Geographical coverage	All 26 regions for most activities (~16,000 govt primary schools). Some exceptions noted below where 7 EQUIP-T regions or 4 Tusome Pamoja regions are excluded.
Main activities in 2016 & 2017 (relevant to EQUIP-T)	Teacher INSET
	Training of 31,966 std 3&4 teachers (2 per school), & selected regional/district officials, on revised std 3&4 curriculum; ADEM led in zonal ADEM centres, ~1 week, trainers TTC tutors & TIE officials
	Training of 16,075 pre-primary teachers (1 per school) on revised pre-primary curriculum; zonal training, ~ 1 week
	Training of 697 teachers who teach learners with intellectual impairments & 1,120 teachers of learners with visual or hearing impairments on 3Rs
	Training of limited number of std1&2 teachers from large schools that missed out on std1&2 curriculum training in 2015 [<i>selected large schools, not EQUIP-T regions</i>]
	Materials distribution
	Distribution to schools of: additional copies of std 1&2 curriculum package (syllabus, teachers' guides)
	Distribution to schools of: standards 3&4 curriculum package (syllabus and teachers' guides); std 1 textbooks (6 titles); std 2 textbooks (5 titles); std 3 textbooks (6 titles); Guide for 'talking classrooms'; Primary School Leaving Examination (PSLE) item analysis booklets
	Materials development and procurement of (not distributed yet): pre-primary books (7 titles); pre-primary curriculum package (syllabi, teachers' guides etc); story books (25 titles) [<i>story books not Tusome Pamoja regions</i>]
	Disbursement modality finalised for direct grant to schools for establishing 'talking classrooms'
	External school support
	Materials development for school quality assurance (draft stage): school quality assurance framework, inspection tools for school quality assurers (SQAs, formerly regional and district inspectors) & district and ward-level officials, plus some vehicles & funds for school inspection
	Training of 2,480 WEOs on school supervision of 3Rs [<i>Not EQUIP-T regions</i>]
	Funds for regional and district (including WEO) officials school visits, based on a monitoring guide [<i>Not 7 EQUIP-T regions</i>]
	2,894 Motorbikes procured but not yet delivered for WEOs [<i>Not EQUIP-T regions</i>]
	Community engagement
Orientation of 129,609 School Management Committee members & distribution of SMC guideline (2 per school); TOT model, ward level meeting, 2 days, trainers district/regional officials [<i>Not EQUIP-T regions</i>]	

Sources: (1) MOEST (2017b); (2) Interview with LANES Co-ordinator, February 2018. Note: (1) LANES started in July 2014 with a planned duration of 3 years (2014/15 to 2016/17); this was extended by 1.5 years to December 2018.

H.2 BRN-Ed/EPforR

Table 54: BRN-Ed programme activities in 2014 and 2015

Overview	2014-2018 funded by World Bank, SIDA, DFID and GoT US\$416m budget
Objectives	To improve education quality in Tanzanian primary and secondary schools. Aim to see gains in the following indicators: (i) national average performance of std 2 students in reading; (ii) national average performance of std 2 students in numeracy; (iii) percentage of teachers found in classrooms during unannounced visit in primary and secondary schools; and (iv) percentage of primary teachers with minimum knowledge in mathematics and languages
Expected outputs	Outputs are linked to 9 priority activities for education, devised to be quick wins: (1) Official school ranking; (2) National 3Rs assessment; (3) School incentive scheme (financial and non-financial rewards); (4) Teacher motivation (non-financial performance incentives for teachers and clear backlog of claims); (5) School improvement toolkit ¹ ; (6) 3R teacher training ² ; (7) Student teacher enrichment programme (STEP) ³ ; (8) Capitation grants; (9) Basic facilities construction. There is a performance-based funding mechanism in place for delivery on (1) to (8).
Geographical coverage	All regions, but some interventions target groups of schools ⁴
Main activities in 2014 & 2015	(1) Official school ranking published based on public examination results (2015)
	(3.) Cash grant given to 120 schools on basis of ranking ⁵
	(4) Clearing of some of the backlog of teacher claims older than 3 months (2015); Teacher awards announced to high-performing teachers (2014)
	(5) Delivery of school improvement toolkits to 9,431 schools (2015) [NB: this is a subset of the LANES activities already detailed above]
	(6) 4,175 teachers participating in 3R training programme (2014) [NB: this is a subset of the LANES activities already detailed above]
	(7) 4,337 primary schools conducting STEP (2015); 1,317 secondary schools conducting STEP (2015)

Sources: World Bank (2015); World Bank (2014). Notes: (1) A guide on best practices to manage a school, as well as training for head teachers to drive quality improvement. (2) Training on how to teach basic skills effectively for std 1 and std 2 teachers in 40 low-performing districts via a cascade model. (3) Training teachers to identify and support low performing students, via diagnostic tests and additional classes (focused at upper primary and lower secondary). (4) Information is not readily available on the location of these target schools. At baseline, the IE survey sample took care to exclude any of the 60 districts that at the time were listed BRN programme areas. (5) DFID provided this information (it was not mentioned in the Oct 2015 implementation status report). Information is not readily available on where the 120 schools are located.

Table 55 EPforR activities in 2016 and 2017

Overview	2014/15-2020/21 funded by World Bank, SIDA, DFID and GoT US\$416m budget ¹
Objectives	To improve student learning outcomes at primary and lower secondary education levels in Tanzania
Expected outputs	System-level improvements including more efficient utilization of financial and human capital resources at central and local levels; better education service delivery via improved accountability and incentive mechanisms at school-level
Performance areas (related to disbursement-linked indicators in 2016/17 & 2017/18)	(i) Improved educational outcomes: Std 2 reading fluency in Swahili (16/17, 17/18) & addition/subtraction skills (17/18)
	(ii) Adequate and timely resource flows from govt education budget: capitation grants, textbook expenditure, teacher non-salary claims, monitoring & evaluation, school incentive grants (16/17 & 17/18); and receipt of textbooks in schools (17/18)
	(iii) Improved results monitoring and information management: timely EMIS data on-line, & annual education sector performance report (16/17 & 17/18)
	(iv) More equitable teacher deployment across and within districts: primary PTRs in acceptable range (16/17 & 17/18)

	(v) School incentive grants (SIG) for most improved or best performance in national exams: 60 (3000) primary schools got monetary (non-monetary) awards (July 2016); 311 (~3,000) primary schools got monetary (non-monetary) awards (July 2017)
	(vi) Improved student retention: district primary retention rates (16/17 & 17/18) & regional girls' transition rates from primary to secondary (17/18)
	(vii) Improved school quality: whole school quality assurance visits (WSVs) (17/18)
	(viii) Stronger national capacity for planning, policy and innovation: commissioned assignments (17/18)
Geographical coverage	National, except for (v) SIGs—monetary and non-monetary awards to schools with most improved or best exam performance ²

Sources: (i) MOEST/PO-RALG (2017); (ii) World Bank (2018). Notes: (1) EPforR (then BRN-Ed) started in July 2014 with a planned duration of ~4 years (2014/15 to 2018/19); this was extended by ~2.5 years to December 2020 (2019/20 and 2020/21). (2) SIG monetary awards went to <0.5% and 2% of primary schools in July 2016 and July 2017 respectively.

H.3 Tusome Pamoja

Table 56 Tusome Pamoja (TP, let's read together) activities in 2016 and 2017

Overview	2016 to 2021 funded by USAID¹
Objectives	Improved age-appropriate, curriculum defined levels of reading and writing at stds 2 & 4 for at least 75% of classrooms in target areas
Expected outputs	Improved quality of early grade basic skills instruction; strengthened skills delivery and assessment system; effective engagement of parents and community in education
Geographical coverage	4 mainland regions: Iringa, Morogoro, Mtwara and Ruvuma ² (2,754 schools); all 11 districts of Zanzibar. All number in the activities summary below refer to mainland.
Main activities in 2016 & 2017 (relevant to EQUIP-T)	Teacher INSET
	Training of 7,183 std 1&2 teachers on curriculum delivery & readers; 4 + 3 days (Feb & July 2017), trainers district officials
	Training of HTs (2,830), WEOs (682), Academic teachers (2,552) on curriculum delivery & readers; 4 days, trainers district officials
	Training of 255 pre-primary teachers/volunteers on curriculum delivery & story books [Mtwara only]
	<i>Planned</i> INSET: std 1&2 teachers on maths materials; std 3&4 on non-fiction titles
	Materials distribution
	Distribution to schools: std 1&2 levelled classroom supplementary readers (10 titles); std 1&2 teacher read alouds/big books (5 titles)
	<i>Planned</i> distribution to schools: std 1&2 decodeable readers; std 1&2 maths materials; stds 3&4 non-fiction titles
	External school support
	Training of 163 SQAs on decentralised periodic learning assessment (DPLA)
	<i>Planned</i> distribution of tablets to schools/WEOs for school information system (SIS)
	<i>Planned</i> grants to WEOs/District official to support school visits
	Community engagement
	Training of 2,754 head teachers & school committee members & 702 WEOs on Parent teacher partnerships (PTP) set-up ; trainers district officials, Dec 2016; Setting up 2,754 PTPs
Training of Community Education Mobilisers (CEMs, 2 per community); Community education mobilisation and action plans (CEMAP) completed ³	
Distribution of 2,754 school noticeboards	

Source: RTI (2017) 'USAID Tusome Pamoja Draft Annual Report 01 October 2016 to 30 September 2017'. Notes: (1) Tusome Pamoja started in January 2016 with a start-up phase, implementation started in October 2016, planned to run until January 2021 (~ 5 years). Total budget is not specified but estimated budget for Year 2 is ~US\$29m. (2) Ruvuma is an EQUIP-T IE control district. (3) Examples of CEMAP activities include reading at home, remedial reading classes, OOSC initiatives, anti-truancy, school feeding, and infrastructure.

About the project

The independent Impact Evaluation of the Education Quality Improvement Programme in Tanzania (EQUIP-T) is a study funded by the United Kingdom Department for International Development (DFID). It is designed to: i) generate evidence on the impact of EQUIP-T on primary pupil learning outcomes, including any differential impacts for girls and boys; ii) examine perceptions of effectiveness of different EQUIP-T components; iii) provide evidence on the fiscal affordability of scaling up EQUIP-T after the programme ends; and iv) communicate evidence generated by the impact evaluation to policy-makers and key education stakeholders.

EQUIP-T is a six-year Government of Tanzania programme, funded by UK DFID, which seeks to improve the quality of primary education in nine regions of Tanzania, and thus to improve learning outcomes, particularly for girls. It focuses on strengthening performance of teachers, school leadership and management, systems which support district management of education, and community participation in education.



Oxford Policy Management

